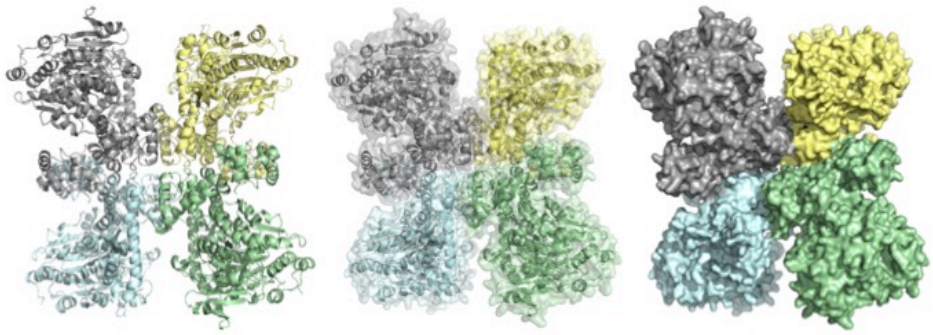
Molecular Analysis of Human and Archaeal DNA Repair Helicases HelQ and Hel308



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

Sarah J. Northall, B.Sc. (Hons.)

March 2017

Abstract

Completion of genome duplication by DNA replication catalyzed at stable replisomes is essential for life by facilitating cell division. Replisomes encounter physical blockage and chemical damage to DNA that frequently threatens to derail replication by inhibiting replisome enzymes. Multiple systems support DNA replication, by detection and repair of DNA damage and removal of physical blocks. Homologous recombination (HR) is one example. Archaeal Hel308 and metazoan HelQ DNA helicases are implicated in DNA repair by HR, in response to toxic DNA interstrand crosslinks that block replication forks. HelQ and Hel308 are single-strand DNA (ssDNA) stimulated ATPases with 3' to 5' translocase/helicase activity, most effective at unwinding forked DNA. Their helicase activities are likely to be crucial for promoting DNA replication and repair, but little is known about how. Here, I have been able to generate high yields of highly active human HelQ protein, compared to previous published strategies, and used *in vitro* biochemistry to show multiple oligomeric states of HelQ that are sensitive to reducing agents. I show that HelQ preferentially targets branched DNA molecules for DNA unwinding similarly to existing data for archaeal Hel308. HelQ and Hel308 demonstrated conservation of function between HelQ and archaeal Hel308 in winged helix domains of both proteins that led to a model for bi-modal DNA binding. This highlights how archaeal Hel308 may also be used as a model for HelQ function.

Table of Contents

List of Figures

List of Tables

List of abbreviations

A	Adenine
ADP	Adenosine diphosphate
AGF	Analytical gel filtration
AP sites	Apurinic sites
ATP	Adenosine triphosphate
ATM	Ataxia-telangiectasia mutated
ATR	Ataxia telangiectasia and Rad3-related
BER	Base excision repair
BIR	Break induced replication
BME	β-mercaptoethanol
BN-PAGE	Blue native polyacrylamide gel electrophoresis
BSA	Bovine serum albumin
COTH	CO-threader
CRISPR	Clustered regularly interspaced short palindromic repeats
C	Cytosine
D-loop	Deoxyribonucleic acid loop
DEAE	Diethylaminoethanol
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
dHJ	Double Holliday Junction
dsDNA	Double stranded deoxyribonucleic acid
DSB	Double strand break
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
EMSA	Electrophoretic mobility shift assay
FA	Fanconi anaemia
FBS	Fetal bovine serum
FRET	Förster resonance energy transfer
G	Guanine

G1	Gap 1
G4	G quadruplex
GFP	Green fluorescent protein
GST	Glutathione S-transferase
HJ	Holliday junction
HMM	Hidden markov model
HR	Homologous recombination
ICL	Interstrand crosslink
IMAC	Immobilized metal affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IR	Infrared
I-TASSER	Iterative threading assembly refinement
LB	Luria Bertani
M-phase	Mitotic phase
MMC	Mitomycin C
MMEJ	Microhomology mediated end joining
MST	Microscale thermophoresis
NAP	Nucleoid associated protein
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining
NPF	Nucleoprotein filament
NTA	Nitrilotriacetic acid
OB fold	Oligonucleotide binding fold
Ori	Origin of replication
ORC	Origin recognition complex
PAGE	Polyacrylamide gel electrophoresis
PDB	Protein data bank
PCNA	Proliferating cell nuclear antigen
PEG	Polyethylene glycol
Phyre 2	Protein homology/ analogy recognition engine version 2.0
RDR	Recombination dependent repair
R-loop	Ribonucleic acid loop

ROS	Reactive oxygen species
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SANS	Small angle neutron scattering
SAXS	Small angle X-ray scattering
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate poly acrylamide gel electrophoresis
SDSA	Synthesis dependent strand annealing
SDW	Sterile distilled water
SF2	Superfamily 2
SUMO	Small ubiquitin like modifier
ssDNA	Single stranded deoxyribonucleic acid
smFRET	Single molecule Förster resonance energy transfer
S-phase	Synthesis phase
T	Thymine
TBE	Trisaminomethane borate ethylenediaminetetraacetic acid
TLS	Translesion synthesis
TIRF	Total internal reflection fluorescence
TROLOX	6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid
TRIS	Trisaminomethane
UV	Ultraviolet
WHD	Winged helix domain
XP	Xeroderma pigmentosum

Chapter 1: Introduction

1.1 Historical context and origins of genetic material

It is thought that ribonucleic acid (RNA) was the original form of replicative life on Earth. Over time, storage of hereditary information transitioned from RNA to deoxyribonucleic acid (DNA) as options evolved for enzyme catalyzed chemical modification of ribonucleotides to deoxyribonucleotides[1]. The 'choice' for the overwhelming majority of most organisms to encode their genomes using DNA lies in the inherent stability of the DNA molecule, at least in comparison to RNA, and this transition likely has viral origins. By default, proteins that catalyse DNA metabolism, namely those that carry out DNA replication and repair, evolved parallel to this transition.

It is critical for bacteria, archaea and eukaryotes that genomic integrity of the DNA molecule(s) is maintained; this ensures faithful transmission of protein coding sequences and their associated regulatory sequences through generations. Genome maintenance occurs by mechanisms that orchestrate complex genome dynamics. Involved in DNA replication, transcription and repair, genome dynamics are required for these essential processes that underpin the central dogma of molecular biology: DNA to RNA to protein. Unchecked dynamics can lead to instability, requiring repair. As a result, intricate mechanisms have evolved to repair DNA damage and errors that may occur during DNA replication, and these are central to genomic integrity.

DNA repair can be global, or may be coupled to essential processes of DNA replication, gene transcription or protein translation. Comprehensive understanding of DNA repair is essential, as genome stability underpins DNA replication. This is relevant, because deregulated DNA replication can cause cancer in humans.

However, despite the requirement for genomic integrity, genomes require an element of plasticity. This generates variation that can be acted upon by selection, such as a random change to DNA that has remained unrepaired to generate a mutation or a change to the original DNA base sequence that is transmitted to future generations. Programmed genetic variation is also essential and is generated through manipulation of DNA to create genetic diversity during V(D)J recombination[2], to develop immunity, and homologous recombination during meiosis in eukaryotes[3].

The subject of this thesis especially focuses on DNA repair associated with DNA replication, with the major focus being processes collectively called homologous recombination. The DNA repair mechanisms in eukaryotes, and in particular metazoans, remain unclear in comparison to well-understood bacterial DNA repair systems. This is explained in part by lack of sequence homology between bacterial DNA repair proteins and those in eukaryotes and archaea making them difficult to identify. However, some progress is being made as a result of genetic and proteomic analyses that are identifying important genes involved in DNA repair – one of these encodes the DNA helicase HelQ, which is conserved in metazoans and archaea[4,5].

However, before we can understand the complexities of maintaining genomic integrity and a putative role for HelQ, we must first consider the DNA molecule itself.

1.2. The chemical structure of DNA

Hereditary information is stored in the form of RNA or DNA in all organisms. Protection of these biological molecules is critical as they are the foundations of the central dogma of molecular biology. Discovery of the double helix structure of DNA in 1953[6] revealed that this molecule intrinsically protects itself, in addition to providing hints as to how this molecule may be processed, first stated by Watson and Crick; 'It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material'.

A typical double stranded DNA molecule is comprised of two antiparallel strands formed from combinations of deoxyribonucleoside triphosphates (dNTPs) that contain purine and pyrimidine bases: adenine (A), guanine (G), thymine (T), and cytosine (C), Figure 1. A single nucleotide is formed from two condensation reactions, both requiring hydroxyl groups from deoxyribose (or ribose, for RNA) sugar. First, the sugar hydroxyl positioned at carbon-1 condenses with a secondary amine from a nitrogenous base. A second reaction is between hydroxyl at sugar carbon-5 and phosphoric acid, generating a nucleotide.

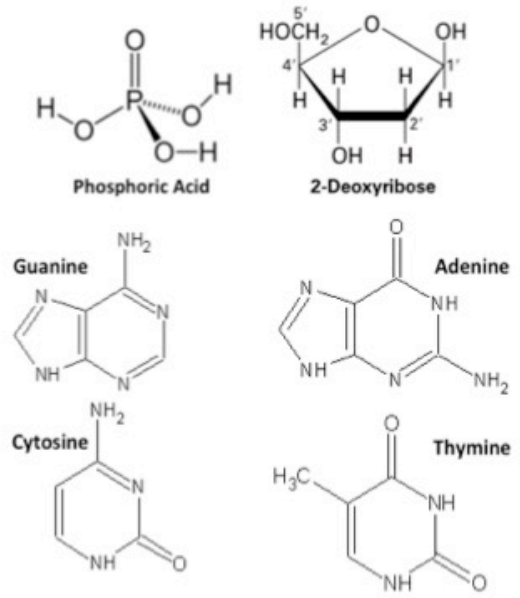


Figure 1.1. Molecular structures of the components of a nucleotide within DNA. Figure adapted from Fisher Scientific and opencurriculum.org.

Deoxyribonucleotides polymerise *via* covalent bonds between the 3'-OH on the deoxyribose moiety and the phosphate group, forming a polar, negatively charged phosphodiester backbone. The polar nature of the DNA backbone is important for the mechanisms by which many DNA processing enzymes manipulate DNA. The two strands are held together by hydrogen bonds that occur between A: T and G: C bases forming Watson-Crick base pairs, Figure 1.2, that can be transiently separated, for example by DNA helicases.

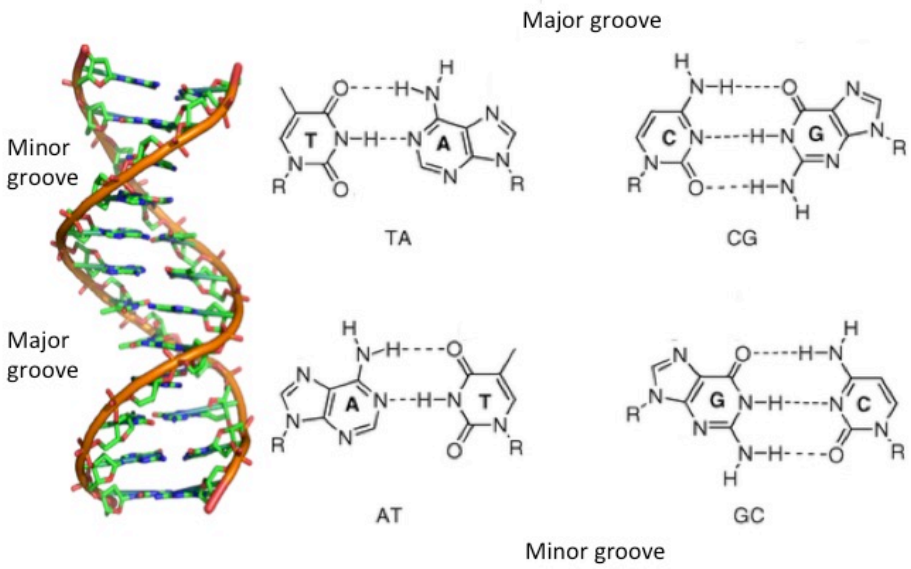


Figure 1.2. Left: cartoon of the double-helical structure of B-form DNA highlighting the sites of major and minor grooves, and showing the phosphodiester "backbone" in orange and base-pairs in green, blue and red, denoting atoms of carbon, nitrogen and oxygen respectively. Right: Atomic structures of each prevalent base pair found in B-form DNA, either adenine (A) to thymine (T) or guanine (G) to cytosine (C). Images adapted from researchgate.net and molecular coordinates downloaded from PDB 1BNA, deposited by Drew *et. al.*[7]

Duplex DNA can adopt several helical conformations: A, B or Z that are defined by helical sense, pitch, rise and diameter[8], summarized in Table 1.1. The most common native form is right handed type B. The helical topology of B-form duplex DNA gives rise to features known as major and minor grooves that are utilized by DNA binding proteins for recognition and manipulation of DNA. Such proteins can engage with major grooves *via* insertion of an α-helix into the major groove where amino acid R-groups make specific contacts with exposed hydrogen bonds for A:T, T:A, C:G and G:C pairs. This is known as direct readout where the

hydrogen bond code can be read by the protein. Different combinations of hydrogen bonds are exposed in the major grooves according to the base sequence. Alternatively, proteins can engage with DNA *via* the minor groove in a structure dependent manner known as indirect readout. However, DNA damage can often disrupt the typical topology of B-form DNA. This is because the chemical bonds in DNA (e.g. C=O, N-H and C-P-O) have chemistry susceptible to chemical rearrangement *via* nucleophilic attack, oxidation and hydrolysis. This can be dealt with by DNA repair

Table 1.1. A summary of the measurements and relevance of common forms of DNA topology.

	A	**B**	**Z**
Shape	Broad	Intermediate	Narrow
Rise per base pair	2.3Å	3.4 Å	3.8 Å
Helix diameter	25.5 Å	23.7 Å	18.4 Å
Screw sense	Right handed	Right handed	Left handed
Glycosidic bond	anti	Anti	Alternating anti, syn
Base pair per turn of helix	11	10.4	12
Pitch per turn of helix	25.3 Å	35.4 Å	45.6 Å
Tilt of base pairs from normal to helix axis	19°	1°	9°
Major groove	Narrow and very deep	Wide and quite deep	Flat
Minor groove	Very broad and shallow	Narrow and quite deep	Very narrow and deep

Diverse DNA repair proteins have evolved to recognize the many chemically specific types of DNA damage that may arise spontaneously or result from errors during DNA processing (e.g. in DNA replication). Repair of DNA damage to restore DNA into its original form is discussed later, but is necessary to prevent multiple forms of genomic instability, including point mutations, replication slippage, and other DNA and chromosomal re-arrangements that arise from DNA strand breaks. DNA repair proteins catalyze physical and chemical manipulation of compromised DNA molecules by duplex strand separation (e.g. ATP-dependent helicases), strand cutting and resection (e.g. nucleases and translocases), strand extension (e.g. polymerases) and joining together of DNA strands (e.g. ligases and annealases). These processes aim to restore DNA back to its original form with minimal disruption to the DNA base sequence.

Although the helical conformation can be considered to protect the information contained within the molecule by 'storing' the bases inside the double helix, these intrinsic protective properties of the DNA molecule are not sufficient to avoid chemical or physical insult. The cell must therefore utilize DNA protection, packaging, monitoring, signaling and repair processes to preserve genome stability, topics that are dealt with in detail in later sections.

1.3 DNA genomes in two or three domains of life

The chemical topology of DNA is universally conserved across all domains of life. DNA sequences however, are responsible for differences

between organisms and therefore this can be used to classify organisms. In the 1970s and 1980s various studies of 16S ribosomal RNA (rRNA) DNA sequences had suggested an evolutionary division between prokaryotes such that archaeabacteria were phylogenetically distinct from bacteria[9]. This was consolidated by biochemical analyses of RNA polymerase enzymes, and studies of antibiotic resistance, which indicated fundamental differences in DNA processing between bacteria and archaebacteria[10,11]. This eventually led to the split in prokaryotes from a monophyletic domain into two domains, Bacteria and Archaea[12].

Like many bacteria, archaea are notable for being extremophiles based on their environmental niches: psychrophiles, which inhabit unusually cold environments, thermophiles, which thrive at optimum temperatures between approximately 70°C and up to 100°C, halophiles, which tolerate high salt habitats, and methanogens, which generate methane by two major metabolic pathways and inhabit a broad spectrum of extreme environments. How macromolecules within these organisms remain stable under these extreme conditions is particularly interesting, and gives rise to understanding of properties that can be exploited for laboratory and biotechnological purposes[13]; for example, many DNA polymerases used on an everyday basis in the laboratory are derived from those found in thermophilic archaea. Halophiles typically grow optimally in highly saline conditions, typically around 4.5 M NaCl. These organisms are highly amenable to genetic analysis in archaea. Again, due to the extreme conditions that biological molecules exist in these organisms, in

this case, high salt, proteins derived from halophiles can be used for biotechnology.

The first archaeal genome sequence in 1996, of *Methanococcus janaschii*[14], and subsequently hundreds of prokaryotic and eukaryotic genome sequences, established about 300 identifiable genes or open reading frames that were predicted to encode proteins that are unique to archaea[15]. These genes encompass diverse processes and reflect the genetic diversity of archaea, forming into several clades over evolutionary time, discussed below. Similar comparative analyses of archaeal genomes against eukaryotes gave clear identification that essential components of DNA replication (e.g. GINS, MCM)[16] and transcription (e.g. RNA polymerase sub-units) are conserved between archaea and eukaryotes. DNA replication and aspects of DNA repair in archaea relevant to this thesis are detailed later. In addition, more recent studies have shown widespread well-conserved systems in archaea and eukaryotes that encode proteins required for complex cellular life, most notably and widespread are genes encoding systems for ubiquitination, actin and tubulin cytoskeletons[17], cell division[18] and RNA interference. This is exemplified by the recent identification of Lokiarchaeota[19]. These archaea are found within the TACK superphylum that originally consisted of the Thaumarchaeota, Aigarchaeota, Crenarchaeota and the Korarchaeota but has since been found to encompass more such as the Euryarchaeota (Figure 1.3). Eukaryotes are predicted to have diverged from a complex

cell of archaeal origins that acquired mitochondria *via* horizontal transfer through an early form of phagocytosis following symbiosis[20].

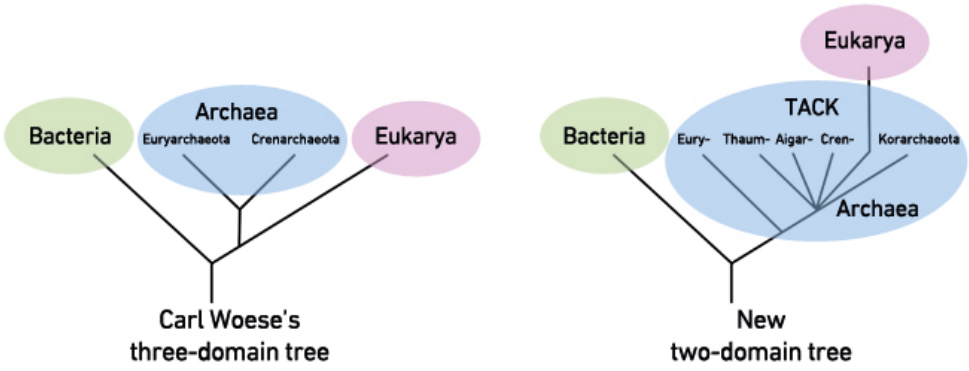


Figure 1.3. Summary of the original three-domains of life hypothesis, compared to the new two-domain hypothesis, arising from identification of lokiarcheota containing eukaryote-like genes. The major difference is that in the two-domains hypothesis eukaryotes arose directly from one clade of archaea. Image courtesy of Thorsten Allers, University of Nottingham.

1.4. Genome storage and packaging: chromatin and euchromatin

The amount of genetic material that is present within a single cell varies hugely between species; the *Escherichia coli* genome is 4.6 Mb of DNA in comparison to *Homo sapiens* genomes of 3.08 Gb. This poses a storage issue, concerning how to package DNA within a single cell in some organisms relative to their size (for example, *E. coli*: 1 μm^3, mammalian cell: 2,000 μm^3). For example in bacteria the contour length of genomic DNA is in the millimeter range, making it 1,000+ times longer than a single cell in the micrometer range. Similar statistics are observed for archaeal and eukaryotic cells. DNA binding by packaging proteins is utilized to overcome this problem, forming chromatin complexes, which also act as a

signals and regulators of fundamental processes such as DNA replication and transcription.

Bacteria lack a nuclear membrane, but genome organization is achieved in a nucleoid in which "Nucleoid associated proteins" (NAPs) effect DNA packaging[21]. In addition to packaging of DNA, NAPs have a profound effect on gene regulation and the chromosome through protein-protein interactions that re-organize DNA packaging into fluid topological domains[22] to allow access to other DNA binding proteins such as transcription factors in response to environmental signals[23]. Archaea also lack a cell nucleus and have archaea-specific NAPs such as Alba[24] that also contribute to genome organization through regulation of DNA processing during DNA replication and repair. In-line with aforementioned similarities between fundamental DNA processing in archaea and eukaryotes (section 1.5), it is interesting to note that many archaeal organisms utilize histone proteins for DNA packaging and regulation, reviewed in Lawrence *et. al.*, 2016[25]. Histone proteins are more usually associated with eukaryotic cells, in which in addition to passive DNA protection and packaging roles, histones are well documented to regulate gene expression[26] and signal DNA damage[27]. However, evidence for regulatory or signaling roles for archaeal histones is currently lacking. Histones from all species contain a signature motif, known as a histone fold (HF) that is required for protein oligomerisation[28]. HFs are distinguished by a topology generated from three α-helix and two short β-strand loops that display amino acids for binding to the DNA backbone, which is wound around histone octamers[29].

Positively charged 'histone tails' are present that overcome repulsion from the DNA backbone on packing. Histone-DNA spooling structures are known as nucleosomes, and are linked by regions of free DNA; this is known as the 'beads on a string' model[30]. Nucleosomes pack together to form a 30 nm fibre[31] that undergoes further folding into euchromatin, that is actively transcribed or densely compacted further into heterochromatin, and ultimately assembles into a chromosome in higher eukaryotes, Figure 1.4. Post-translational modification of eukaryotic histones, especially of the histone tails, is widely observed to regulate DNA binding by histones, part of a broad range of "chromatin re-modeling" activities that occur in all cells. Common histone chemical modifications include phosphorylation, methylation or acetylation of lysine side-chains, having the effect of reducing histone-DNA binding[26]. This can provide access to DNA for recombination, replication and DNA repair proteins, in response to genotoxic stress[32] or as part of a normalized cell cycle.

Chromatin remodeling is also undertaken actively, catalysed by chromatin remodelers that displace nucleosomes, releasing DNA prior to DNA replication and transcription[33]. Chromatin remodelers localize to chromatin *via* physical interaction with nucleosomes and use adenosine triphosphate (ATP) dependent helicase activity to modify chromatin non-covalently. By re-arranging chromatin proteins, the remodelers assemble, dissemble, exchange or space out nucleosomes in response to the aforementioned post-translational chemical modifications of histones.

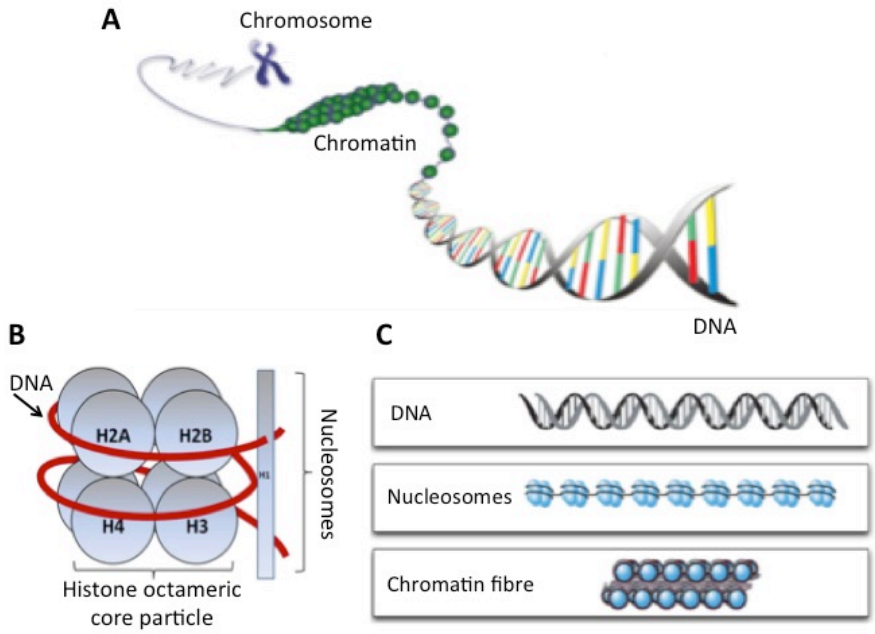


Figure 1.4. Summary of how DNA is packaged in to chromatin in genomes, requiring DNA packaging proteins that include histones forming nucleosomes. Images adapted from Bio-Rad.com.

1.5. DNA replication: Bacteria - Archaea - Eukaryotes

1.5.1 Overview

Cellular (non-viral) DNA replication follows the same basic principles of symmetrical synthesis of nascent DNA strands at a replication fork, catalyzed by a multi-protein replisome machine. The identities of protein components of bacterial and eukaryotic (and archaeal) replisomes are different. However, the major protein structural and enzymatic activities of these proteins are analogous to carry out efficient replication from distinct origins. The assembly and molecular events at the replication fork are probably best understood in bacteria, from early genetic analysis and the ability to fully reconstitute the bacterial (*E. coli*) replisome *in vitro*[34]. Essential components of DNA replication in

eukaryotes were initially identified by genetic analysis in yeasts[35]. Since then a multitude of studies have produced robust models for how eukaryotic replisomes function, including recently the ability to reconstitute eukaryotic replication forks *in vitro*[36,37]. In archaea, conservation of the proteins required for eukaryotic replication suggests that it functions in a similar way, but with reduced complexity[38], as detailed below. As mentioned previously, this difference between bacteria and archaea was one reason leading to the phylogenetic division of bacteria from "archaeabacteria". In the next sections I present a summary of the major parts of DNA replication illustrated in bacteria, archaea and eukaryotes. The topic has had entire collections of review articles and encyclopedia written about it, as an example for further detailed information the interested reader is directed to a recent collection in Cold Spring Harbour Perspectives in Biology: Principles and Concepts of DNA Replication in Bacteria, Archaea and Eukarya, O'Donnell *et. al.* 2013[39].

1.5.2. The process of DNA replication

DNA replication is carried out in a semi-conservative manner[40] in bacteria, archaea and eukaryotes, requiring separation of duplex DNA to release a single stranded DNA (ssDNA) template to be copied, Figure 1.5.

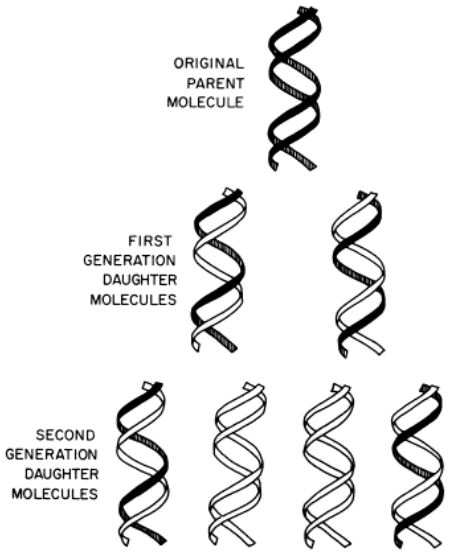


Figure 1.5: Semiconservative DNA replication, as illustrated in the original research paper by Meselson and Stahl[40].

DNA replication is closely linked to the cell cycle, to limit it to once per cycle. For this reason, proteins regulate DNA replication during the synthesis (S) phase of the cell cycle that is common to bacteria and eukaryotes, and presumably archaea.

DNA replication is dependent on the generation of a DNA replication fork that is activated by a multi-protein complex known as the replisome, described in detail in the next section. The replisome comprises DNA processing and DNA binding proteins that interact physically and functionally to replicate DNA, and these proteins must be assembled as a stable complex to resume replication. This ordinarily occurs at sites on the genome known as replication origins in response to specific DNA sequences called origins of replication (*Ori*; "origins"), although replisomes can also be assembled at sites remote from origins, under circumstances where replication forks are broken or blocked away

from origins, as described later. In bacteria, one replication origin is located on the circular chromosome, whereas multiple origins of replication in many species of archaea, typified by *Haloferax volcanii*[41] and *Sulfolobus solfataricus*[42,43], and in eukaryotes are distributed throughout genomes and sometimes under different regulatory control. Under normal circumstances during S phase, replisome components typically assemble at origins, which are typically AT- rich to assist with local melting of duplex DNA that is required to facilitate helicase loading onto single-stranded DNA (ssDNA). Mechanisms of origin firing in bacteria are quite different to archaea and eukaryotes, and even within the bacterial phylum. It should be noted that nomenclature for bacterial DNA replication proteins varies between bacterial species, but for the purposes of this thesis bacterial replication proteins for *E. coli* are described.

In canonical DNA replication initiating from an origin, replication resumes by origin DNA being bound by an origin recognition complex (ORC) in archaea and eukaryotes, or by binding of the DnaA protein in *E. coli*, which functions analogously to ORC. Helicase loading next occurs; in eukaryotes this is a toroidal, heterohexameric MCM 2-7 (for mini-chromosome maintenance) in eukaryotes, and a homohexameric MCM in archaea, reviewed in Bochman and Schwacha, 2009[44]. These proteins encircle DNA strands in readiness for ATP dependent translocation and separation of duplex DNA[45]. MCM 2-7 is loaded onto dsDNA[46] at an origin where it associates with Cdc45 and GINS that assemble to form the CMG helicase complex, in which MCM 3‘ to 5’ translocation is activated[47].

Although there are archaeal homologues/ orthologues for GINS, archaeal sequence homologues for Cdc45 are less clear, but are detectable in genomes as catalytically inactive RecJ homologues in some species[48]. By MCM helicase translocating anti-parallel DNA with 3' to 5' polarity it generates a bidirectional replication fork[46]. Mechanisms for strand separation by MCM remain controversial, but is thought to occur by one of three models: ploughshare, where a β pin is dragged through duplex DNA thereby separating it; steric exclusion, where one of the strands of the duplex DNA is physically excluded from the MCM central channel, and finally a concerted model which included components of steric exclusion and a ploughshare through a pumping action of the hexameric rings, reviewed in Bochman and Schwacha, 2009[44]. In *E. coli*, the replicative helicase is DnaB, which is loaded onto ssDNA *via* interaction with DnaC protein and DnaA located at an origin. Although this scenario in *E. coli* is accepted as the general bacterial model, this is not representative in bacteria overall. Other different ring helicase loading mechanisms exist within the bacterial kingdom that includes the use of helicase co-loaders, for example in low G-C content gram-positive fermicutes such as *Bacillus subtillis*. These are reviewed extensively in Soultanas, 2012[49]. DnaB is structurally and functionally analogous to MCM from archaea, in that it forms a homo-hexameric ring encircling ssDNA. However, interestingly, DnaB translocates with the opposite polarity (5' to 3') to MCM. A more detailed analysis of DNA helicases and their activities is presented later in this thesis.

During unperturbed DNA replication, the *E. coli* replisome has been shown to progress at a rate of 500-1,000 base pairs per second[50]. The major components of the replisome are DNA helicase, polymerase and topoisomerase enzymes, which re-model DNA to achieve synthesis of nascent strands. Replicative DNA helicases MCM (archaea and eukaryotes) or DnaB (bacteria) catalyze ATP-Mg^{2+} dependent unwinding of duplex DNA to expose template DNA for a DNA-dependent RNA polymerase (primase) and subsequently, DNA-dependent DNA polymerases. A primase is required to generate RNA primers from ssDNA after helicase loading has exposed the ssDNA substrate, because replicative DNA polymerases are unable to synthesize DNA *de novo*, whereas RNA can be generated by primases[51]. RNA primers are generated on both strands of melted DNA in the canonical 5' to 3' direction for polymerases, thereby priming both lagging and leading strand DNA synthesis for the initiating replication fork. The leading strand RNA primer is generated once, and is extended into DNA synthesis only, whereas RNA priming reactions on the lagging strand are repeated multiple times to form Okazaki fragments of 1-2 kb in bacteria and 100-200 bp in eukaryotes.

In bacteria, DnaG primase interacts with DnaB as a complex to initiate DNA replication from the RNA primer at an origin in a co-ordinated fashion[52]. In eukaryotes the priming polymerase activity is provided by a multi-subunit Polα-primase, comprising a heterodimer of PriS/PriL, the extending polymerase, Polα, and a regulatory B sub-unit (p70). PriS,

regulated by PriL, generates RNA that is extended by Polα establishing the replication fork. This complex may or may not interact with MCM replicative helicase depending on the organism studied. In archaea the priming reaction for DNA replication is poorly understood, and both DnaG and PriS-PriL are present, but there is no Polα. DnaG physically interacts with MCM[53] and PriSL[54] in at least one archaeal species (e.g. *S. solfataricus*). Archaeal PriS-PriL homologues are associated with a third factor, PriX, which catalyses efficient RNA primer extension, and together are likely to form the primase activity required at archaeal replication origins[55].

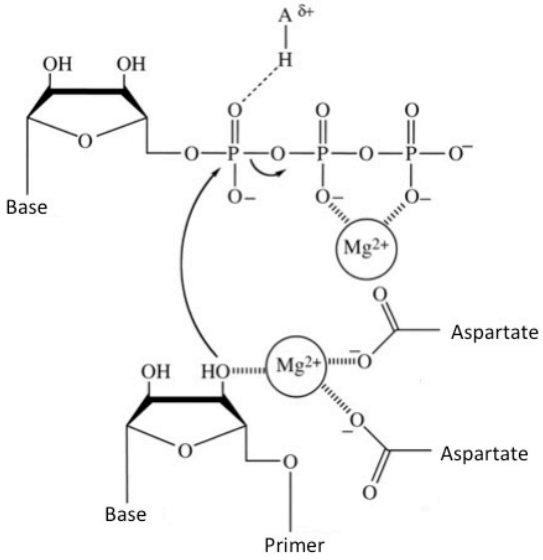


Figure 1.6. Summary reaction chemistry of a generalized primer extension reaction catalysed by a "two-metal ion mechanism" at the active site of a DNA polymerase. Activated by the chemistry of bound magnesium ions, the 3'-OH of a growing DNA strand nucleophilically attacks the alpha phosphate group of a deoxynucleoside triphosphate (dNTP), with resulting bond re-arrangement leading to incorporation of a deoxynucleoside monophosphate (dNMP) into the growing DNA strand. Adapted from Steitz and Steitz, 1993[56].

DNA polymerases are classified into seven families (A, B, C, D, X, Y and RT) of which members of families A-D are associated with stable DNA replication at replisomes in bacteria, archaea and eukaryotes. In bacteria DNA Polymerase III (PolIII), a family C enzyme, generates both leading and lagging strands at the replisome, reviewed in McHenry, 2011[57]. In addition, DNA polymerase I (PolA), a family A polymerase, can synthesize short DNA strands from RNA primers on the lagging strand, although PolA is more usually associated with DNA gap repair at sites of DNA damage. Eukaryotic DNA replication relies on three polymerases from the B-family, Polα-primase, Polδ and Polε, which are primarily associated with DNA synthesis on lagging and leading strands, respectively. These are multi-subunit complexes and are reviewed in detail in Johansson and Dixon, 2013[58], reflecting different types of replicative regulation in the respective species. Replicative polymerase activity in some archaea has complexity reminiscent of eukaryotes, in utilizing PolB, a B-family polymerase and PolD (D-family) at the replisome. There is currently debate over the relative contributions of these polymerases to stable DNA replication, and whether one or both contribute to leading and lagging strand DNA synthesis. Genetic analysis suggests that PolD is essential for replication because this gene is unable to be deleted, whereas PolB can be[59]. However, biochemical evidence has suggested that PolB may be the leading strand polymerase whilst PolD is responsible for lagging strand synthesis[60].

DNA synthesis by polymerases is dependent on the presence of two active-site Mg^{2+} ions, which provide the catalytic chemistry for a "two-metal ion" reaction[56], that is generally applicable to diverse DNA processing enzymes, summarised in Figure 1.6. In this reaction, one metal ion interacts with two acidic amino acid R-groups (usually aspartate) that serves to activate the 3'-OH group of the primer as a nucleophile. Another metal ion interacts with phosphate groups of the nucleoside triphosphate substrate and is thought to stabilize the charged intermediate that forms on nucleophilic attack by the 3'-OH. Recent reviews including Johansson & Dixon, 2013[58] detail advances in understanding the molecular biology of replicative DNA polymerases.

Structural components of the replisome modulate the activities of replisome helicases and polymerases, for example by loading these enzymes onto DNA, and by bridging together various components of the replisome. In eukaryotes and archaea, Proliferating Cell Nuclear Antigen (PCNA), acts as a clamp to tether DNA polymerases to DNA, for loading and conferring processivity to DNA synthesis within the replisome[61]. PCNA is a target for protein modifications that are subsequently responsible for recruiting proteins to the replisome during replication and also during DNA repair at sites of stalled replication, to be discussed later. In bacteria, an analogous polymerase tethering function to PCNA is provided by the β-clamp protein[62].

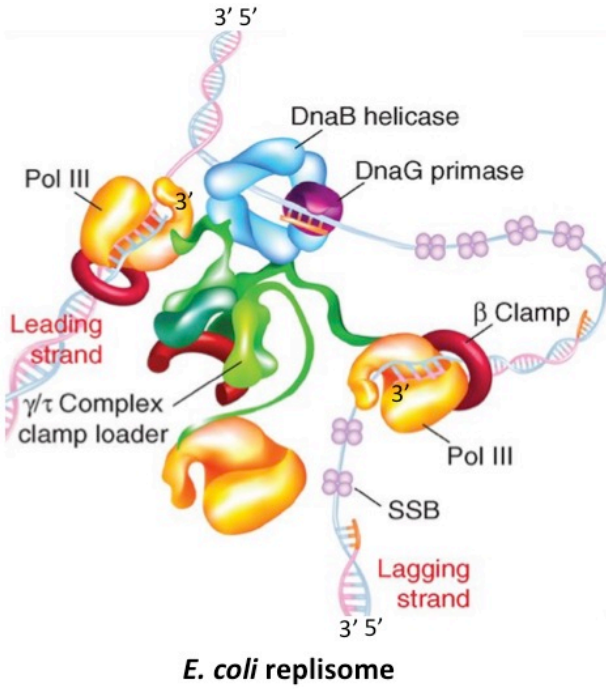


Figure 1.7. Illustration of an *E. coli* replisome highlighting essential proteins required for active DNA replication. Details of the enzymes labeled are given in the main text. Adapted from Yao and O'Donnell, 2010[63].

Other essential replisome components are DNA topoisomerases and ssDNA binding proteins, Replication Protein A (RPA) in archaea and eukaryotes[64,65] and ssDNA binding protein (SSB) in bacteria[66]. Topoisomerase enzymes, TopoIA in bacteria and TopoIA and TopoII in eukaryotes, are required to overcome torsional strain in DNA that arises from its supercoiling as the replicative helicases unwind the double helix ahead of the fork branch-point. These, and other topoisomerases (e.g. TopoV, TopoVI, DNA gyrase and reverse gyrase) form a group of enzymes that are varied in their context and mechanisms of action[67]. However, they are universal in being able to catalyse relaxation of supercoiled DNA

into a less-supercoiled form, which in the context of the replisome allows DNA replication forks to proceed unhindered by "knotted" DNA.

The single stranded DNA binding proteins RPA and SSB are crucial for all aspects of DNA metabolism, and act as DNA protectors and signalers within the replisome. In eukaryotes, ssDNA generated by the replicative helicase is sequestered by heterotrimeric RPA, comprising RPA70, RPA32 and RPA14, denoting approximate molecular weights of each subunit[65]. The RPA complex assembles in a 5' to 3' direction to occlude exposed ssDNA, with binding occuring through the oligonucleotide binding (OB) fold[68], and a winged helix (WH) fold found within the complex and can occur in different modes where affinity to ssDNA and number of bound nucleotides can vary. Archaea also utilize RPA but with reduced complexity, with many species having a single fused protein[69]. The bacterial equivalent of RPA, SSB, acts analogously, and like RPA has important roles in orchestrating protein-protein interactions at the replication fork by stimulating or inhibiting various enzyme activities[66].

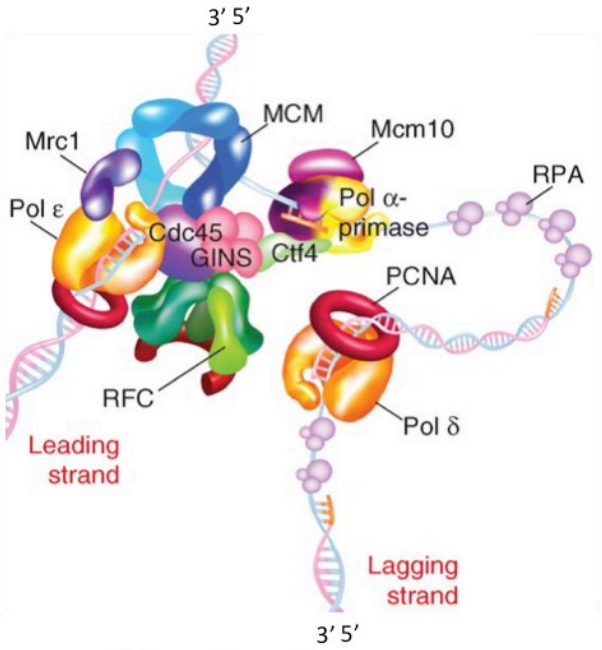


Figure 1.8. Illustration of a replisome from eukaryotes highlighting essential proteins required for active DNA replication. Details of the enzymes labeled are given in the main text. Many of the proteins highlighted here are also conserved in archaea, as detailed in the main text. Adapted from Yao and O'Donnell, 2010[63].

1.6. Accessory helicases promote active replisomes

As discussed in later sections, the chemical damage or alteration of DNA templates at replication forks can lead to stalling of replicative helicase and/or polymerase enzymes, or other DNA processing enzymes because the modified DNA, or protein-DNA complexes cannot be accommodated into enzyme active sites. This blocks forward progression of the replisome unless the obstacle is removed, an action that for protein-DNA complexes may be achieved by "accessory helicases" that co-migrate with replisomes, currently best characterized in the bacterium *E. coli*[70,71]. *E. coli* Rep helicase interacts physically and functionally with DnaB

replicative helicase within the replisome where it translocates along ssDNA opposite to the strand on which the replicative helicase translocates in a 3' to 5' direction displacing protein roadblocks thereby allowing progression of the replisome[34]. UvrD, another 3' to 5' DNA helicase, is also thought to act as an accessory helicase to promote replisome progression, in this instance by physical interaction with blocked RNA polymerase enzyme to facilitate its removal[72,73]. In this respect accessory helicases such as Rep and UvrD act as a "rapid response" first line of DNA repair at the replisome to ensure DNA replication is ongoing. A third bacterial helicase, RecG, may also be involved in processing blocked replisomes, in this case by remodeling the replication fork DNA rather than the associated proteins[74]. These actions of accessory helicases may be preferable over replication fork stalling leading to replication collapse and subsequent processing by recombination, because they are faster and avoid any local genetic re-arrangement that may be associated with recombination, detailed later. Eukaryotes and archaea are also emerging as having candidate helicases with analogous roles to bacterial Rep, UvrD and RecG. Genetic studies in yeast indicate that Pif1[75] and Rrm3[76], both 5' to 3' helicases, are required to overcome engineered replication fork blocks[77], and biochemical analysis of the isolated proteins is consistent with an ability to unwind DNA forks and other substrates that may mimic blocks, such as G4 (quadruplex) DNA[78,79]. I now survey types of DNA damage before

returning to the theme of DNA repair, with particular emphasis on the roles of DNA repair helicases.

1.7. Overview of DNA damage as a source of genome instability

Maintenance of genomic stability underpins productive DNA replication and transcription. DNA is subject to chemical and physical attack from endogenous and exogenous sources and manifest in the form of alkylation, oxidation and covalent crosslinking[80], targeting nucleotide bases and the phosphodiester backbone, Figure 1.9. For example, abasic sites can arise from chemically induced hydrolysis of the glycosidic bond that removes a nitrogenous base from the DNA backbone. Endogenous chemical attack on DNA can occur as a by-product from essential cellular processes most prevalently those that release reactive oxygen species (ROS) and endogenous alkylating agents[81]. Other lesions arise from exogenous sources, such as UV irradiation.

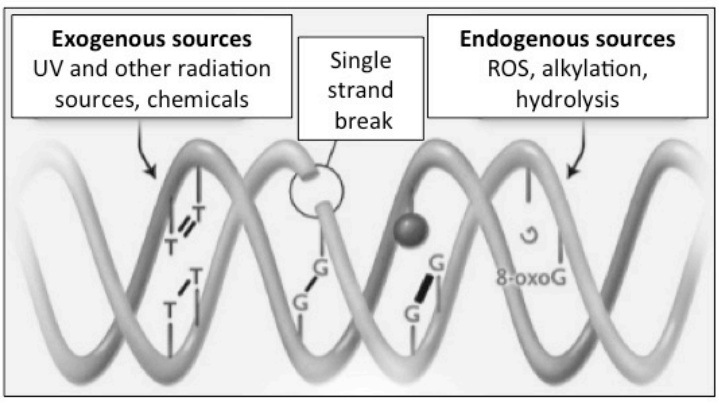


Figure 1.9. Illustration of types of damage inflicted on duplex DNA, and its possible sources, including ultraviolet light (UV) damage, and damage caused by "reactive oxygen species". Adapted from Hoejimakers, 2001[82].

Some chemical conversions of DNA bases can lead to mismatches during DNA replication, such as the formation of 7,8-dihydro-8-oxoguanine by ROS that is able to base pair with adenine as well as cytosine. Following two rounds of DNA replication, this can cause a mutation from G-C to T-A[82]. Each type of damage forms unique DNA lesions for which specific repair mechanisms have evolved for their removal, and are conserved among all domains of life, although to varying degrees of complexity.

Failure to accurately repair DNA lesions can lead to mutation. Although this is perhaps dependent on how extensive the damage and the resulting mutation is, this can lead to cell death. In humans, unresolved DNA lesions can cause mutations that, if located in oncogenes or tumour suppressor genes, can often lead to cancer. For example, mutations carried within *p53* give rise to Li-Fraumeni syndrome that predisposes patients to cancer as a result of inactive p53 that normally assists in cell cycle progression, apoptosis and can sense DNA damage to ultimately prevent tumourigenesis[83]. Mutations can arise from the different repair methods used to repair DNA lesions, perhaps counter intuitively. This is a result of different repair pathways that are employed at different stages of the cell cycle. Alternatively, it can be beneficial to employ a low fidelity pathway in which mutation may arise in favour of restoring DNA to an intact form quickly, or to prevent from further, more extensive damage and therefore mutation arising from gross chromosomal rearrangement.

DNA damage manifests in a variety of forms, resulting from multiple chemical reactivities of DNA bases and of the phosphodiseter backbone. Consequently, processes that repair DNA are diverse, dependent on the precise chemistry of the DNA damage encountered and its general disposition within a DNA molecule, for example if it is present at an active replication fork or elsewhere. Numerous lesion-specific DNA repair mechanisms have evolved. These include nucleotide excision repair (NER) and base excision repair (BER), which remove tracts of DNA (NER) or DNA bases (BER) that contain chemically damaged DNA[84]. The most toxic DNA lesions are double strand breaks (DSBs) to DNA strands that result in collapse of replication forks or transcription complexes and generate mutagenic DNA ends. Homologous recombination and non-homologous end joining (NHEJ) are effective for repair of DNA DSBs, dependent on the context in which the DSB is encountered, influenced by the availability of homologous DNA according to the stage of the cell cycle. However, as we will see in later sections, HR repair of DNA can take on several guises generating specific outcomes.

DNA lesions caused by chemical means are not the only challenges to genome stability. DNA replication is also perturbed by protein complexes bound to DNA that form physical roadblocks to replication. These are often dealt with by replisome accessory helicases such as the aforementioned *E. coli* Rep and UvrD that physically displace the protein from DNA, allowing replisome progression[34,73]. RNA-loops (R-loops), in which an RNA strand is base-paired to a complementary DNA strand from

a duplex, thereby displacing a ssDNA loop, also provoke genome instability in a number of ways. They are natural intermediates in cellular processes such as transcription, in which a growing RNA chain invades the proximal DNA duplex[85], and clustered regularly interspaced short palindromic repeats (CRISPR) immunity, during interference reactions[86]. However, if R-loops are allowed to persist beyond their usefulness then this can lead to inappropriate recombination or RNA primed replication arising from the R-loop. In the latter, it has been suggested that R-loop primed origin independent DNA replication may utilise error prone (translesion) polymerases, provoking mutagenesis. It is also possible that nucleoprotein complexes that generate R-loops themselves may present a physical block to the progressing replisome, as RNA polymerase or CRISPR interference complexes[86]. Similarly as for repairing physical DNA lesions, there are numerous helicases and nucleases that specifically target R-loops to dissociate them or degrade RNA, thus restoring the integrity of the DNA duplex and removing a potential block to replication. These systems and enzymes are reviewed in Santos-Pereira and Aguilera, 2015[87].

The importance of the role of DNA repair pathways in human health is reflected in the severity of several genetic syndromes whose phenotypes arise as a result of improper DNA repair. These include Fanconi Anaemia (FA)[88], Bloom's Syndrome and Werner's syndrome[89] that predispose patients to cancers. These disorders arise from mutations carried in genes encoding DNA repair machineries such as the Fanconi anaemia factors, Bloom helicase and Werner helicase respectively, and

can subsequently predispose patients to cancers[90] and or mental retardation as a result of an inability to accurately repair DNA lesions. Inadequate DNA repair is therefore associated with the onset of mutations in genes that maintain genome stability, exemplified by the prevalence of cancers (estimated at 40%) associated with unrepaired damage to genes encoding the tumour suppressor p53[83].

1.8. Chromatin as part of the DNA damage response

As described in section 1.4, cellular DNA is packaged in the form of heterochromatin or euchromatin. At active replication forks or transcription complexes DNA is already subject to controlled denuding so that replicative enzymes can gain access to ssDNA. In the event that DNA damage that is not associated with these processes, DNA repair can only be achieved through signaling and chromatin re-arrangements that facilitate access of repair enzymes to DNA damage; DNA must be released from chromatin to facilitate repair. This problem was noted by Smerdon who proposed the general 'access-repair-restore' model following detection of a lesion[91].

The chemical and physical insult inflicted on DNA as described in section 1.7 can manifest in ssDNA or double stranded DNA (dsDNA). Typical examples that lead to characterised DNA damage responses are physical roadblocks to replication that generate aberrant ssDNA tracts when uncoupling occurs between the replicative helicase and polymerase[92], and the generation of ssDNA or dsDNA breaks, by free

radicals and other reactive products of cellular metabolism. These two types of DNA damage elicit a response that begins with recruitment of sensors and mediator proteins, followed by recruitment of transducer proteins, two very well studied being the kinase enzymes ataxia telangiectasia-mutated (ATM), and ATM and Rad3 related protein (ATR). ATM and ATR are phosphoinositide-3-kinase related protein kinases that are major contributors to the DNA damage response by coordinating all aspects of DNA processing activities within the cell. ATR responds to aberrant ssDNA during origin firing and processing of stalled replication forks in the event of replication stress, and ATM primarily co-ordinates DNA repair at dsDNA breaks (DSBs)[27].

On exposure of ssDNA, for example in the event of damage, single stranded binding proteins coat DNA[93]. In humans, this occurs by the heterotrimeric RPA complex that is described in section 1.5.2. RPA-DNA binding prevents secondary structure formation and inappropriate DNA processing by nucleases. The RPA complex also contains protein-protein interaction modules that participate in signaling, therefore due to this, and the affinity of RPA for ssDNA that can occur following DNA damage, RPA is crucial to DNA damage responses. ATR is recruited to ssDNA that is unusually abundant when replication forks suffer uncoupling of DNA helicase and polymerase activities, such that the helicase continues to unwind duplex DNA ahead of the fork but polymerases do not synthesise new DNA[94]. ATR can also interact with RPA, mediated by ATR Interacting Protein (ATRIP)[95]. ATR is activated for DNA binding by further protein

complexes; TOPBP1, Rad9-Hus1-Rad1, Mre11-Rad50-Nbs1[96], RHINO and ETAA-1[97]. These proteins also localize to sites of stalled replication, but will not be discussed in detail here. Activated ATR phosphorylates the H2AX histone to γH2AX, a modification known to primarily flag damage at sites of DSBs by ATM. However, γH2AX has also been observed at stalled replication forks during early stages prior to fork collapse into DSBs. Phosphorylation of H2AX spreads along chromatin, surrounding the site of damage[98], recruiting DNA repair proteins. Subsequent histone modifications by ATR can occur.

ATR also targets components of the replication machinery in response to replication fork stalling, in an effort to restart replication or to stabilize the fork to prevent collapse and formation of potentially toxic DSBs. ATR dependent phosphorylation of recombination proteins has also been reported, which would provide a higher level of control for recombination mediated replication restart, however these are not well understood[32]. Interaction of ATR with CHK1, a DNA replication checkpoint kinase, leads to CHK1 activation that eventually triggers cell cycle arrest *via* a phosphorylation cascade, thus co-ordinating cell division events with responses to stalled replication.

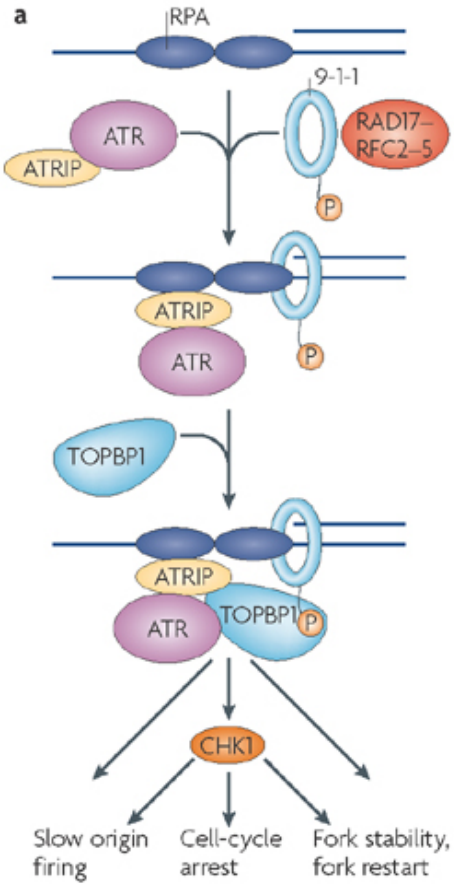


Figure 1.10 DNA damage signaling events by ATR at ssDNA

ATR interacts with RPA that coats ssDNA *via* ATRIP. This subsequently recruits DNA damage response molecules *via* phosphorylation that trigger a response to facilitate repair such as slowing of origin firing, cell cycle arrest and replication fork stabilization and restart. Adapted from Cimprich and Cortez, 2008[99].

ATM is a protein kinase that responds to dsDNA breaks, that when disrupted in humans leads to the genome instability syndrome ataxia-telangiectasia. Although the precise mechanism for how ATM responds to dsDNA breaks is not understood it does require functioning Mre11-Rad50-Nbs1, with there being physical interaction between ATM and Nbs1 proteins in some systems. ATM also interacts in multi-protein complexes

with other DNA repair effectors, including MDC1, which seems to be crucial for activation of ATM kinase activity. ATM acts as at the hub of a phosphorylation cascade, altering functionality of other proteins, including several other protein kinases that then act on their own phosphorylation substrates. Interested readers are directed to a detailed review by Shiloh and Ziv, 2013[100]. Coordination of the DNA damage response by ATR and ATM allows for DNA damage proteins to localize onto chromatin to enable DNA repair by one or more processes dependent on the DNA lesion. In addition to the proteins described above, DNA helicases are targets for post-translational modifications to confer higher levels of control for DNA repair.

1.9. DNA helicase enzymes

DNA helicases are utilized during DNA replication, repair and during chromatin remodeling. DNA helicases are defined as enzymes that utilize energy released by hydrolysis of nucleoside triphosphates (typically ATP) to separate duplex DNA or RNA. Helicases may track or translocate along one strand of the duplex with directionality (3' to 5' or 5' to 3') or may utilize pumping mechanisms without ssDNA directionality, to traverse the duplex backbone whilst separating strands[101]. Helicases are characterized by ATP-binding and hydrolysis within RecA-like domains, which form a "helicase core" activity along with a DNA interacting domain that directs translocation directionality. "Accessory domains" to the core helicase domains give helicases additional functionality, such as substrate

specificity to unwind specialized DNA or RNA structures (e.g. quadruplexes or displacement loops, D-loops) or to give protein-protein interaction surfaces. The motive power generated by translocation of a helicase can also be utilized to displace bound proteins that may otherwise impede DNA processing, such as the previously described role of bacterial accessory helicases Rep and UvrD, section 1.6. Based on amino acid sequence motif similarities, and conserved helicase functions, helicases are currently categorized into major superfamilies 1 and 2 (SF1 and SF2), and three smaller superfamilies 3-5. The major AAA^{+} superfamily of proteins also includes some DNA helicases, as a distinct sub-group of helicases that are usually functional as multimeric ring structures, for example MCM[45]. The figure below illustrates general roles of DNA helicases, especially applicable to DNA repair processes.

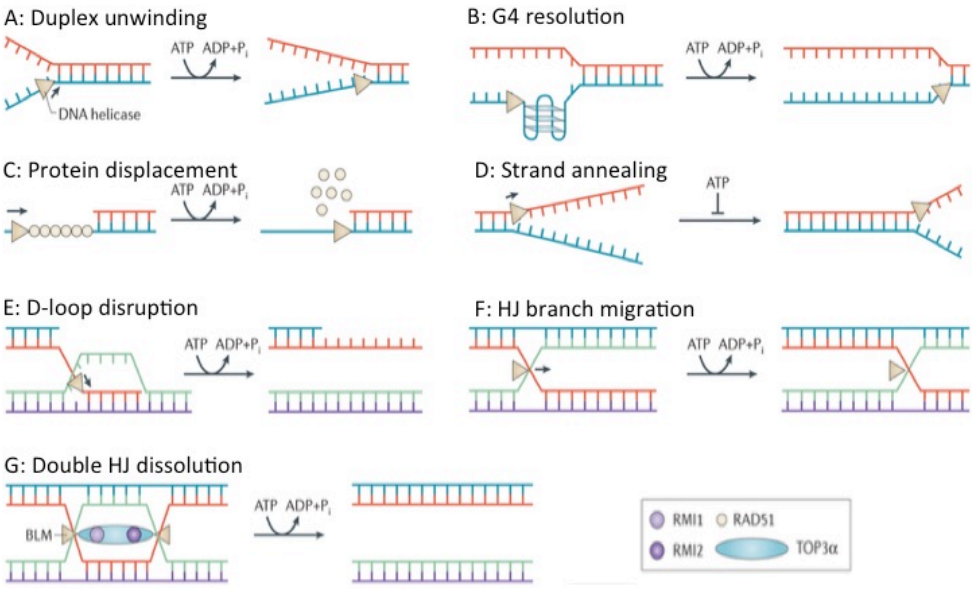


Figure 1.11. A summary of DNA processing reactions catalysed by DNA helicase enzymes. Adapted from Brosh, 2013[102].

1.10. Roles of helicases during DNA repair

Helicases are critical for manipulation of DNA during repair processes, particularly during processing of stalled replication forks and HR where they are required to remodel complex branched DNA substrates to release ssDNA to promote replication, transcription or repair[103]. For example, the RecQ helicase family has homologs in bacteria and eukaryotes and are involved in unwinding of G-quadruplex secondary structures that block replication fork progression, D-loops and DNA substrates at stalled replication forks[104]. RecQ family helicases BLM (*H. sapiens*) and Sgs1 (*S. cerevisiae*) also catalyse branch migration[105] and dissolve Holliday Junctions (HJs) as part of larger protein complexes to ensure formation of non-crossover products. In the context of some repair processes, these may help avoid loss of heterozygosity, thereby promoting genetic diversity.

UvrD and Srs2, part of the UvrD helicase family and found in *E. coli* and *S. cerevisiae* respectively, act to displace RecA and Rad51 filaments to prevent D-loop formation hence promoting synthesis dependent strand annealing (SDSA), an alternative route of repair that forms non-crossover products, detailed in section 1.22[106,107]. Mph1, the yeast homolog of human FANCM, is able to disrupt Rad51 catalysed D-loops to drive recombination pathways towards SDSA[108].

Pif1 helicase has homologs in bacteria, yeast and humans[109]. In particular, Pif1 shares unique motifs with bacterial RecD. In yeast, Pif1 unwinds G4 structures, migrates D-loops and functions with polymerase δ

during lagging strand synthesis and actively displaces telomerase from telomeres to promote replication[75]. Rrm3 helicase, a homolog of Pif1 found in yeast, facilitates replication fork progression by numerous mechanisms including removal of proteins from DNA, enhancing replication through chromosomal sites and unwinding of DNA-RNA structures[79]. Interactions with PCNA support a role for Rrm3 as an accessory helicase for replication fork progression.

1.11. Superfamily 2 DNA helicases

Helicases couple free energy released on ATP hydrolysis to drive conformational change within the protein to separate regions of duplex DNA into ssDNA that is a substrate for further processing during replication, transcription or repair. Some helicases, such as the Ski-2 family, are also active on RNA[110]. Helicases can be divided into six super-families according to their direction of translocation, substrate, motifs and the biological process they facilitate. The focus of this thesis is on Hel308 and HelQ DNA helicases that are classed as Superfamily 2 (SF2) helicases. SF2 is the largest class of helicases and is further subdivided, although only two divisions will be focused on here: Ski-2 like helicases and those that contain a DExH/D box motif, both features that Hel308 and HelQ share. SF2 helicases possess 9 motifs that are characteristic of ATPases: Q, Ia, Ib, II, III, IV, V and VI. These motifs assemble at the interface of the RecA like folds forming a cleft, where ATP is bound and hydrolysed. These domains are in contact with DNA and so this region acts as a motor

domain. Motifs I and II are also referred to as Walker A (-G-xxxx-GK-TS) and Walker B motifs, the latter containing the sequence "DEAD/DEAH" or close variants. Motif I/Walker A is required for ATP binding and Motif II DEAD/DEAH is required for ATP hydrolysis through co-ordination of metal ions.

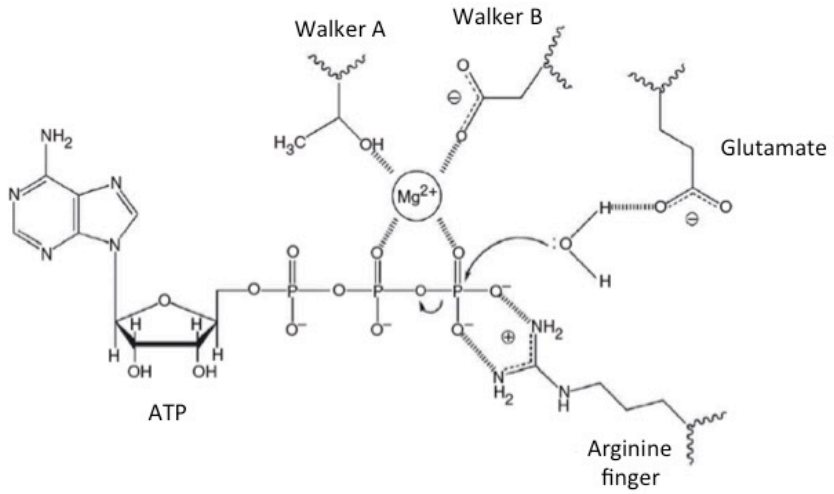


Figure 1.12: Diagram summarizing chemical reactions of ATP hydrolysis during helicase activity, requiring a metal co-factor and protein motifs that include Walker A, Walker B and an arginine finger. In this reaction a water molecule acts as a nucleophile that attacks the γ phosphate of ATP that is also bound by a metal ion. Conformational movement of the protein can accompany resulting ATP hydrolysis to drive helicase/translocase activity. Image courtesy of Edward Bolt, University of Nottingham.

HelQ and Hel308, the major subject of this PhD thesis, belong to superfamily 2 (SF2). SF2 helicases are typically monomeric, in comparison to helicases that function as oligomers such as the SF6 eukaryotic replicative heterohexameric ring helicase MCM 2-7. Ski-2 like RNA helicases are typical of SF2 helicases. Interestingly, Hel308 shows

similarities to the RNA processing Ski-2 like family. The current model for SF2 helicase translocation is based upon that of NS3 in which ssDNA runs along the top and through the RecA fold interface stabilized by residues which act as 'bookends' on the phosphodiester backbone[111]. These residues move upon ATP binding and hydrolysis at the RecA domain interface causing translocation along ssDNA in a ratchet type mechanism. The features used to classify Hel308 into SF2 lie within the core helicase domains that are characteristic of many helicases. However, Hel308 has three additional accessory domains that enable it to carry out its DNA processing functions by mechanisms that are still being elucidated. The later sections of the introduction provide a full review of Hel308 and HelQ helicases, and their putative roles in repair of DNA replication that is particularly linked to DNA damage in the form of ICLs. In the next sections I describe DNA repair mechanisms that utilize these and other DNA helicases as major components of their respective repair pathways.

1.12. DNA repair mechanisms: Nucleotide excision repair

Initially identified in bacteria, nucleotide excision repair (NER) systems repair damaged nucleotides that manifest as bulky adducts that destabilize duplex DNA[84]. Typical examples of damage tackled by NER are cyclobutane pyrimidine dimers and 6-4 photoproducts that result from ultraviolet (UV) irradiation (100-400 nm), and base mismatches that result from chemical damage brought on by genotoxic agents. Repair is achieved step-wise by damage recognition, excision of a nucleotide and adjacent

oligonucleotides followed by synthesis of nascent DNA across the gap. In humans, mutations in genes encoding NER proteins cause syndromes, such as xeroderma pigmentosum that predisposes patients to skin carcinomas. NER is an example of DNA repair that is described as global because it is not linked to any particular stage of the cell cycle or DNA processing event, but is activated at any stage in response to chemically damaged DNA.

In *E. coli* NER is carried out by the UvrABC complex in which dimeric UvrA detects damage *via* indirect readout by testing DNA flexibility to detect distortions in the DNA backbone[112,113]. This method of detection means NER can remove an array of different DNA lesions. DNA binding specificity of UvrB for disrupted DNA backbone structures verifies the damage and recruits UvrC, nuclease, that cleaves at specific sites 3' and 5' either side of the site of damage[114]. UvrD, a helicase that also contributes to other forms of repair, clears the fragment using a 3' to 5' direction of ssDNA translocation, thereby allowing polymerase access to the gap for synthesis of nascent DNA that is ligated to the existing strands by DNA ligase[114]. Essentially this process is the same in archaea and eukaryotes, however in typical eukaryotic fashion, at a level of greater complexity in which more proteins engage in the process, imposing a greater level of control[115,116].

Proteins that participate in eukaryotic NER show little sequence homology to those involved in bacterial NER. Some archaea possess homologues of both UvrABC and XPD systems, but the extent of their

functionality is not clear. XPD helicase has 5'-3' polarity and is a SF2 member. It is required to open tracts of damaged DNA during NER to facilitate repair of chemically damaged DNA bases. Two separate protein complexes carry out cleavage of the exposed DNA backbone adjacent to the damage: XPF/ ERCC1 and XPG nucleases[117]. These carry out, respectively, the 5' and 3' incisions and independently of one another. The analogous situation is simpler in bacteria, where UvrC executes both incisions. XPA helicase is then responsible for removing the cleaved fragment of damaged DNA, in a process analogous to that carried out in bacteria by UvrD. Following exposure of ssDNA, PCNA clamp is recruited with Polymerases δ and ε that synthesise nascent DNA from the 3'-OH released following cleavage by XPF-ERCC1 before ligation. The helicase activity of XPD is essential for eukaryotic NER, but is dispensable for its role within the transcription initiation complex TFIIH. In humans, mutations in XPD give rise to the diseases xeroderma pigmentosum (XP), trichothiodystrophy and Cockayne's syndrome.

1.13. DNA repair mechanisms: Interstrand crosslink repair

Covalent crosslinks that occur between bases in DNA are toxic lesions which most commonly occur following exposure to endogenous byproducts of metabolism, crosslinking drugs typically used for cancer treatment, such as cisplatin or mitomycin C (MMC) or by exposure to radiation such as UV which causes formation of pyrimidine dimers that distort the DNA backbone. Crosslinks can occur between two stands of

DNA or between adjacent bases on the same strand, known as interstrand and intrastrand crosslinks respectively. Both types of crosslink block macromolecular machineries that carry out DNA replication and transcription. Intrastrand crosslinks occur more frequently than interstrand crosslinks and prevent replication by blocking DNA polymerases. To overcome this, translesion polymerases are employed, however, these are typically error prone in comparison to replicative polymerases and as a result lead to mutagenesis, discussed below. In comparison, interstrand crosslinks (ICLs) that covalently connect the two strands of duplex DNA normally held together by collectively strong but separable hydrogen bonds, prevent strand separation by DNA helicases thereby blocking progression of the replisome resulting in replication fork stalling and collapse. Crosslinks also affect progression of transcription machinery in a similar manner. As a result, specialized repair machinery must be recruited to the site of damage for ICL removal and replication fork recovery.

There is no single identified pathway for specific repair of ICLs. ICL repair tends to involve interplay between identifiable components usually associated with other major DNA repair pathways[118]. For instance, removal of ICLs in *E. coli* is achieved by cooperation between different pathways as identified in models in 1973[119]; In this scenario, nucleotide excision repair (NER) occurs to initially 'unhook' the lesion from one strand and then excise it from the second strand, followed by homologous recombination (HR) priming of DNA replication to fill in the gap. This

bacterial recombination dependent pathway of repair can only occur following DNA replication when a sister chromatid is available for homology searching during HR, detailed later. In metazoans in the FA pathway of ICL repair, enzymatic activities for nucleotide excision, base excision, DNA replication and general nucleases are needed to complete repair, detailed in the next section.

In both of these instances in bacteria and metazoans, and for DNA repair processes more generally, the repair mechanisms are dependent on the type of the lesion, and on its context, for example whether the repair process is physically or functionally coupled to DNA replication or gene transcription. Recombination-independent pathways involving translesion synthesis (TLS) following NER is also known to occur in *E. coli* when DNA replication has not yet taken place.

1.14 ICL repair in humans: The Fanconi Anaemia pathway

Although not the only route of ICL repair in humans, the Fanconi Anaemia (FA) pathway is a major repair mechanism activated upon detection of ICLs, which drives towards HR[120]. While relatively well studied, due to the complex nature of the FA pathway, major questions remain to be answered with regard to this mechanism of repair. Understanding of the FA pathway initially began by identification of several different complementation groups from FA patients by Guido Fanconi. 16 complementation groups (FANCA-FANCQ) have been

identified to date, in which each is distinguished from the others by the presence of mutations in a different gene required in the pathway. FA patients are predisposed to development of cancers, typically squamous cell carcinomas of the head and neck.

The FA pathway is typically activated upon detection of ICLs *via* replication fork stalling. Proteins involved in this pathway can be grouped into major complexes, each of which has been identified to carry out independent roles. The FA core complex, composed of FANCA-M, possesses E3 ligase activity (FANCL) that activates the FANCD2-FANCI heterodimer by ubiquitination of FANCD2 upon detection of ICLs *via* recognition of stalled replication forks by FANCM. FANCM has been identified as the human homolog of the archaeal Hef protein which possesses both endonuclease and helicase activity[121]. Hef generates DSBs for repair by HR from stalled replication fork substrates. FANCM is able to migrate branched DNA intermediates including HJs (described below), in addition to replication fork reversal for resetting by HR[121,122]. Once activated, the FANCD2-FANCI heterodimer localizes to chromatin adjacent to sites of damage. FANCD2 is known also to localize with proteins essential for repair by recombination such as Rad51 and BRCA2 (FANCD1). The FA pathway has close links to BRCA proteins in which BRCA2 is important for formation of the Rad51 nucleoprotein filament (NPF). Unhooking of the ICL by nucleases still occurs in the FA pathway prior to HR as described by the simplistic Cole model[119]. However, numerous nucleases have been linked to this process. Deans and West propose that

this is due to different ICL structures as a result of various types of damage, which therefore require different proteins for removal[80].

1.15. DNA repair mechanisms: Base excision repair

Nucleotide bases can be modified by exposure to chemical agents or upon irradiation. Damaged bases are removed during base excision repair (BER). Chemical modifications to DNA bases include methylation and deamination that for example can convert from cytosine to uracil that becomes thymine at replication, and these typically occur at particular sites on the molecule prone to attack. Chemical modifications to DNA may also occur on exposure to free radicals. These give rise to oxidized bases, the most prevalent being 8-oxoGuanine that is mutagenic, and they cause DNA strand nicks that can collapse replication and transcription if left unattended. A-purinic (AP) sites are formed on hydrolysis of the N-glycosidic bond leading to loss of purines. These lesions can lead to mismatching of base pairs during DNA replication, or even blockage of replication or transcription machinery[123]. BER differs from NER as suggested in its name, in which rather than excising whole oligonucleotide fragment containing the DNA lesion *via* cleavage of the DNA phosphodiester backbone (unless for repair of AP sites), BER functions by hydrolyzing the N-glycosidic bond between the deoxyribose moiety and the purine or pyrimidine molecule, catalyzed by DNA glycosylases, creating an AP site[124,125].

Unlike NER damage identifying enzymes that detect damage in an indirect manner, BER utilizes a repertoire of DNA glycosylases that are used for specific recognition of different lesions[124].

1.16. Lesion bypass by translesion synthesis and template switching

Translesion synthesis (TLS) is an error prone method of tolerating DNA lesions that promotes continuation of DNA replication across sites of damage, known as lesion bypass[126,127]. Eukaryotic TLS is carried out by any of five known DNA polymerases, Rev1, Pol ζ, Pol κ, Pol η, and Pol ι, reviewed in Waters *et. al.*, 2009[128]. Typically TLS operates by actions of two polymerases in concert, one polymerase inserting nucleotides opposite a lesion and a second polymerase molecule extends this nucleotide from the 3'-OH beyond the damage, enabling lesion bypass.

The Y-family of polymerases is the largest group comprising TLS polymerases, and these differ slightly in structure to typical replicative polymerases, enabling lesion bypass. Although TLS polymerases adopt the typical polymerase 'hand-like' structure consisting of palm, hand and thumb domains, they also feature an additional domain referred to as the 'little finger' that is unique to Y-family polymerases[129,130]. Ultimately, TLS polymerases have a larger active site in comparison to replicative polymerases in order to accommodate often bulky adducts on DNA such as cyclobutane pyrimidine dimers. Although TLS polymerases can incorporate nucleotides to bypass DNA lesions, they cannot read damaged bases to correctly incorporate the appropriate nucleotide opposite the

damage during synthesis. This leads to elevated levels of mis-incorporation of nucleotides, which is the basis of the error prone nature of TLS. However, some TLS polymerases are capable of replicating DNA at particular lesions in order to incorporate correct bases accurately, particularly when thymidine dimers are encountered in the case of Pol η for example[131]. TLS polymerase active sites may be tolerant to diverse types of DNA damage whilst remaining active for DNA replication, probably reflecting the diversity of chemical damage that can arise on DNA molecules. In some instances a TLS polymerase has strong preference for a specific type of chemical modification, however, suggesting evolution in response to prevalent or chronic types of damage.

TLS polymerases are recruited to a stalled replication fork *via* PCNA in eukaryotes[132]. TLS polymerases act in concert with replicative polymerases, when on blockage of a replicative polymerase, a TLS polymerase is recruited where it replicates past the lesion before being replaced by the original replicative polymerase. However, it has been suggested that TLS polymerases are required to synthesize DNA opposite DNA lesions across gaps left by the replisome post replicatively. Recruitment of TLS polymerases to synthesise DNA where the usual replicative polymerases are unable, could allow us to suggest that leading strand synthesis is dis-continuous. Recruitment of error prone TLS polymerases is carefully controlled to ensure they are only utilized for lesion bypass only, and not misused for replication of undamaged DNA. TLS polymerases have inherent properties to protect against this; for

example, they are deliberately less processive as a mechanism to prevent them form carrying out extensive error prone DNA replication, in favour of ensuring high fidelity replicative polymerases take over.

Lesion bypass can also occur by template switching, which unlike TLS, is almost error free. Template switching involves copying the DNA sequence from the intact homologous region, rather then the original damaged strand. This is achieved by extensive remodeling of a stalled replication fork[133]. For example, this may occur following fork regression in which the replication fork is effectively reversed into a structure reminiscent of a Holliday Junction. Alternatively, this process is carried out similarly to recombination, in which the ssDNA template invades a homologous duplex that acts as a new template. These pathways are influenced by protein modifications of PCNA.

The DNA repair pathways listed above, apart from HR that may follow ICL repair, are known as global DNA repair pathways. This is in reference to how these pathways can be utilized to remove DNA lesions at any stage of the cell cycle. However, initiation of some repair pathways can only be triggered during particular stages of the cell cycle. These are primarily the recombination dependent repair (RDR) pathways that require homologous DNA from a sister chromatid to copy, thereby often facilitating high fidelity DNA repair, in comparison to the global pathways that are often less accurate as a result of employing translesion synthesis following lesion removal. Due to the requirement of a template from a sister chromatid, by definition, these pathways can only occur during S

phase during or after DNA replication. If DNA lesions fail to be repaired by S phase, it is highly likely that replication forks will encounter the offending DNA causing them to stall and potentially collapse.

1.17. Stalled replication forks

It is possible during synthesis phase (S-phase) of the cell cycle that the replication fork might collide with unrepaired DNA lesions. As alluded to previously, the replication machinery is highly vulnerable to DNA damage because the active sites of replisome components cannot accommodate bulky DNA lesions on the bases or the backbone. This is often less of a problem for replicative helicases that indiscriminately tear apart duplex DNA *via* various mechanisms, however replicative DNA polymerases have small active sites that have evolved to accommodate native bases to carry out high fidelity DNA replication. This is discussed comparatively with the low fidelity translesion polymerases that can traverse DNA lesions on one strand, in the event of replication stress in section 1.16. It must be noted that major DNA lesions such as covalent interstrand crosslinks pose a physical block even for replicative DNA helicases. As a result, DNA lesions and protein complexes impede progression of the replisome and can lead to replication fork stalling, Figure 1.13, and if not stabilized, subsequent collapse[134,135].

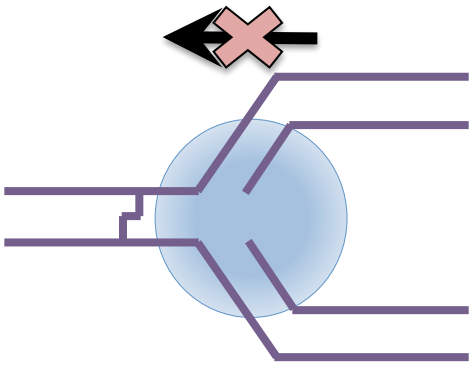


Figure 1.13: Schematic showing an inter-strand cross-link blocking the replisome (blue circle) ahead of the replication fork branch point.

However, although typically considered dangerous, in some cases stalled replication forks are 'programmed' to occur at specific sites, for example mechanisms by which bacterial DNA terminates replication. Bacterial *ter* sites are numerous sequential DNA sequences located on the bacterial chromosome approximately opposite origins of replication on circular chromosomes, and which "capture" the advancing replisome at the terminus of bi-directional DNA replication. A Tus protein binding to *ter* with high affinity can create a protein-DNA roadblock that impedes replisome progression, triggering replication fork stalling, and thereby replication termination[136,137]. This Tus protein-based replication termination may also be controlled by relative activities of other DNA repair factors that prevent over-replication (e.g. PriA and RecG helicases) and by specific DNA structure or sequence elements that discourage re-activation of a replisome at *ter* sites.

In the event of replication fork stalling during replication stress, it is vital that replication forks are stabilized and reset to prevent further

damage in order to allow DNA replication to resume[133]. For example, unattended replication forks can collapse, into highly toxic double strand breaks. Stalled replication forks can be recovered and replication resumed by means of recombination dependent restart, or in some organisms, by direct replication fork restart.

1.17.1. Replication fork restart and origin-independent DNA replication

It is vital that sites of stalled replication are rescued quickly: if stalled forks remain unattended then fork collapse is likely to ensue that can result in further DNA damage such as highly toxic DSBs. Recently, proteins such as Mre11 typically associated with initial resection stages of DSBs, have been found to interact at sites of stalled replication. The state of stalled replication forks upon arrest is likely to be different according to the context of the DNA lesion(s) that has provoked stalling. For this reason, replication forks must often undergo extensive remodeling prior to repair and subsequent replication restart[133]. DNA remodeling activities are carried out by DNA helicases that release ssDNA for loading of proteins required for restart of replication. This occurs during a phenomenon known as direct replication fork restart.

Direct replication fork restart has been observed in bacteria. Genetic studies in *E. coli* and *Bacillus subtillis* implicate PriA in direct replication fork restart by facilitating replisome reloading following unwinding of the lagging strand[138]. PriA is a SF2 3’ to 5’ DNA helicase that targets branched DNA structures and recognizes the 3’-OH of the nascent

DNA *via* a 3'-OH binding pocket, thereby orientating it for assembly of replication components at the parental duplex[139,140]. This, and interaction with SSB at branch sites, ensures PriA does not unwind the parental duplex.

PriA cannot achieve replisome reloading alone, and interacts with different proteins in the pathway depending on the type of branched stalled fork structure that is present: PriA-PriC, PriA-PriB, PriC-Rep[141]. PriA remodels the stalled fork *via* clearance of the lagging strand to enable loading of replicative helicases DnaB onto ssDNA, Figure 1.14, mediated by the protein complexes listed above. However, in the case of PriC-Rep, Rep helicase acts in place of PriA[141]. Ultimately, PriA facilitates assembly of the primosome that allows replication to resume in an origin independent manner *via* remodelling of a stalled replication fork. However, PriA is known to facilitate repriming of replication at branched structures known as at displacement-loops (D-loops) hand in hand with PriB and DnaT, again, independently of an origin of replication[142]. This is an attractive mechanism of replication fork restart in all organisms, however, there are no sequence homologues of PriA in eukaryotes and a functional orthologue is yet to be found.

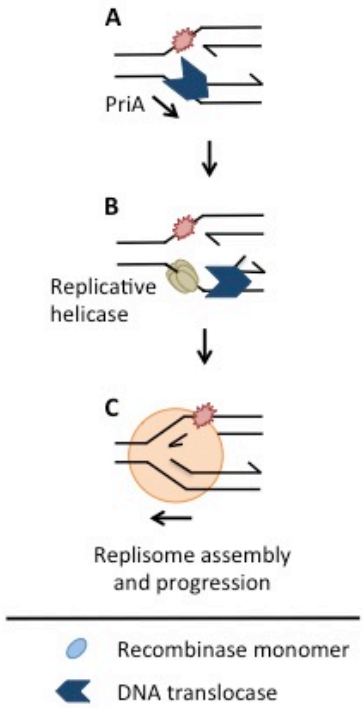


Figure 1.14. PriA helicase in bacteria can re-activate stalled replication at sites remote from replication origin sequences by remodeling a forked DNA structure to promote loading of DnaB replicative helicase, resuming replication. Adapted from Northall *et. al.*, 2016[143].

ATR, a DNA damage response molecule is known to target replication machinery at sites of stalled replication. Likewise, Rad51 paralogues, typically required during homologous recombination, have also been linked to replication fork stabilization and possible replication restart[144], in addition to roles in the DNA damage response[145]. Roles of Rad51 paralogues are discussed extensively later.

Origin independent replication fork restart at sites of stalled replication avoids the need for replication restart by full HR, therefore preventing risk of genetic rearrangement, discussed later, while also saving time and energy. The many different processes that are

intertwined to form a network that enable this demonstrate the importance of direct replication fork restart.

If direct replication fork restart is not an option, then often the stalled fork must undergo extensive remodeling to either (a) enable replication past the lesion *via* template switching or (b) to generate branched DNA substrates suitable for recombination mediated fork restart. Some suggested remodeling activities remain controversial, such as the fork regression theory that fills gaps in our understanding of fork restart and has been observed *in vitro*, but not *in vivo*. This involves helicase mediated regression of the nascent DNA strands that base pair with each other, and subsequent re-annealing of the parental duplex to form a structure reminiscent of a Holliday Junction, described below, referred to as a chicken foot. For example, DNA helicases RecG, Hef and FANCM from bacteria, archaea and eukaryotes respectively have been observed to reverse replication forks *in vitro*[74]. This is a theoretical way in which non-canonical template switching may occur for lesion bypass at the fork. Alternatively, stalled replication forks may be remodeled for restart by homologous recombination (HR). This is a high fidelity pathway that utilizes an intact homologous strand as a template to fill in gaps that may occur following DNA damage. Although a high fidelity means of repair and replication restart, HR is not risk free. In order to understand this we must consider the HR pathway in its entirety.

1.18. Recombination dependent repair: Homologous recombination

HR is an essential mechanism to support and repair compromised replication forks and other DNA damage and to generate genetic diversity in bacteria, archaea, eukaryotes and viruses. Initiation of HR requires ssDNA. This ssDNA may need to be generated, for example by resection from DNA ends at a double strand break (DSB), a process that is highly controlled as this commits the cell to repair by recombination over NHEJ[146], or by replication fork remodeling at sites of stalled replication. HR can be divided into three major stages known as pre-synapsis, synapsis and post-synapsis, but there is significant complexity and redundancy in these stages, dependent on the exact disposition of the type of DNA damage to be repaired, and the prevailing physical and functional protein-protein interactions.

1.19. Pre Synapsis: Recombinases and the nucleoprotein filament

Exposed ssDNA, for example if generated ectopically at sites of blocked DNA replication, is sequestered by single stranded DNA binding proteins that protect it and recruit DNA damage response molecules. For HR to proceed, single stranded DNA binding proteins must be removed to enable loading of recombinases onto ssDNA during pre-synapsis to form a nucleoprotein filament that is comprised of recombinase monomers. Recombinases are universal in their importance and function in bacteria, archaea and eukaryotes, and are known as RecA, RadA and Rad51[147] respectively. These assemble to form a nucleoprotein filament[148] that

catalyses strand invasion into homologous duplex, known as synapsis. Deletion of *rad51* is lethal in yeast and higher eukaryotes. An alternative eukaryotic recombinase, Dmc1, carries out the same function, but is required only for meiosis to generate genetic variation[149]. They are conserved, although primarily between the core ATP hydrolysis units. However, there are some differences that arise as a result of mechanistic variations and these can often be attributed to the varying levels of control required in each system. This is relevant to formation and stabilization of recombinase filaments described later.

RecA loading in context of DSB or single strand break repair in bacteria, is carried out by RecFOR, RecBCD and AddAB helicase- nuclease complexes. *Via* extensive resection activities, these macromolecular complexes generate recombinogenic ssDNA onto which RecA is loaded to stimulate HR. These could be described as HR initiation machines. Recombinase loading occurs very differently in eukaryotes, relying on a multitude of proteins. *S. cerevisiae* Rad52 is a ring shaped oligomer that that interacts with RPA and assists in loading of Rad51 onto ssDNA. However human Rad52, though similar in structure, does not seem to be important for Rad51 loading. Instead, other proteins such as BRCA2 facilitate Rad51 loading and filament formation[150] following resection of DNA at DSBs. BRCA2 is a carrier of up to eight Rad51 monomers *via* interaction *via* BRC repeats. BRCA2 displaces RPA and loads Rad51 onto ssDNA where it polymerises to form a nucleoprotein filament. BRCA2 deficient mice are not viable, and in humans, mutations in the BRCA2

gene are linked to several cancers, resulting from catastrophic loss of control or activity of accurate DNA repair by homologous recombination.

Recombinases nucleate on ssDNA and polymerise cooperatively on ssDNA forming a right handed nucleoprotein filament[148], Figure 1.15, that extends along DNA in either a 3' to 5' or 5' to 3' direction, although assembly typically occurs faster 3' to 5'. Each recombinase monomer binds three nucleotides, and filament assembly by recombinase enzymes requires bound ATP[148]. The state of this nucleotide cofactor dictates whether the filament is in the active ATP bound form, or the inactive adenosine diphosphate (ADP) bound form that stimulates filament disassembly. Mechanistic differences between isolated RecA, RadA and Rad51 proteins is exemplified by each recombinase binding and hydrolysing ATP with different affinities[151,152], probably reflecting mechanistic adaptation to interaction with modulating proteins during the complex events that occur during recombination synapsis. For example, RecA is stable in complex with ATP and forms a stable nucleoprotein filament. However, RadA and Rad51 hydrolyse ATP easily and as a result, the nucleoprotein filament is less stable and prone to disassembly. To overcome this, there are several mediator proteins that assist with recombinase nucleoprotein filament (NPF) formation that assist with recombinase loading, nucleation and subsequent filament stabilization. Within the nucleoprotein filament, DNA adopts an extended form, 150% longer than standard B-form DNA[153], thereby enhancing the efficiency of

the subsequent homology search by providing enhanced access to base-sequences during the search.

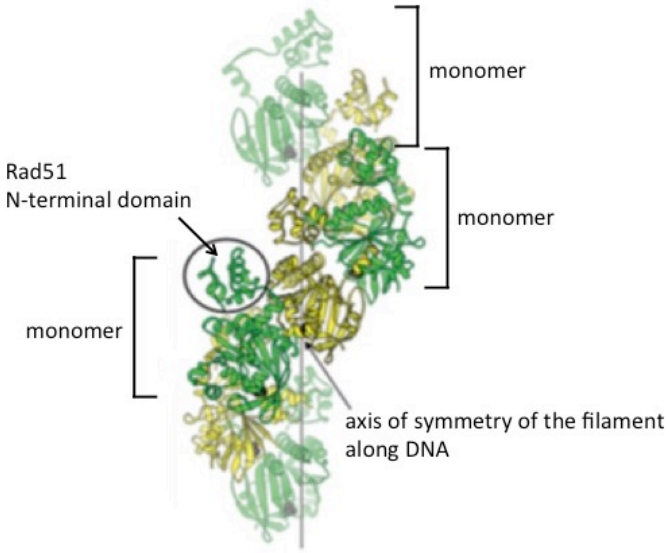


Figure 1.15: Picture of the atomic resolution structure of *S. cerevisiae* Rad51 nucleoprotein filament. The diagram shows a vertical axis around which the filament forms (grey line) and seven molecules of Rad51 protein, with the N-terminal highlighted by a circle. Approximate positions of each Rad51 monomer are indicated, and alternate in the figure as green and yellow in colour. Figure adapted from Conway *et. al.*, 2004[148].

Rad51 paralogues, discussed extensively later, found in eukaryotes, archaea and recently, bacteria, have been revealed to play an important role in NPF formation[154,155]. Studies in yeast have yielded most information about the Rad51 paralogues. A heterodimer of Rad55-Rad57 has been shown to stabilize the Rad51 NPF[156] *via* possible incorporation into the Rad51 filament. Likewise, the human Rad51B-Rad51C complex stabilizes the Rad51 filament[157]. It has been suggested that through this incorporation, the filament adopts characteristics of the bacterial RecA filament, in which ATP hydrolysis is reduced, thereby stabilising the

filament. In addition, studies in *Caenorhabditis elegans* suggest that Rad51 paralogues also open up the Rad51 filament, assisting with the homology search[158]. Single molecule studies have revealed that *S. cerevisiae* Rad51 interrogates for regions of homology by sampling eight nucleotides at a time[159]. If this minimum requirement is met and extended to nine, then strand invasion occurs into homologous duplex by the NPF in increments of three nucleotides[159]. It is likely that RadA and human Rad51 act in a similar way.

1.20. Synapsis: Strand invasion

On strand invasion, a strand of DNA is displaced forming a D-loop structure in a process known as synapsis, Figure 1.16. The invading strand forms Watson-Crick base pairs with complementary regions of DNA to form heteroduplex DNA. Rad51 is displaced from heteroduplex DNA by Rad54 to allow DNA synthesis by DNA polymerases δ and η, primed by the invading strand and using the intact homologous sequence as a template. Replicative extension of the invading strand can progress in concert with other specialized HR helicases, into a Holliday junction (HJ) *via* 'branch migration'[105].

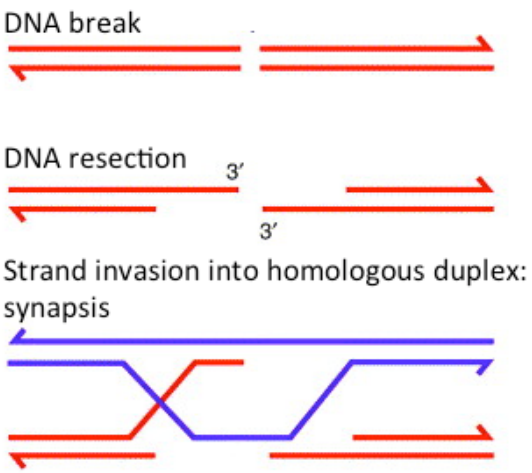


Figure 1.16. The process of synapsis initiates homologous recombination. A broken DNA strand or strands is resected by nuclease activity to generate 3'-OH ended ssDNA tails that can be loaded by a recombinase enzyme to begin a homology search for an intact DNA duplex that has sequence similarity to the broken strand. This generates a "joint" DNA molecule in which the recombinase coated ssDNA is invaded into the duplex. RecA in bacteria, RadA in archaea and Rad51 in eukaryotes catalyse the strand invasion part of synapsis. Figure adapted from genetics.org.

1.21. Post Synapsis: D-loop extension and branch migration into Holliday Junctions

The extent of HR branch migration and HJ formation depends on the context in which HR is required in a cell. Control of the extent of synapsis during HR is important to ensure replication can be restarted rapidly and to prevent gene conversion that can arise from HJ resolution if HR is extensive. Branch migration is controlled by helicases, some of which also dissociate HJ recombination intermediates to generate a substrate suitable for replication fork restart. Meanwhile, the other 3'-OH overhang at the other broken DNA end forms base pairs with the

displaced strand during 'second-end capture'. Rad52 is required for second end capture of the remaining 3'-OH ssDNA strand that anneals to the homologous region on the displaced strand of the D-loop. During canonical HR, the extended invading strand is ligated to its original strand, forming a joint molecule, and a HJ structure, Figure 1.17. HJs are mobile and their movement and direction can be catalyzed by DNA helicases. Following extension, strands are ligated to their respective original strands to form intact duplex DNA creating a double HJ. To complete HR the joint DNA molecules must be separated by 'resolution' to give crossover and non-crossover products, or 'dissolution' leading to non-crossover products[160]. Crossover refers to reciprocal exchange of genetic material and can result in gross chromosomal rearrangement; where as non-crossover at most can result in gene conversion following copying of homologous DNA during synapsis. There is some debate over which protein nuclease complexes carry out HJ resolution in humans. It is likely that the complex used is dependent on the context in which HR is required, however, candidates include MUS81-EME4 and SLX1-SLX4 to form the SLX-MUS complex, Mlh1-Mlh3-EXO1 (crossover products only) and GEN1[161,162]. Alternatively, HJ dissolution may occur to give rise to non-crossover products, carried out by BLM-TopoIIIα-RMI1-RMI2[163]. The way in which this is achieved can lead to different genetic outcomes for the cell. HJs can be processed by resolution or dissolution mechanisms that are carried out by different protein complexes. Although genetic exchange is useful during mitosis to introduce variation into the genome, it can be

dangerous in context of DNA repair and in humans can lead to cancers. For this reason, extensive HR is often blocked prior to HJ formation in order to divert toward other recombination dependent DNA repair pathways at the pre-synaptic and synaptic checkpoints mentioned earlier.

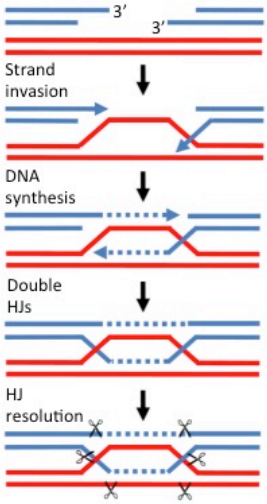


Figure 1.17. Summary of events occurring classical Double-Strand Break Repair (DSBR), a model of homologous recombination. DSBR is initiated by synapsis (see Fig. 1.16) and post-synaptic events include extension of the invaded DNA strand within the joint molecule by DNA replication, forming a D-loop. The process of "branch migration" can lead to double Holliday junctions (dHJ) formation, to extend nascent DNA through the region of resection. HJ resolution is required to separate the joined homologous DNA molecules, catalysed by structure-specific nucleases called HJ resolvases, here indicated by scissor icons. Depending on the positions of the resolvase DNA strand cutting, products may be "crossovers" or "non-crossovers", as explained in the main text. Adapted from Hiom, 2000[164].

Not all proteins involved in the HR pathway promote HR, for example, Rad54 in yeast antagonizes annealing of the second 3'-OH overhang preventing second end capture. This promotes other

recombination dependent repair pathways such as SDSA, discussed below.

The pre- synaptic and synaptic stages where the NFP and D-loop is formed are important 'checkpoints' during the HR process. The structures generated during these stages can be remodeled to divert toward other recombination dependent repair pathways, as this structure is common to all three pathways, to be discussed later.

1.22. Homologous recombination dependent repair by SDSA and BIR

Although a highly accurate DNA repair process in comparison to other repair mechanisms, extensive HR can be dangerous as it has the potential to generate gross chromosomal rearrangements through generation of crossover products following HJ processing. However, there are alternative pathways to HR that stem from the synapsis stage of HR. These pathways are known as synthesis-dependent strand annealing (SDSA) and break induced replication (BIR). SDSA is beneficial in the context of DNA repair as it produces non-crossover products, whereas BIR is a recombination dependent form of replication fork restart[165,166]. Therefore it is in the interest of the cell to divert HR away from HJ formation and towards these alternative recombination dependent repair pathways, in which BIR would be useful in the event of stalled replication. There are variations on BIR such as micro-homology mediated BIR (MMBIR)[167]. Regulation of HR, SDSA and BIR from synapsis is discussed later.

SDSA occurs if second end capture of the second 3'-OH DNA end does not occur, and proceeds following extension of the 3'-OH of the invading strand and then unwinding of the D-loop structure formed during synapsis. This process depends on annealing of complementary DNA strands. It may be necessary to extend DNA to fill-in gaps before ligation, Figure 1.18A. Genetic analysis in *Saccharomyces cerevisiae* suggests diversion to SDSA from HR when Srs2 helicase is defective and leads to hyper-recombination, as a result of extensive homologous recombination. Srs2 helicase is known to disassemble Rad51 filaments[168], thus supporting the idea that diversion away from full HR occurs. This demonstrates that helicases are major players in diversion to SDSA, however they are supported by other proteins that antagonize HR if synapsis is allowed to occur.

In brief, SDSA occurs by melting of heteroduplex DNA following extension of the invading strand prior to second end capture during HR. Dissociated strands are annealed and DNA synthesis occurs as required[169].

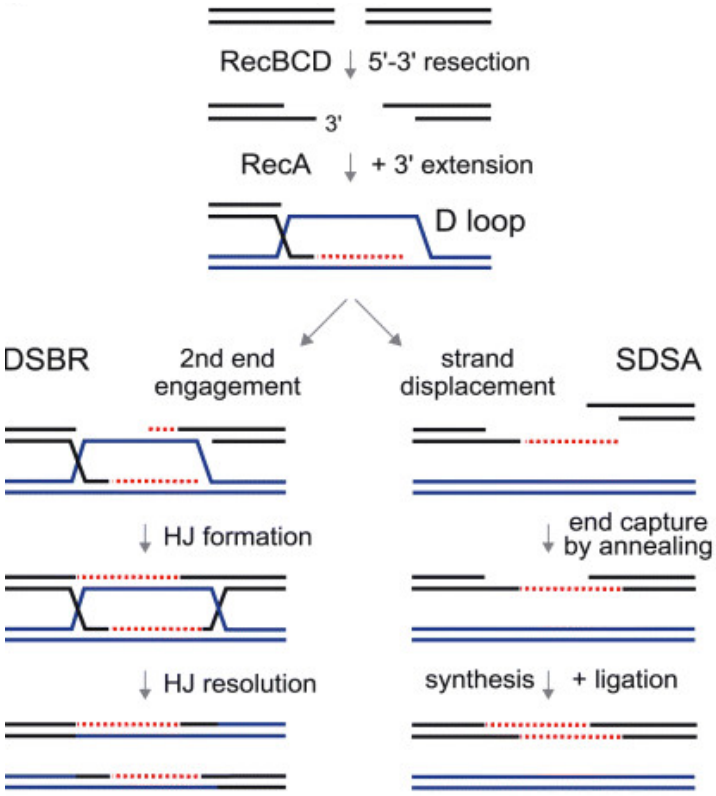


Figure 1.18A. Summary contrasting post-synaptic events during DSBR (see Fig. 1.17) and an alternative called Synthesis Dependent Strand Annealing (SDSA). In both processes the invaded DNA at synapsis is extended by DNA replication. In SDSA, HJ formation is avoided by dissociation of the joint molecule containing nascent DNA, which can be annealed to its homologue leading to further replication through "gap-filling", DNA strand ligation and completion of DNA repair without recourse to full HR. Adapted from Meddows *et. al.*[170].

Several independent research groups first suggested that replication could be initiated or restarted by means of recombination more than seventy years ago and was observed during studies of DNA replication in bacteriophage[171]. Following from this, replication was demonstrated to be started by means of recombination by Kogoma in *E. coli*[172–175]. Replication by BIR utilizes the standard replication machineries

described earlier (MCM and Pol α). BIR proceeds from the point of the D-loop, and therefore is recombination-dependent[165,166], Figure 1.18B. Although BIR is useful for replication fork restart, subsequent replication is not high fidelity and is mutagenic. BIR Is linked to human diseases, as this process can be highly mutagenic.

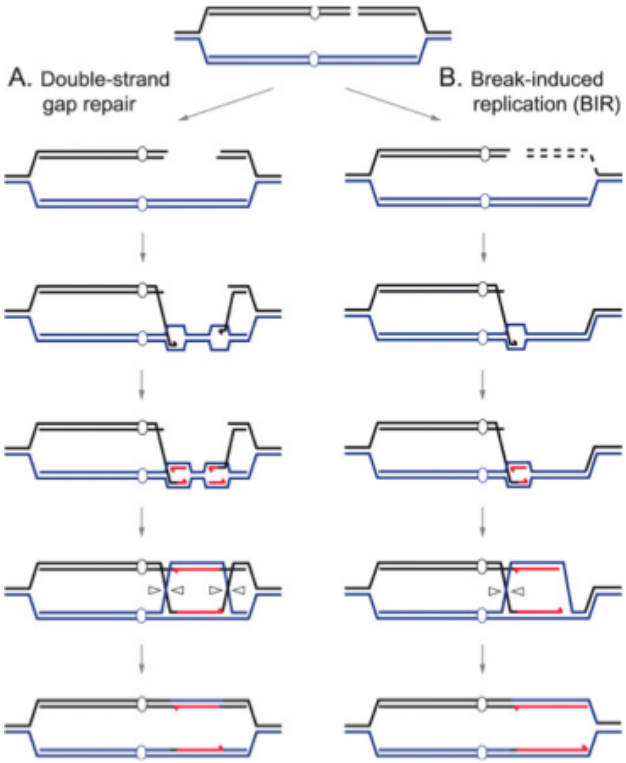


Figure 1.18B. Summary contrasting post-synaptic events during DSBR (see Fig. 1.17) and an alternative called Break-Induced Replication (BIR). In the latter a single HJ is sufficient for repair by avoiding a second DNA end capture event that is characteristic of DSBR. Adapted from Meddows *et. al.*[170].

As previously described, BIR and SDSA stem from the strand invasion stage of HR, and so that can be regarded as an important checkpoint as to which pathway is followed. For this reason, there are a

number of proteins that target this stage of HR. The majority of these are the DNA helicases and translocases.

1.23. Regulation of synapsis

Synapsis can be considered to be a checkpoint during HR as this structure is common to HR, SDSA and BIR. Pathways such as SDSA require D-loop disassembly to restore DNA integrity. There are many proteins, several of which are found in the HR pathway such as Rad54, that can be described as pro- or anti-recombinogenic. The majority described here are anti-recombinogenic DNA helicases or translocases. These are important in the event of stalled replication to ensure recombination-mediated fork restart and not extensive HR. There are two approaches by which D-loop regulation can be achieved: (a) prevention of synapsis *via* targeting of pre-synaptic intermediates and (b) D-loop disassembly. D-loop formation and disassembly is regulated by DNA translocases. Initially, prevention of D-loop formation will be discussed.

As previously described, the recombinase NPF is required to catalyse strand invasion into homologous duplex, to form a D-loop, therefore the NPF is a major target that dictates whether or not D-loops can form. If the NPF is disassembled from ssDNA, then HR cannot proceed. DNA translocases can disrupt the NPF in an active (ATP-dependent) or passive (ATP-independent) manner, although in some cases there seems to be some crossover. Translocases such as UvrD that disassemble the NPF actively use the energy released from ATP hydrolysis

to translocate along ssDNA and physically displace recombinase monomers from DNA[106]. Other translocases, such as the well characterized *Saccharomyces cerevisiae* Srs2, do not require ATP driven translocation to catalyse NPF disassembly, but rather rely on physical interaction between the translocase and the NPF[168,176]. This can stimulate ATPase activity within the filament, therefore switching the filament from the active ATP bound form to the inactive ADP bound form and therefore triggering release of recombinase monomers. Examples of translocases that function in this way are found among bacteria, archaea and eukaryotes.

If the NPF is not targeted and synapsis is allowed to proceed, again, this process can be reversed. In these circumstances, the recombinase filament may still be in complex with double stranded DNA, or the NPF may have already been removed. DNA helicases such as RTEL1 and RECQ5 are able to displace invading DNA strands from D-loop structures[177,178], releasing ssDNA for processing during other pathways. Alternatively, a DNA helicase named HelQ is capable of removing Rad51 NPFs from dsDNA *in vitro*[179]. In a physiological context, HelQ may be required for removal of such filaments from heteroduplex DNA within a D-loop. This demonstrates how DNA translocases and helicases can remodel both DNA-protein complexes and D-loops in order to prevent HR from occurring and diverting towards pathways that generate safer non-crossover products.

HelQ is implicated in repair of stalled replication forks and has been shown to have possible pro- and anti- recombinogenic properties. This interesting protein whose physiological function remains elusive is focused on from now on as the major experimental work of this thesis, with its archaeal homologue, Hel308.

1.24. Hel308 and HelQ DNA helicases

The Hel308 family of DNA repair helicase homologues was unearthed by genetic analysis of cellular sensitivity to DNA crosslinking agents in *Drosophila melanogaster*[180]. Mutation of one gene, *mus308*, was found to confer sensitivity to nitrogen mustard, a chemical causing inter- and intra- strand DNA crosslinks. *PolQ*, the human homolog to *mus308,* encodes a protein containing a N-terminal helicase region, and a C-terminal polymerase[181]. Biochemical analysis of PolQ, otherwise known as DNA Polθ, reveals the helicase region is not active, although the C-terminal shows functional polymerase activity[182]. PolQ has been linked to translesion synthesis, as it displays some interesting non-canonical polymerase properties such as base insertion and extension from AP-sites and is also linked to micro-homology mediated end joining (MMEJ).

Designated as part of the *mus308* family of proteins, two more sequence homologs of *PolQ* were identified in humans: Hel308[4] (a SF2 DNA helicase) and PolN[183] (a putative translesion polymerase) which contain only the helicase and polymerase portions of PolQ respectively, Figure 1.19. Hel308 sequence homologs have since been identified in

several species of archaea[5] and metazoans[4], although they are absent in yeasts and bacteria[184]. Genetic studies in archaea and mice revealed that Hel308 deficiencies confer sensitivity to ICL inducing agents similarly to *mus308* in *D. melanogaster*[185]. HelQ deficient mice also show slowed replication [185].

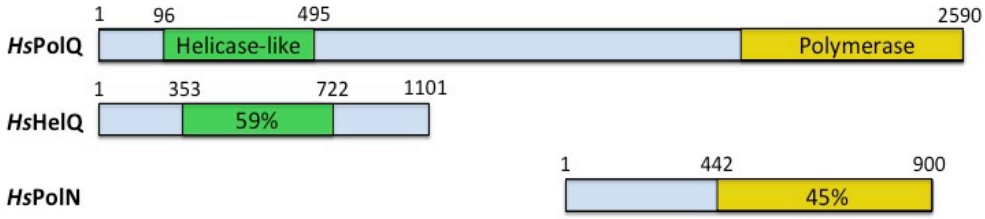


Figure 1.19: Cartoon to highlight similarities in domain organization between human proteins PolQ, HelQ and PolN. HelQ helicase shares 59% amino acid sequence similarity with the helicase-like domains of PolQ, and the latter has 45% sequence homology to the polymerase domains of PolN.

1.25. *In vitro* activity of Hel308 and HelQ

Human Hel308, now renamed HelQ, is a 3' to 5' ATP-dependent helicase which shows ssDNA-dependent ATPase activity[4]. It has a minimal helicase substrate of a partial duplex with a 3'-OH overhang, but preferentially targets branched structures[5]. Stimulated by interaction with RPA[185,186], both archaeal Hel308 and HelQ unwind lagging strands of branched substrates consistent with DNA structures found at stalled replication forks[5,187]. This perhaps is indicative of a functional role for Hel308 and HelQ *in vivo* by which they may expose ssDNA for loading of proteins involved with replication fork restart *via* HR mechanisms[187,188]. *In*

vitro evidence supports this, as described in results sections. Hel308 from the archaeon *Methanothermobacter thermautotrophicus* has also been shown to dissociate D-loops[188]. The oligomeric state of Hel308 and HelQ is unclear, although a recent atomic structure of the PolQ helicase domain, homologous to HelQ, is tetrameric, and it was proposed that it may function as a tetramer for MMEJ[189].

Caenorhabditis elegans (*Cel*) Hel308 binds to Rad51 peptide filaments on ssDNA and promotes their disassembly from dsDNA[179], suggesting a possible anti-recombinogenic role by removal of the Rad51 NPF required for strand invasion. The authors suggest that the *Cel*Hel308 C-terminus interacts with Rad51, triggering the Rad51 NPF to depolymerize in an ATPase-independent manner[179]. *Sulfolobus solfataricus* Hel308 is a robust translocase having the ability to displace streptavidin from DNA coated with biotin[190], perhaps indicating potential for Hel308 and HelQ to remove proteins such as Rad51 or RPA from DNA in a role that remodels branched DNA molecules.

1.26. Atomic resolution structures of Hel308

Crystal structures of archaeal Hel308 from *Archaeoglobus fulgidus*, *Sulfolobus solfataricus* and *Pyrococcus furious* reveal a five domain architecture[190,191], Figure 1.20. No structural data is currently available for HelQ, although close sequence homology would suggest a similar structure. An atomic resolution structure for the closely related PolQ helicase domain has recently been elucidated[189], discussed in Chapter 4.

In addition, HelQ sequences in higher eukaryotes have a 300 residue N-terminal 'tail' that is absent from archaeal Hel308.

There is only one isoform of Hel308 present in *Methanothermobacter thermautotrophicus*, however, in some archaeal species such as *Haloferax volcanii*, there are at least two isoforms. Multiple isoforms also exist in metazoans such as *Adineta vaga* and *Homo sapiens* that have four isoforms of HelQ.

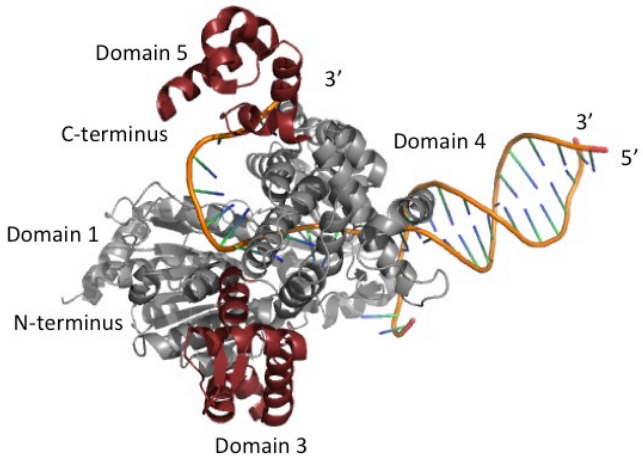


Figure 1.20: Illustration of the atomic resolution crystal structure of archaeal Hel308 from *Archaeoglobus fulgidus* bound to a 3'-OH tailed DNA duplex. The "core" helicase domains (1, 2 and 4) are shown in grey, comprising a helicase ratchet and two RecA-like ATPase domains. Two "accessory" domains are shown in scarlet, a winged helix domain (domain 3) and the other, domain 5, that binds ssDNA and is thought to be a brake on helicase activity. Coordinates were downloaded from PDB accession number 2P6R, deposited by Büttner *et. al.*[191], and were rendered using MacPymol.

Biochemical analysis carried out on archaeal Hel308 has identified possible functional roles for each domain. Sequence homology comparisons reveal that domains 1 and 2 are RecA-like and form a motor

domain confirmed by crystal structure evidence[191]. The seven superfamily 2 helicase motifs (including Walker A and B) are located at the RecA-like domain interface where an ATP binding and hydrolysis site is formed[191]. Domain 2 contains a β hairpin responsible for separation of duplex DNA. Domain 3 forms a winged helix domain (WHD) that is a helicase accessory domain, whose role is not yet determined. However, close packing of domain 3 against domains 1 and 4 place it in an ideal position for coordinating between domains, perhaps coupling ATP hydrolysis to another function of the protein[192]. A coordination role has been assigned to a WHD in the Brr2 RNA helicase which bears structural resemblance to Hel308[193]. The WHD therefore remains an interesting structural feature of Hel308. Domain 4 has been identified as a helicase ratchet that contributes to the SF2 helicase core that confers directional 3' to 5' translocation, and accessory domain 5 has been identified as auto inhibitory, perhaps functioning as a molecular brake[191,194]. Domains 1, 3 and 4 adopt a ring like conformation, encircling 3'-OH ssDNA likely contributing to Hel308 processivity[191].

Structural data from AfuHel308 in complex with partial duplex DNA provides valuable information regarding contacts between Hel308 and DNA, although this structure is not in complex with its preferred substrate. However, extensive contacts with up to 60 nucleotides are observed inside the central channel.

Although the Hel308 translocation mechanism is not fully understood, it is thought to be similar to that observed for SF2 NS3 helicase described in section 1.11 that functions with a helicase ratchet, similar to that found in Hel308 domain 4.

In vivo, HelQ deficient U2OS cells show increased radial chromosome formation and chromatid breaks when treated with ICL inducing MMC[185,195]. Radial chromosome formation following such treatment is characteristic of FA cells[195]. Although occasionally described as a possible FA related gene[195], through use of mouse models HelQ has been shown to act in parallel to the FA pathway[185], although some overlap of function may exist.

Fluorescence co-localization studies show HelQ localises to sites of MMC-induced stalled replication forks where it colocalises with RPA, Rad51 and FANCD2 foci, key proteins involved in processing of stalled replication forks by HR[187]. Evidence suggests HelQ acts in HR following Rad51 recruitment as Rad51 foci remain at damage sites for longer in HelQ deficient cells treated with MMC[195].

Co-precipitation assays from two independent studies show HelQ interacts with endogenous forms of the Rad51 paralog complex BCDX2, DNA damage signaling molecule ATR, RPA and the FANCD2-FANCI heterodimer[185,195], Figure 1.21. These interactions occur in both MMC treated and untreated cells. No interaction with PolN was observed in recent studies, conflicting with an interaction reported by Moldovan *et. al.* in 2010[196].

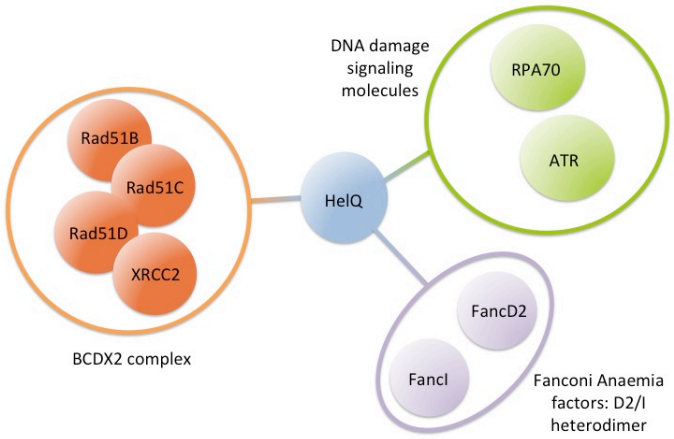


Figure 1.21: Summary of detected interactions between HelQ protein and other proteins of DNA repair. Adapted from Adelman *et. al.*[185].

Identification of human HelQ protein-protein interactions provides an interesting basis for further research, which may give insight into HelQ function with some or all of these proteins. Interaction with ATR and RPA suggests HelQ may influence signaling during the ICL repair process. Similarly, a direct physical interaction has been detected between Hel308 and archaeal RPA supporting this[186]. ATR, a checkpoint kinase, is known to be involved in signaling in response to ICL repair following cellular MMC treatment *via* RPA-ATR-CHK1 signalling[197], in which downstream signaling effects of phosphorylated CHK1 cause cell cycle arrest in G2/M phase. Following MMC treatment of HelQ deficient cells, CHK1 phosphorylation was reduced and fewer cells were found to be in the Gap 2 (G2)/ Meitotic (M) phase[195]. Interaction with the BCDX2 complex poses interesting potential roles for HelQ as Rad51 paralogue functions are not yet fully understood.

1.27. Rad51 paralogue complexes: BCDX2 and CX3

Rad51 paralogues Rad51B[198], Rad51C[199], Rad51D[200], XRCC2[201] and XRCC3[202] have been identified to have approximately 20-30% sequence homology to Rad51. Cellular deficiency of any one of these causes sensitivity to ICL inducing agents[203]. These proteins show no recombinase activity, unlike Rad51, although they do have weak ATPase activity[157] that may be involved in activating replication fork restart[144]. Rad51 paralogues assemble to form two major multi-protein complexes: Rad51C and XRCC3 to form CX3; and Rad51B, C, D and XRCC2 to form BCDX2[203]. Smaller sub-complexes also exist: Rad51B-Rad51C, Rad51C-Rad51D and Rad51C-Rad51D-XRCC2[203].

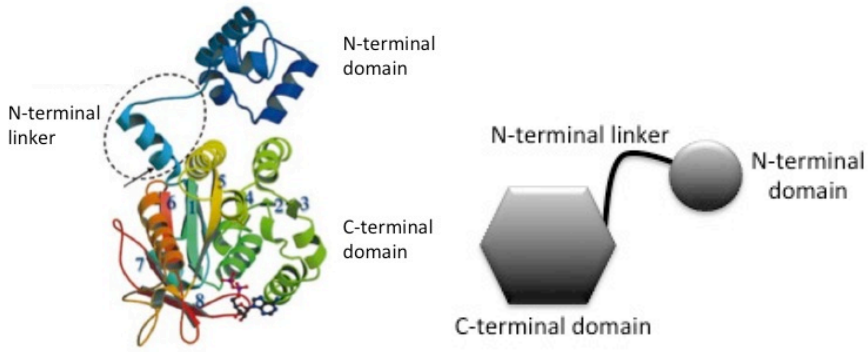


Figure 1.22: Left: Modeled structure of Rad51B, where the C-terminal domain connected to the N-terminal domain *via* an N-terminal linker can clearly be seen. Adapted from Miller *et. al.*[204]. **Right:** Generalized overview of domain structure in Rad51 paralogues. A large C-terminal domain is connected to a smaller N-terminal domain *via* a N-terminal linker. Adapted from Suwaki *et. al.*[203].

No complete structures are available for the Rad51 paralogues, however molecular modeling predicts all but XRCC2 to have N and C-terminal domains connected by a linker[204], Figure 1.22. Yeast two-hybrid analysis and mutagenesis suggests that Rad51 paralogues interact with each other *via* N-terminal linker - C-terminal domain interactions, see Figure 1.23. How this fits in with circular complex structures observed by electron microscopy (EM)[205] remains unclear.

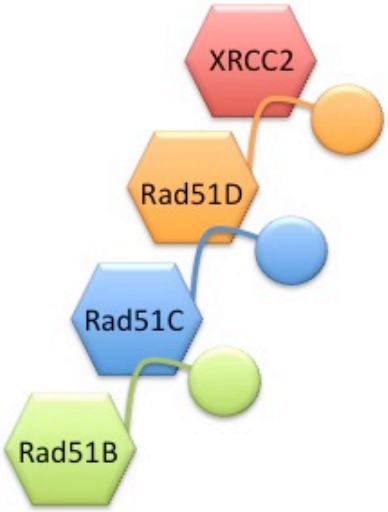


Figure 1.23: Cartoon illustrating predicted interactions of proteins Rad51B, Rad51C, Rad51D and XRCC2 that form the "BCDX2" paralogue complex. N-terminal linkers interact with C-terminal domains as shown. Adapted from Suwaki *et. al.*[203].

1.28. Rad51 Paralogues: Roles in homologous recombination

Although a clear role for Rad51 paralogue function remains elusive, evidence indicates that they influence certain DNA repair pathways which stem from HR DSB repair such as SDSA and single strand annealing (SSA)[203] perhaps to regulate the extent of which gene

conversion occurs. SDSA is a repair pathway that branches from the central HR pathway to form non-crossover products. XRCC2 and XRCC3 have been linked to regulation of the extent to which gene conversion is allowed to occur[206].

In order for Rad51 paralogues to influence such pathways it might be expected that they participate at various stages of HR. Indeed, formation of Rad51 foci at sites of damage is not efficient in the absence of Rad51 paralogs[207] implying an early role in HR, whereas the CX3 complex has been implicated in HJ resolution during late stages[208]. Electrophoretic mobility shift assays show both CX3 and BCDX2 bind branched DNA structures, such as those found at stalled replication forks or HJs[209]. EM confirms a preference for branched DNA substrates by BCDX2 and CX3. In addition, EM studies revealed possible ring like formations of the paralogue complexes which may give insight into how they bind DNA[205]. BCDX2 preferentially binds branched DNA structures reminiscent of stalled replication forks[209], similarly to HelQ, and also displays strand annealing activity[209], although whether this has any *in vivo* significance remains to be seen.

The Rad51C-Rad51B sub-complex has been identified as a mediator of Rad51 NPF formation at pre-synapsis[210] similarly to Rad55-57 in *Saccharomyces cerevisiae*[211]. *In vitro* experiments show the Rad51B-Rad51C complex stabilizes the Rad51 NPF in the presence of BLM[157], an anti-recombinogenic helicase which dissociates D-loops[105] and is capable of Rad51 NPF disassembly[157]. The authors speculate that Rad51B-Rad51C

complexes integrate into the Rad51 NPF *via* interaction with the Rad51 N-terminal ATP cap.

1.29. HelQ, Rad51 paralogues and cancer

Recently linked to head and neck cancers by genome wide association studies[212], human HelQ has become a medically important target for research. Genetic studies in mice also revealed increased incidence of ovarian cancer and fertility defects in females deficient in HelQ[185]. Rad51C, present in both major Rad51 paralogue complexes, has also been linked to ovarian and breast cancers[203]. These proteins are the latest in a list of HR related proteins that have been linked to cancer, highlighting the importance for HR repair in protection from cancer. HelQ defects are linked to the onset of several cancers[213,214], making it relevant to understand its precise role in maintaining genome stability.

1.30. Rad51 paralogues stabilise stalled replication forks

Depletion of Rad51 paralogues leads to increased frequency of double strand breaks, a product of replication fork collapse from unstable replication forks[144]. This implicates a role for Rad51 paralogues in replication fork stabilisation. Activation of ATP hydrolysis by these proteins appears to mediate activation of replication fork recovery.

1.31. Rad51 Paralogues: Roles in signaling

Rad51C, recently identified as a potential new FA complementation group[215, 216] that may be involved in ICL repair, seems to play a crucial role out of all Rad51 paralogs. Found within both major complexes, Rad51C has been implicated in both early and late stages of HR. Rad51C is the only Rad51 paralog to contain a nuclear localization signal[203] which may be relevant to a role in signal transduction. Rad51C has been implicated in ATM signaling on detection of damage by influencing phosphorylation of CHK2 and inducing cell cycle arrest[145].

1.32. Summary

Current evidence points to involvement of HelQ during processing of DNA and protein:DNA structures at sites of stalled replication at early stages of HR where it may participate in resetting of replication forks following collapse. Rad51 paralogs are also believed to act during early HR. Reported interactions between HelQ and components of the FA pathway, DNA damage signaling molecules and major Rad51 paralog complex BCDX2 suggests an intriguing role for HelQ in recombination in downstream ICL repair and DNA repair networks.

1.33. Aims and development of the project

The initial project aims were to begin to define the molecular interactions that occur between HelQ and the Rad51 paralog complex BCDX2, and their effects on HelQ activity. To achieve this required reliable

and reproducible expression and purification of human proteins HelQ, Rad51, and Rad51 paralogs Rad51B, Rad51C, Rad51D, XRCC2, described in Chapter 3. This would then enable (a) more detailed characterization of HelQ helicase activities, especially in comparison with the archaeal homologue Hel308, (b) attempts to reconstitute the BCDX2 complex and investigate its effect on HelQ activity *in vitro*, and (c) attempts to investigate relative activities of HelQ and Rad51 during synapsis and D-loop formation. As would become clear during the project, handling and analysis of human Rad51 paralogues was not possible. However, biochemical characterization of HelQ *in vitro* and modeling of some new observations was possible (Chapters 3 and 4). The project also developed an analysis of the winged helix domains of Hel308 and HelQ (Chapter 5), which had not been studied but are an essential part of the Hel308/HelQ DNA repair activity.

Chapter 2: Materials and Methods

2.1 Chemicals

All chemicals were supplied by Sigma or Fisher Scientific unless stated otherwise.

2.1.1 Antibiotics.

Table 2.1. Antibiotics used in this work, as referred to in the main text.

Antibiotic	**Supplier**	**Stock concentration**	**Working concentration**
Carbenicillin	Melford	20 mg/ml in sterilized water	50 µg/ml
Kanamycin	Sigma	4 mg/ml in sterilized water	40 µg/ml
Spectinomycin	Sigma	10 mg/ml in sterilized water	80 µg/ml
Chloramphenicol	Sigma	7 mg/ml in ethanol	15 µg/ml
Penicillin - Streptomycin	Sigma	10,000 Units Penicillin 10 mg/ml Streptomycin in sterilized water	100 Units Penicillin 0.1 mg/ml Streptomycin

Table 2.2: *E. coli* strains used as referred to in the main text.

***E. coli* strain**	**Supplier**	**Relevant parts of the genotype for each strain**
XL1 Blue	Agilent Technologies	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F´ proAB lacIq ZΔM15 Tn10 (Tetr)]*
DH5α	Invitrogen	*F⁻ Φ80lacZΔM15 Δ(lacZYA-argF) U169 recA1 endA1 hsdR17(r_k^-, m_k^+) phoA supE44 thi-1 gyrA96 relA1 λ⁻*
BL21 AI	Invitrogen	*F⁻ompT hsdS_B (r_B^- m_B^-) gal dcm araB:T7RNAP-tetA*
Lemo 21(DE3)	NEB	*fhuA2 [lon] ompT gal (λ DE3) [dcm] ΔhsdS/ pLemo*(Cam^R) *λ DE3 = λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5* pLemo = pACYC184-*PrhaBAD-lysY*
Rosetta 2 (DE3) pLacI	Novagen	*F⁻ ompT hsdS_B (r_B^- m_B^-) gal dcm (DE3)* pLacIRARE2 = *(Cam^R)*
Rosetta (DE3)pLysS	Novagen	F⁻ *ompT hsdS_B*(r_B^- m_B^-) *gal dcm* (DE3) pLysSRARE *(Cam^R)*
BLR (DE3)	Novagen	F⁻ *ompT hsdS_B*(r_B^- m_B^-) *gal dcm* (DE3) Δ(srl-recA)306::Tn*10* (Tet^R)
BL21-CodonPlus (DE3)-RIL strain	Agilent Technologies	Additional rare codons provided by: *argU* (AGA, AGG), *ileY* (AUA), *leuW* (CUA)
OmniMax	Thermo Fisher Scientific	{*pro*AB *lac*I^q *lac*ZΔM15 *Tn*10(Tet^R) Δ(*ccdA*B)} *mcr*A Δ(*mrr hsd*RMS-*mcr*BC) Φ 80(*lac*Z)ΔM15 Δ(*lac*ZYA-*arg*F)U169 *end*A1 *rec*A1 *sup*E44 *thi*-1 *gyr*A96 *rel*A1 *ton*A *pan*D

E. coli strains used for cloning have been engineered to optimize cloning efficiency by stabilizing plasmid propagated; for example deletion of *recA* minimizes potential recombination events between plasmid and

chromosome, which would lead to loss of plasmid. Cloning strains may carry additional properties such as features to enable 'blue – white' screening *via* insertional inactivation of *lacZ*, described in more detail in section 2.17.

E. coli strains used for protein expression are engineered to be deficient in proteases such as omp and ION, which may degrade heterologous proteins during their expression. BL21 A.I. (Arabinose Inducible) is a specialized expression strain derived from BL21. Expression of chromosomally inserted T7 RNA polymerase is under control of the *ara*Bad promoter, resulting in its inducible expression when L-arabinose is added to growth media, or its repression by addition of glucose. This approach helps to minimize basal protein expression and is suited for expression of toxic proteins that cause problems for *E. coli* growth. Lemo21 (DE3) offers IPTG inducible over-expression of proteins encoded on T7 based expression vectors. Extra control of protein expression in this strain is provided by inducible expression of T7 lysozyme, an inhibitor of T7 RNA polymerase, which is under control of the $P_{rha\text{BAD}}$. T7 lysozyme, carried on pLEMO, is therefore induced on addition of L-rhamnose if required, acting to prevent leaky basal expression. Chloramphenicol is required to maintain pLEMO. Rosetta2(DE3)pLacI *E. coli* also offers tight regulation of protein over-expression. pLacI carries the *lacI* gene for additional repression of the lac operator. The T7 RNA polymerase is under control of P_{lac}. Rosetta strains usually carry an additional plasmid, pRARE, that encodes tRNAs for seven rare codons that aims to improve protein

yields. For Rosetta2(DE3)pLacI the tRNAs are encoded by the pLacI plasmid.

Table 2.3: Cell lines were used as referenced in the main text.

Cell line	**Supplier**
Sf9, derived from IPLB-SF21-AE	Invitrogen

Transfection reagents

FuGene HD (Promega) was used to transfect bacmid DNA into Sf9 cells to generate baculovirus for protein production as described in section 2.29.

Table 2.4. Compositions of general buffers

Buffer or reagent	**Composition**
50x TAE:	2 M TRIS 1 M Acetic Acid 50 mM EDTA
10x TBE	1 M TRIS 1 M Boric acid 20 mM EDTA
6x SDS-PAGE loading dye:	375 mM TRIS-HCl pH6.8 6% SDS 48% (v/v) Glycerol 5 mM DTT 0.03% (w/v) Bromophenol Blue
10x SDS-PAGE running buffer:	250 mM TRIS 1.92 M Glycine 1% (w/v) SDS
10x TG buffer	250 mM TRIS 1.92 M Glycine
10x WB buffer	100 mM TRIS-HCl pH 7.5 1.5 M NaCl

	(Add 0.1% v/v TWEEN® 20 when 10x WB diluted to 1x)
10x SSC buffer	300 mM sodium citrate, pH 7.0 1 M NaCl
5x HB	100 mM TRIS-HCl pH 8.0 0.5 mg/ml BSA 0.5 M NaCl 5 mM DTT
5x Stop buffer	20 mg/ml proteinase K 5% SDS (w/v) 0.5 M EDTA
TIRF Imaging buffer	6% (v/v) glucose 0.1 mg/ml glucose oxidase 0.02 mg/ml glucose catalase 1 mM Trolox 1 mM ATP 1 mM $MgCl_2$

Table 2.5: Media used as when referenced throughout. All media was sterilized by autoclaving before use.

Media	**Supplier**	**Composition**
LB	Melford	Tryptone 10 g/L NaCl 5 g/L Yeast extract 5 g/L
LB agar	Sigma	LB broth with agar (Miller)
Power Broth	Molecular Dimensions	Proprietary
Overnight Express Instant TB medium	Novagen	Proprietary; this media was prepared according to the manufacturer's instructions.
Insect-XPRESS™ Protein-free Insect Cell Medium with L-glutamine	Lonza	Proprietary

2.2. Commercial enzymes

All enzymes used were sourced from NEB unless stated otherwise.

Table 2.6: Commercial enzymes used in molecular cloning and DNA manipulations. These were used as referred to in the text.

Enzyme	Other
XhoI restriction endonuclease	
DpnI restriction endonuclease	Acts on methylated DNA only
KpnI restriction endonuclease	
Q5 DNA polymerase	
Phusion Flash DNA polymerase	Thermo Fisher Scientific
Vent DNA polymerase	
KOD Xtreme DNA Polymerase	Sigma Aldrich
T4 DNA ligase	
Antarctic phosphatase	
T4 Polynucleotide kinase	

2.3. General protein purification procedures using FPLC

Table 2.7: FPLC chromatography columns used as referred to in the main text. Columns were from GE Healthcare.

FPLC chromatography column	Type
5 mL His-trap	Immobilised metal affinity chromatography
1 mL Q-sepharose	Anionic exchange chromatography
1 mL and 5 mL heparin	Affinity chromatography
Superdex 200 24/ 60	Size exclusion chromatography
Superose 12 10/30	Size exclusion chromatography

Protein purification exploits the chemical and physical properties of the protein(s) to be purified. To enable easier protein purification, tags may be engineered onto the N- or C- terminus of a protein using molecular cloning. Commonly used affinity tags include hexahistidines (His), streptactin (strep), glutathione-S-transferase (GST), and maltose binding protein (MBP). Some of these tags, e.g. MBP are noted for enhancement of protein solubility, and other tags, e.g. green fluorescent protein (GFP), confer properties such as fluorescence for downstream applications. Here, protein purification approaches used during this project are described.

All soluble proteins overexpressed in this project, (Rad51, HelQ, Rad51B, Rad51C, Rad51D and XRCC2) carried a hexahistidine (His) affinity tag and/or a Small Ubiquitin-like MOdifier (SUMO) affinity tag to aid protein purification and enhance protein solubility respectively. His-tagged proteins are typically purified using immobilized metal affinity chromatography (IMAC). This relies on immobilized metal ions, usually Ni^{2+} or Co^{2+}, which are bound to resin *via* nitrilotriacetic acid (NTA), to which histidine rich protein regions bind. Large numbers of histidine residues are not typically found in native proteins; therefore most untagged proteins from clarified cell lysate will not bind to the Ni^{2+}-nitrilotriacetic acid (NTA) column, although some non-specific binding to the column cannot be avoided. His-tagged proteins become enriched on the column and are eluted from the column using imidazole, a molecule structurally similar to histidine that has a stronger affinity to the metal

ions immobilized on the column than the His-tag. Therefore, imidazole outcompetes the His-tagged proteins and they elute from the column, typically using up to 1 M imidazole. Majority of non-specifically bound proteins can washed off the IMAC column using a low concentration (e.g. 10 mM) of imidazole.

Ion exchange chromatography relies on the isoelectric point (pI) of a protein, which is most often determined from the amino acid sequence as a theoretical value. The net charge on a protein can be altered by varying buffer pH, which requires careful consideration of buffers because they are effective as buffers at different pH ranges. Q- and diethylaminoethanol (DEAE)- sepharose columns have anionic, positively charged matrices, whereas S-columns have a cationic, negatively charged matrix. These are utilized for binding of negatively and positively charged proteins respectively.

In some cases, other protein-specific properties can be exploited for purification, for example a protein's ability to bind nucleotide cofactors or DNA. Heparin, a type of affinity chromatography, contains a matrix of sulfated polysaccharides that mimic the phosphodiester backbone of DNA and is therefore suitable for enrichment of many DNA binding proteins and was used extensively during all protein purifications described in the appropriate sections of this thesis.

Proteins can be separated by their size and shape by size exclusion chromatography, although the resolution of protein separation that can be achieved depends on several properties of the types of column used.

Resins used for molecular sieving consist of different polymers such as sephacryl, superdex and sepharose, and each possesses different properties that affect non-specific binding of proteins thereby affecting the rate of protein mobility through the column, which determines the effective resolution of any size exclusion chromatography procedure. The rate at which proteins pass through columns is also affected by the pore size of the resin.

Differential protein precipitation ("salting out") can be used as a means of protein purification by exploiting the tendency of different proteins to precipitate out of solution by hydrophobic interactions typically achieved in response to increasing ammonium sulfate concentration in aqueous solution. Ammonium sulfate is typically added to a solution to achieve desired percentage saturation. This method can be used to concentrate proteins or in buffer exchange by, respectively, re-suspension of the precipitated protein into a small volume of buffer or into a different buffer. Precipitated proteins stored as a pellet after salting out usually remain stable and therefore this can be used as a means of storage.

2.4. Oligonucleotides

All oligonucleotides used for primers or for constructing model DNA substrates were sourced from Eurofins MWG Operon or Sigma and were supplied lyophilized, before dilution in sterile distilled water (SDW) to 100 μM. Oligonucleotides are listed in Table 2.9.

2.5. Primers

Table 2.8 below lists primers used in polymerase chain reactions for (a) generating DNA for cloning into plasmids for protein over-expression at The OPPF, (b) verifying DNA sequences of cloned genes, (c) site-directed mutagenesis using the NEB Q5 method *via* NEB-BaseChanger. Listed are primer pairs for each gene cloned, in each case (1). is the forward primer for amplification from the gene start, and (2). is the reverse primer. Also given is the name of the plasmid into which each gene was ultimately cloned from each primer pair. All primers used for cloning into pOPIN vectors at OPPF-UK were designed by Dr L. Bird (OPPF-UK).

Table 2.8. List of DNA primer sequences used for cloning in this work, where genes were amplified from GeneArt delivery plasmids unless otherwise stated. Appendices at the end of this thesis provide further sequences and information about GeneArt constructs that were used as templates for PCR reactions.

Gene/ plasmid	**Oligonucleotide sequences 5'-3'**
rad51b/ pOPINE	1. AGGAGATATACCATGGGTAGCAAAAAACTGAAACGTGTTGGTCTG 2. GTGATGGTGATGTTTTTAAAAGATCAGCTGGGTATGATGACACATCTGC
rad51b/ pOPINF	1.AAGTTCTGTTTCAGGGCCCGGGTAGCAAAAAACTGAAACGTGTTGGTCTG 2.ATGGTCTAGAAAGCTTTAAAAGATCAGCTGGGTATGATGACACATCTGC
rad51b/ pOPINS3C	1.AAGTTCTGTTTCAGGGCCCGGGTAGCAAAAAACTGAAACGTGTTGGTCTG 2.ATGGTCTAGAAAGCTTTAAAAGATCAGCTGGGTATGATGACACATCTGC
rad51c/ pOPINRSE	1.AGGAGATATACCATGCGTGGTAAAACCTTTCGTTTTGAAATGCAG 2.GTGATGGTGATGTTTTTACAGTTCCTCTTCCGGATCACGGC
rad51c/ pOPINE	1. AGGAGATATACCATGCGTGGTAAAACCTTTCGTTTTGAAATGCAG 2. GTGATGGTGATGTTTTTACAGTTCCTCTTCCGGATCACGGC
rad51c/ pOPINF	1. AAGTTCTGTTTCAGGGCCCGCGTGGTAAAACCTTTCGTTTTGAAATGCAG 2. ATGGTCTAGAAAGCTTTACAGTTCCTCTTCCGGATCACGGC
rad51c/ pOPINS3C	1.AAGTTCTGTTTCAGGGCCCGCGTGGTAAAACCTTTCGTTTTGAAATGCAG 2. ATGGTCTAGAAAGCTTTACAGTTCCTCTTCCGGATCACGGC
rad51d/ pOPINCDE	1.AGGAGATATACCATGGGTGTTCTGCGTGTTGGTCTGTGTC 2.GTGATGGTGATGTTTTTAGGTCTGATCACCCTGCAGTGTTG
rad51d/ pOPINE	1. AGGAGATATACCATGGGTGTTCTGCGTGTTGGTCTGTGTC 2. GTGATGGTGATGTTTTTAGGTCTGATCACCCTGCAGTGTTGC
rad51d/ pOPINF	1. AAGTTCTGTTTCAGGGCCCGGGTGTTCTGCGTGTTGGTCTGTGTC 2. ATGGTCTAGAAAGCTTTAGGTCTGATCACCCTGCAGTGTTGC
rad51d/ pOPINS3C	1. AAGTTCTGTTTCAGGGCCCGGGTGTTCTGCGTGTTGGTCTGTGTC 2. ATGGTCTAGAAAGCTTTAGGTCTGATCACCCTGCAGTGTTGC
xrcc2/ pOPINE	1. AGGAGATATACCATGTGTAGCGCATTTCATCGTGCAGAAAGC 2. GTGATGGTGATGTTTGCAAAATTCCACACCGCTTTCACCG
xrcc2/ pOPINF	1. AAGTTCTGTTTCAGGGCCCGTGTAGCGCATTTCATCGTGCAGAAAGC 2. ATGGTCTAGAAAGCTTTAGCAAAATTCCACACCGCTTTCACCG
xrcc2/ pOPINS3C	1. AAGTTCTGTTTCAGGGCCCGTGTAGCGCATTTCATCGTGCAGAAAGC 2. ATGGTCTAGAAAGCTTTAGCAAAATTCCACACCGCTTTCACCG
xrcc3/ pOPINE	1. AGGAGATATACCATGGATCTGGATCTGCTGGACCTGAATCC 2. GTGATGGTGATGTTTTTAATGGCTCTGTGTACCCGGTGTACCAC

xrcc3/ pOPINF	1. AAGTTCTGTTTCAGGGCCCGGATCTGGATCTGCTGGACCTGAATCC 2. ATGGTCTAGAAAGCTTTAATGGCTCTGTGTACCCGGTGTACCAC
xrcc3/ pOPINS3C	1. AAGTTCTGTTTCAGGGCCCGGATCTGGATCTGCTGGACCTGAATCC 2. ATGGTCTAGAAAGCTTTAATGGCTCTGTGTACCCGGTGTACCAC
xrcc3/ pOPINRSE	1. AGGAGATATACCATGGATCTGGATCTGCTGGACCTGAATCC 2. GTGATGGTGATGTTTTTAATGGCTCTGTGTACCCGGTGTACCAC
rad52/ pOPINE	1. AGGAGATATACCATGGCCGGTACGGAAGAAGCAATTCTG 2. GTGATGGTGATGTTTTTAGCTCGGGTCATATTTGCGTTTTTTCATAT
rad52/ pOPINF	1. AAGTTCTGTTTCAGGGCCCGGCCGGTACGGAAGAAGCAATTCTG 2. ATGGTCTAGAAAGCTTTAGCTCGGGTCATATTTGCGTTTTTTCATATC
rad52/ pOPINRSE	1. AAGTTCTGTTTCAGGGCCCGGCCGGTACGGAAGAAGCAATTCTG 2. ATGGTCTAGAAAGCTTTAGCTCGGGTCATATTTGCGTTTTTTCATATC
helq/ pOPINE	1. AGGAGATATACCATGGATGAATGTGGTAGCCGTATTCGTCGTCG 2. GTGATGGTGATGTTTTTATGCTTTATCGGTTGAGCTTGCAACTGC
Helq/ pOPINF	1. AAGTTCTGTTTCAGGGCCCGGATGAATGTGGTAGCCGTATTCGTCGTCG 2. ATGGTCTAGAAAGCTTTATGCTTTATCGGTTGAGCTTGCAACTGC
Helq/ pOPINS3C	1.AAGTTCTGTTTCAGGGCCCGGATGAATGTGGTAGCCGTATTCGTCGTCG 2.ATGGTCTAGAAAGCTTTATGCTTTATCGGTTGAGCTTGCAACTGC
helq core helicase (amplified from pSN50)/ pBADHisA	1. CTGGGTCCGCTCGAGTTTTATAGTCTGCCGAGCAAAG 2. GTGATGTTTGGTACCTTATGCTTTATCGGTTGAGC

T7 forward sequencing primer	TAATACGACTCACTATAGGG
pTriEx-Down sequencing primer	TCGATCTCAGTGGTATTTGTG
pBAD forward sequencing primer	ATGCCATAGCATTTTTATCC
Mutagenic K133R Rad51 (designed for use with pEBPS2)	1. TCGTACCGGCCGTACCCAGATTTG 2. AATTCACCAAACATTTCGG
Mutagenic D222A Rad51 (designed for use with pEBPS2)	1. GCTGATTGTTGCTAGCGCAACCG 2. AATTCACCAAACATTTCGG

Table 2.9. List of oligonucleotides used to form for model DNA substrates in this work

Oligonucleotides used for model DNA substrates. '*' Indicates oligonucleotides that were either fluorescently labeled or radiolabeled at the 5' end.

Substrate	Description	Oligonucleotide name	Oligonucleotide sequence 5'-3'
ssDNA		ELB 58*	TCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTC GCCATCAACT AACGATTCTGTCAAAA ACTGACGCGT
Linear duplex	Linear duplex	ELB 40*	GGAGCTCCCTAGGCAGGATCGTTCGCGACGATGGCCTTCGAAGAGCTCCAGTTACGGATACGGAT CCTGC
		ELB 41	GCAGGATCCGTATCCGTAACTGGAGCTCTTCGAAGGCCATCGTCGCGAACGATCCTGCCTAGGGA GCTCC
Fork 2	Splayed duplex	MW 14*	CAACGTCATAGACGATTACAGTTCTACATGGAGCTGTCTAGAGGATCCGA
		MW 12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC
Fork 3	Leading strand only model replication fork	MW 14*	CAACGTCATAGACGATTACAGTTCTACATGGAGCTGTCTAGAGGATCCGA
		MW 12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC
		PM 16	TGCCGAATTCTACCAGTGCCAGTGAT
Fork 4	Lagging strand only model replication fork	MW 14*	CAACGTCATAGACGATTACAGTTCTACATGGAGCTGTCTAGAGGATCCGA
		MW 12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC
		PM 17	TAGCAATGTAATCGTCTATGACGTTG
Fork 5	Fully base paired model replication fork	MW 14*	CAACGTCATAGACGATTACAGTTCTACATGGAGCTGTCTAGAGGATCCGA
		MW 12	TCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC
		PM 16	TGCCGAATTCTACCAGTGCCAGTGAT
		PM 17	TAGCAATGTAATCGTCTATGACGTTG
D-Loop 1	D-loop with untailed invading strand	RGL 19*	GACGCTGCCGAATTCTACCAGTGCCTTGCTAGGACATCTTTGCCCACCTGCAGGTTCACCC
		PM 4	CTGCGACGGCTTAAGATGGTTGATTGGATGCTACTCGTCGGCGGGTGGACGTCCAAGTGGG

		PM 5	AAAGATGTCCTAGCAAGGCAC
D-loop 2	D-loop with 3' tailed invading strand	RGL 19*	GACGCTGCCGAATTCTACCAGTGCCTTGCTAGGACATCTTTGCCCACCTGCAGGTTCACCC
		PM 4	CTGCGACGGCTTAAGATGGTTGATTGGATGCTACTCGTCGGCGGGTGGACGTCCAAGTGGG
		PM 7	AAAGATGTCCTAGCAAGGCACGATCGACCGGATATCTATGA
		RGL 19*	GACGCTGCCGAATTCTACCAGTGCCTTGCTAGGACATCTTTGCCCACCTGCAGGTTCACCC
		PM 4	CTGCGACGGCTTAAGATGGTTGATTGGATGCTACTCGTCGGCGGGTGGACGTCCAAGTGGG
		PM 6	TAAGAGCAAGATGTTCTATAAAAGATGTCCTAGCAAGGCAC

2.6. Use of GeneArt® customized DNA synthesis

Human genes encoding Rad51, Rad52, Rad54, Rad51B, Rad51C, Rad51D, XRCC2, XRCC3 and HelQ were codon optimized for expression in *E. coli* and synthesized by GeneArt®. Synthetic genes were supplied in a generic pMA-T plasmid. Please refer to the appendices at the end of this thesis for further details of the GeneArt projects and sequences that were synthesized.

2.7. General microbiology

2.7.1. Electro-competent *E. coli*

LB (500 ml) containing appropriate antibiotic selection was inoculated at 1/100 from an overnight starter culture and incubated at 37^{o}C with shaking until an OD of 0.4-0.6 was reached. Cells were harvested at 3,000 *g* for 20 minutes at 4^{o}C in pre-chilled centrifuge flasks. Supernatant was discarded and cells were resuspended in 200 ml pre-chilled sterile distilled water (SDW) by swirling. Cells were harvested at 6,000 *g* for 12 minutes and resuspended as before. This was repeated but cells resuspended in 20 ml and then 6 ml SDW. Glycerol was added to 30% v/v and cells aliquoted to 100 µl and flash frozen in liquid N_2 for storage at -80^{o}C.

2.7.2. Transformation of electro-competent *E. coli*

Plasmid DNA (typically 250 ng of DNA from a plasmid "mini-prep") was added to $6x10^9$ electro-competent *E. coli* and incubated on ice for two minutes. Cells were electro-pulsed using the EC1 (1.8 kV) setting on a Bio-Rad Micro-Pulser™, rapidly resuspended in 1 ml LB and incubated at 37^oC with shaking for 1 hour. 50 µl culture was spread onto agar plates containing appropriate antibiotic selection and incubated overnight at 37^OC.

2.8. Molecular cloning and routine DNA cloning and manipulation

2.8.1. Purification of plasmid DNA

Plasmid DNA was purified from transformed *E. coli* cloning strains such as XL1 Blue or DH5α. Cultures (10 ml) containing appropriate antibiotic selection were inoculated with a single colony and incubated overnight at 37^oC with shaking at 180 rpm. Cells were harvested from 5 mL culture using a Heraeus Biofuge primo R benchtop centrifuge. "Mini-prep" purification of plasmid DNA was carried out using an alkaline lysis[217] based method of DNA extraction using a GeneJET Plasmid Mini-prep Kit (Thermo Fisher Scientific) according to the manufacturer's instructions. This involved a series of steps for cell lysis, neutralization of the DNA containing supernatant and column wash steps before elution of purified plasmid DNA in SDW. Concentration and quality of purified DNA was assessed using a ND 1000 Spectrophotometer NanoDrop. This quantifies

concentration of nucleic acids by reading absorption at 260 nm. The Beer-Lambert law is applied to this value to work out nucleic acid concentration in ng/ µl. A reading at 280 nm measures protein content in the sample, to calculate the A260:280 ratio that determines the quality of the sample. A260:280 ratios of approximately 1.8 is indicative of a good quality DNA sample with minimal protein contaminants.

2.8.2. PCR product purification

PCR products were purified using a GeneJET PCR Purification Kit (Thermo Scientific) according to the manufacturer's instructions. This primarily involved a series of column wash steps to remove remaining components of the PCR before elution of purified DNA into SDW.

2.8.3. Agarose gel electrophoresis

1% agarose was dissolved in 1x TAE. 1x SYBR Safe DNA gel stain (Invitrogen) or ethidium bromide 10 µg/ml was added to visualize DNA. Electrophoresis was carried out in 1x TAE at 100 V at constant current for 50 minutes using a BioRad Mini-Sub Cell GP electrophoresis tank. DNA was visualized using a Syngene G:Box gel imager and GeneSnap software.

2.8.4. DNA sequencing

Sanger sequencing was used to verify gene sequences, using standard primers for the plasmid as specified. Due to the large size of the HelQ gene, internal sequencing primers were designed for sequencing,

listed in Table 2.8. DNA sequence alignments were carried out using A Plasmid Editor (APE) software.

2.9. Amplification and cloning of *rad51* and *helQ*

GeneArt® custom DNA synthesis, see section 2.6, was used to generate several of the gene constructs described in this thesis; custom gene sequences and the associated GeneArt® projects are detailed in appendices at the end of this thesis. Some other cloning or mutagenesis procedures for Rad51 and HelQ that utilized PCR are described here.

2.10. PCR based generation of *rad51* mutants

To generate Rad51 mutants, PCR reactions of 50 µl were assembled to contain approximately 100 ng template DNA (pEBPS2), 1x Q5 reaction buffer, 200 µM dNTPs, 0.5 µM forward and reverse primers listed in Table 2.8 and 0.02 U/ µl Q5 DNA polymerase. PCR products for K122R and D222A Rad51 mutants were generated using the following primer sequences specific to pEBPS2, and PCR cycles detailed in Tables 2.10 and 2.10A, using a MultiGene™ Mini thermocycler (Labnet International, Inc.). Q5 mutagenesis (New England Biolabs) was used to generate K133R using primers designed by NEBase Changer for the Q5 mutagenesis kit (New England Biolabs) listed in Table 2.8.

Table 2.10: PCR cycle used to generate HsaRad51 mutant K133R. Stages and temperatures highlighted in bold were repeated for 25 cycles.

PCR stage	**Temperature**	**Time**
Initial denaturation	98°C	30 seconds
Denaturation	**98°C**	**5 seconds**
Annealing	**58°C**	**30 seconds**
Extension	**72°C**	**2 minutes 45 seconds**
Final extension	72°C	2 minutes

Q5 mutagenesis (New England Biolabs) was used to generate His-HsaRad51 D222A using primers designed by NEBase Changer (New England Biolabs) for the Q5 mutagenesis kit (New England Biolabs) listed in Table 2.1.10.

Table 2.10A: PCR cycle used to generate HsaRad51 mutant D222A. Stages and temperatures highlighted in bold were repeated for 25 cycles.

PCR stage	**Temperature**	**Time**
Initial denaturation	98°C	30 seconds
Denaturation	**98°C**	**5 seconds**
Annealing	**64°C**	**30 seconds**
Extension	**72°C**	**2 minutes 45 seconds**
Final extension	72°C	2 minutes

K133R, D222A double mutant was generated using K133R Rad51 template and D222A mutagenic primers and PCR cycle.

Linear PCR products using NEBase Changer (New England Biolabs) primers were subjected to treatment with the Q5 mutagenesis kit (New England Biolabs) according to the manufacturer's protocol. Briefly, this, consists of treatment of the PCR product with 'KDL' mix that consists of

enzymes that catalyse degradation of methylated DNA by Dpn1, phosphorylation of DNA ends by T4 PNK and ligation by T4 DNA ligase, before transformation into an *E. coli* cloning strain.

2.11. PCR amplification of *helq* core helicase gene

DNA encoding the HelQ core helicase was PCR amplified for sub-cloning into pBADHisA. PCR reactions of 50 µl were assembled to contain approximately 100 ng template DNA (pSN50), 1x Q5 reaction buffer, 200 µM dNTPs, 0.5 µM forward and reverse primers and 0.02 U/µl Q5 DNA polymerase. PCR products containing upstream *XhoI* and downstream *KpnI* restriction sites were generated using primers detailed in table 2.8 and PCR cycle detailed in Table 2.11 using a MultiGene™ Mini thermocycler (Labnet International, Inc.).

Table 2.11: PCR cycle used to amplify the *helq* core helicase DNA fragment to contain *Xho1* and *Kpn1* restriction sites. Stages and temperatures highlighted in bold were repeated for 25 cycles.

PCR stage	**Temperature**	**Time**
Initial denaturation	98°C	30 seconds
Denaturation	**98°C**	**5 seconds**
Annealing	**64°C**	**30 seconds**
Extension	**72°C**	**1 minute 15 seconds**
Final extension	72°C	2 minutes

2.12. Cloning of Hel308 and HelQ WHD for over-expression of isolated WHD protein

DNA sequences encoding wild type WHDs from MthHel308 and human HelQ were predicted from sequence comparisons of those proteins with the sequence of AfuHel308, for which there is a crystal structure. Phyre 2 and Pymol software were used to generate molecular models of Hel308 and HelQ. Given below is the resulting amino acid sequence of MthHel308 that was used for generating an N-terminally His-tagged WHD protein:

(M)GEVERTTSRIIENRDALYRQIIAQVASGLSGTTEELADF**F**RNT**F**YGYQMVEGPFS DSFGMDSI**QY**EVE**N**ATEYLMRNRILYP(STOP)

In addition, given below is the amino acid sequence of human HelQ that was determined by using the same sequence alignment process, to generate isolated N-terminally His-tagged human HelQ WHD protein:

(M)LILQEKDKQQVLELITKPLENCYSHLVQEFTKGIQTLFLSLIGLKIATNLDDIYHFM NGTFFGVQQKVLLKEKSLWEITVESLRYLTEKGLLQKDTI**YK**SEEEV**QY**NFHITKLGR ASFKGTIDLAYCDILYRDLKKGLEGLVLESLLHLIYLTTPYDLVSQ(STOP)

In each case, codons for an N-terminal methionine (M) and STOP were introduced into the GeneArt DNA synthesis for cloning, and in bold underlined are amino acids focused on for mutagenesis of these domains.

Mth and human Hel308/HelQ WHD encoding DNA was synthesized by GeneArt and delivered in plasmids (Thermo Fisher) with codons

optimised for protein expression in *E. coli*. Coding sequences for each WHD from the GeneArt plasmids were sub-cloned from these constructs into pET14b *via Nde1* and *BamHI* restriction sites for expression of N-terminally hexahistidine $(His)_6$ tagged WHDs. The Q5 site-directed mutagenesis system (New England Biolabs) was used to generate mutations for all amino acid substitutions described. Archaeal Hel308 with the WHD deleted and replaced by a 22 amino acid "linker" has the wild type WHD sequence given above changed to the following sequence: PSVAVEVAPGVPAVEEGAVPAV.

The *recQ* helicase gene was cloned as described in Guy and Bolt 2005[5], for use in genetic analysis.

Table 2.12. List of plasmids used in the analysis of winged helix domains (Chapter 5).

Name	Description	Reference
pEB310	Mth*hel308* cloned into pT7-7 for genetic analysis *in E. coli.*	(5)
pEB431	pET14b cloned *Mthhel308* for protein expression and purification.	(5)
pEB422	Mutant Mth*hel308* K51L, created by mutagenesis of pEB310.	(5)
pEB417	*E. coli recQ* cloned into T7-7 for genetic analysis in *E. coli*.	(5)
pEB622	pET14b cloned Mth*hel308* winged helix domain. Created from sub-cloning a synthetic sequence from GeneArt.	This work
pEB643	pET14b cloned full length Mth*hel308* winged helix domain mutant WHD^{GAG}. Created from Q5	This work

	mutagenesis of pEB622.	
pEB644	pET14b cloned Mth*hel308* winged helix domain mutant WHD$^{\Delta\alpha20}$. Created from Q5 mutagenesis of pEB622.	This work
pEB664	pET14b cloned human *helQ* winged helix domain. Created from sub-cloning a synthetic sequence from GeneArt.	This work
pEB665	pET14b cloned human *helQ* winged helix domain mutant Q-WHD^{SD}. Created from Q5 mutagenesis of pEB664	This work
pEB666	pET14b cloned human *helQ* winged helix domain mutant Q-WHD^{SS}. Created from Q5 mutagenesis of pEB664	This work
pEB475	pT7-7 cloned mutant Mth*hel308* F434V, created by Q5 mutagenesis of pEB431.	This work
pEB477	pET14b cloned mutant Mth*hel308* F434V, created by sub-cloning from pEB475.	This work
pEB444	pT7-7 cloned mutant Mth*hel308* F439A, created by Q5 mutagenesis of pEB431.	This work
pEB447	pET14b cloned mutant Mth*hel308* F434V, created by sub-cloning from pEB475.	This work

2.13. High throughput DNA cloning and manipulation carried out at the Oxford Protein Production Facility – UK (OPPF-UK)

To assist with protein expression trials for subsequent protein purification of human HelQ and Rad51 paralogues, high throughput cloning was carried out at OPPF-UK into vectors suitable for protein expression in *E. coli* and Sf9 insect cells, described below and in Chapter 3.

2.13.1. Amplification and cloning of human *helQ* and *rad51* paralogues

High throughput PCR amplification of genes encoding human HelQ and Rad51 paralogues used plasmid constructs generated by GeneArt custom synthesis as DNA templates. Details of the gene sequences from these constrcuts are provided in appendices at the end of this thesis. Reactions (25 µl) containing 20 ng template DNA, 0.5 µM forward and reverse primers and 1x Phusion Flash master mix (Thermo Fisher Scientific) that contained reaction buffer, Phusion Flash polymerase and dNTPs, in a 96 well plate format. PCR products for all templates were generated using the cycle detailed in Table 2.13 in a Veriti Thermal Cycler (Applied Biosystems). Primers for Ligation Independent Cloning are listed in Table 2.8.

Table 2.13: PCR cycles used to amplify *rad51B, rad51C, rad51D, xrcc2 and helq* for LIC. Stages and temperatures highlighted in bold were repeated for 25 cycles.

PCR stage	**Temperature**	**Time**
Initial denaturation	98°C	10 seconds
Denaturation	**98°C**	**1 second**
Annealing	**64°C**	**5 seconds**
Extension	**72°C**	**1 minute**
Final extension	72°C	1 minute

2.13.2. *Dpn*1 digestion and PCR product purification

Completed PCR mutagenesis reactions were treated with Dpn1 to remove methylated wild-type template DNA. PCR products in 96-well plates were purified from unincorporated dNTPs, DNA polymerase and

other PCR components by 5 min room temperature incubation of 90 µl (per 50 µl of PCR reaction) of Agencourt AMPure XP magnetic bead solution (Beckman Coulter), followed by transfer to a magnetic block and allowed to stand for a further 5 minutes. Clear supernatant was removed from the wells and discarded. Magnetic beads were washed with 200 µl ethanol and incubated for 30 seconds at RT before ethanol was removed and discarded. This step was repeated and beads were allowed to air dry for 10 minutes, and then resuspended in 30 µl elution buffer. The reaction plate was transferred back onto the magnetic block and eluate was transferred into a fresh 96 well plate.

2.13.3. InFusion™ cloning of PCR products into pOPIN vectors for protein expression in *E. coli* and in insect cells.

PCR products were cloned into pOPIN vectors, to create a series of vectors carrying *helq* and *rad51* paralogue genes with a variety of tags to aid solubility and purification, using InFusion™ Cloning (Clontech), a type of ligation independent cloning (LIC). pOPIN vectors were designed by OPPF-UK and are based on pTriEx vectors that are suitable for protein expression in *E. coli*, Sf9 or mammalian cell based systems either directly, or after further cloning procedures. To enable this, each pOPIN vector therefore carries elements that are recognized by each expression system, including a baculovirus recombination region for cloning to generate baculovirus for insect cell protein expression; a chicken β actin promoter and rabbit globin terminator for protein expression in mammalian cells;

and finally T7 promoter and terminator sequences for protein expression in *E. coli*. Each pOPIN vector also carries sequences that encode HRV-3C cleavable N or C terminal tags onto proteins to aid protein purification or enhance protein solubility. Cloning vectors can be linearized by cutting within *lacZ* using a restriction enzyme (ClonTech).

For cloning into pOPIN vectors, PCR products (2 µl) were transferred into corresponding wells of a fresh 96 well plate, for mixing with 7 µl SDW and 100 ng of linearized plasmid. This reaction mixture was transferred to tubes containing lyophilised InFusion™ enzyme and incubated for 5 minutes before transfer back into the 96 well plate. Reactions were incubated for 30 minutes at 42^{o}C and were terminated by addition of 20 mM EDTA.

Infusion^{TM} cloning (ClonTech) is a form of ligation independent cloning (LIC) for annealing of an insert gene into a vector backbone. A vector is linearised using a restriction enzyme, and is treated with T4 DNA polymerase to utilize its 5' to 3' exonuclease activity to generate single stranded overhangs. This is carried out in the presence of only one dNTP, and exonuclease activity continues until a base complemantary to the selected dNTP is reached. Primers used to amplify each gene carry extension sequences that are complementary to the ends of the treated linearized vector, facilitating annealing of the insert and vector. Gaps are sealed *in vivo* after transformation into *E. coli*. This particular method of cloning is suited to the high throughput 96 well format used at OPPF-UK enabling rapid cloning of multiple constructs.

2.13.4. Detection of successful recombinant plasmid clones *via* insertional inactivation of *lacZ*

Cloning of PCR products into *lacZ* of pOPIN vectors allows blue-white screening of colonies to identify successful clones by plating transformed *E. coli* onto agar medium containing X-gal (40 µg/ml) and 1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG). *lacZ* encodes β-galactosidase, an *E. coli* protein that hydrolyses X-gal into 5-bromo-4-chloro-indoxyl, resulting in blue colonies from intact *lacZ* or white colonies indicating no expression of β-galactosidase as a result of disruption of *lacZ*.

2.13.5. High throughput purification of plasmid DNA

For this procedure, plasmid DNA was purified using an automated high-throughput miniprep procedure using a Qiagen BioRobot 8000 with Qiagen reagents. This technique is based upon the alkaline lysis method of DNA extraction, and follows the same steps as described in section 2.8.1, Purification of Plasmid DNA.

2.14. Protein expression

With the exclusion of HelQ and Rad51, detailed later, protein overexpression from *E. coli* for purification was optimized *via* pilot over expression tests at OPPF-UK. Procedures for assessing over-expression of each protein are described individually below.

2.14.1. Over-expression of His-HsaRad51 in *E. coli*

To over-express N-terminally hexahistidine tagged Rad51, plasmid pEBPS2 encoding His_6-Rad51 was transformed into chemically competent BL21 AI *E. coli*. pEBPS2 is a T7 based expression vector based on pBad-HisA. In this plasmid the *ara*BAD promoter (P_{BAD}) is under tight control by AraC to prevent leaky basal protein expression by the repressor binding and releasing the promoter in equilibrium. Repression of the promoter (P_{BAD}) by AraC is relieved by addition of L-arabinose thereby enabling transcription of downstream genes, therefore expression can be tightly controlled; this is useful for expression of toxic proteins. Here, this works in concert with the BL21 AI *E. coli* protein expression strain. Together, this allowed inducible over-expression of Rad51. A single colony was picked to inoculate an overnight starter culture containing 50 μg/ml carbenicillin. LB (1 L) containing 50 μg/ml carbenicillin was inoculated at 1/100 from the overnight starter culture, and grown until an OD (at λ595 nm) of 0.4-0.6 was reached. Protein expression was induced using 0.2% L-arabinose and expressed for 3 hours. Cells were harvested at 4,000 *g* using a Beckman Avanti J-26 XP centrifuge in rotor JA-10.5 for 20 minutes then flash frozen in liquid N_2 for storage at -80^{o}C.

2.14.2. Over-expression of MthHel308 proteins in *E. coli*

Archaeal MthHel308 wild type and mutant proteins were over-expressed in *E. coli* strains BL21 AI using the same general procedure as described in Guy and Bolt, 2005[5].

2.14.3. Over-expression of MthHel308 and HsaHelQ WHDs in *E. coli*

Isolated WHD proteins from Mth and human were over-expressed and purified in the same way as for full length MthHel308, in *E. coli* strain BL21 AI[5].

2.14.4. Over-expression and optimization of human Rad51 paralogue proteins in *E. coli* at OPPF-UK

Over-expression of Rad51 paralogue proteins was trialed at OPPF-UK, using Lemo21(DE3) and Rosetta 2 (DE3) pLacI *E. coli* expression strains and different *E. coli* culture media. Starter cultures were prepared by inoculating plasmid-transformed *E. coli* cells into 1 ml broth containing appropriate antibiotic in 96 well blocks. The 96 well format was continued for downstream high throughput applications. Cells were grown overnight with shaking at 37°C. In addition, 3 ml cultures were set up in 36 well format using the appropriate selection markers in duplicate using different types of media: Power Broth™ (Molecular Dimensions) and Overnight Express™ Instant TB medium (Novagen). Overnight Express™ Instant TB medium relies upon different rates of metabolism of sugars and other nutrients before metabolism of IPTG by *E. coli* to induce protein expression 'automatically'. Power Broth™ contains higher levels of vitamins and amino acids than LB, and contains buffers to maintain a pH of 6.8 ±0.2. Usual working antibiotic concentrations were reduced by half for co-transformants containing two or more plasmids carrying different

antibiotic resistance markers. Transformed Lemo21(DE3) and Rosetta 2 (DE3) pLacI overnight starter cultures were used to inoculate cultures at inoculums of 1/20 and 1/12 respectively. Cultures were incubated at 37^{o}C with shaking until OD (at 595 nm) 0.4-0.6 was reached. At this point, cultures growing in Power Broth™ media were induced using 1 mM IPTG and incubated at 25^{o}C overnight with shaking. Overnight Express™ Instant TB medium cultures were transferred to 20^{o}C and incubated overnight with shaking.

2.14.5. Micro-scale purification of His-tagged Rad51-paralogue proteins

To identify over-expression of soluble human Rad51 paralogue proteins, micro-scale purification of His-tagged proteins was carried out using Ni^{2+}-NTA magnetic beads. A 1 ml sample from each induced culture was transferred into a fresh 96 well block. Cells were harvested at 4,000 *g* for 10 minutes using a Beckman Allegra X15R centrifuge and rotor SX4750 and supernatant was discarded. Blocks were stored at -80^{o}C for 30 minutes to aid cell lysis. High throughput micro-scale purification was carried out using a Qiagen BioRobot 8,000. Pellets were resuspended in lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, 1% v/v Tween 10 and 1 mg/ml lysozyme) and incubated at RT with shaking for 30 minutes. Cell debris was pelleted by centrifugation at 5,000 *g* for 30 minutes using a using a Beckman Allegra X15R centrifuge and rotor SX4750. The supernatant was transferred to a fresh 96 well plate and 20 µl Ni^{2+}-NTA magnetic beads (Promega) were added to each well and

gently shaken for 30 minutes. The plate was transferred to a magnetic block for 1 minute and the supernatant discarded. The plate was removed from the magnetic block and the beads were washed in 200 µl wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole) and shaken gently for 5 minutes. The plate was transferred onto the magnetic block and the buffer discarded. The wash step was repeated. His-tagged proteins were eluted from the Ni^{2+}-NTA magnetic beads using 50 µl elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole). Elution of soluble His-tag and histidine rich proteins was transferred to a fresh 96 well plate.

2.14.6. Reconstitution of Rad51 paralogue protein over-expression at Nottingham

Following detection of soluble protein expression using *E. coli* based protein expression systems at OPPF-UK, this was reconstituted at Nottingham where samples of soluble and insoluble proteins were analysed. Pilot cultures (50 ml) containing appropriate antibiotic selection were inoculated 1/100 from overnight starter cultures and grown at indicated temperatures until an OD of 0.4-0.6 was reached, and protein expression was induced with 1 mM IPTG. An uninduced culture was used as a negative control. Samples (10 ml) were removed from each pilot culture at specified time intervals and cells were harvested at 4000 *g* for 10 minutes using a Heraeus Biofuge primo R benchtop centrifuge. Pellets were resuspended in 1/10 original volume resuspension buffer (50 mM

TRIS-HCl pH 7.4, 200 mM NaCl) and cells were lysed by sonication on ice using a Soniprep 150 ultrasonic disintegrator (MSE – Measuring and Scientific Equipment) at 4°C. Soluble and insoluble fractions were separated by centrifugation at 16300 *g* for 10 minutes using a Genfuge 24D microfuge (Progen) and the soluble fraction transferred to a fresh eppendorf tube. The pellet was resuspended in 1/10 original volume using 50 mM TRIS-HCl pH 7.5, 100 mM NaCl and 6 M urea. Samples were taken and SDS-PAGE loading dye added for loading onto SDS-PAGE gels containing appropriate acrylamide concentrations for resolution of the target proteins.

2.14.7. Over-expression of His-SUMO-XRCC2 in *E. coli*

Protein over-expression trials and micro-scale protein purification at OPPF-UK showed soluble His-SUMO-XRCC2 could be obtained from Lemo21(DE3) *E. coli* using Overnight Express™ Instant TB medium (Novagen). This was replicated in Nottingham and large-scale over-expression of His-SUMO-XRCC2 was carried out for protein purification. 2 L Overnight Express™ instant TB medium (Novagen) 50 µg/ml carbenicillin was inoculated 1/100 with an overnight starter culture of Lemo21(DE3) *E. coli* carrying expression vectors for His-SUMO-XRCC2. Cultures were incubated for 6 hours at 37°C, before reduction to 18°C for overnight incubation with shaking at 180 rpm. Cells were harvested at 4,000 *g* using a Beckman Avanti J-26 XP centrifuge in a JA-10.5 rotor for 20 minutes then flash frozen in liquid N_2 and stored at -80°C.

2.14.8. Over-expression of His-SUMO-HelQ and Rad51 paralogue proteins by baculovirus infection of insect cells at OPPF-UK

Protein over-expression trials of HelQ and Rad51 paralogue proteins in Sf9 cells showed soluble protein expression could be achieved for His-SUMO-XRCC2, His-SUMO-Rad51B, His-SUMO-Rad51C, His-SUMO-Rad51D and His-SUMO-HelQ proteins. This was subsequently replicated and optimized for large-scale protein over-expression for protein purification.

2.14.9. Optimisation of His-SUMO-HelQ and Rad51 paralogue overexpression in insect cells

Sf9 cells are derived from the *Spodoptera frugiperda* pupal ovarian tissue cell line IPLB-SF21-AE. They are amenable to growth in suspension and infection by *Autographa californica* multiple nucleopolyhedrovirus (AcMNPV) baculovirus to deliver genetic material for expression of mammalian proteins. Sf9 cells were grown in Insect Xpress Protein Free Media (Lonza) supplemented with 100 units penicillin and 0.1 mg/ml streptomycin solution (Sigma), 1% (v/v) Pluronic F-68 solution (Thermo Fisher Scientific) and 2.5 µg/ml Amphotericin B solution (Sigma). Cells were grown at 27°C with shaking at 170 rpm. At approximately $1x10^7$ cells/ml, cultures were 'split' to lower cell densities in fresh media to maintain the cell line for further growth. Cells were counted using a haemocytometer, and on splitting, Sf9 cells were seeded to $1x10^6$ cells/ml

in fresh pre-warmed media before splitting again when cell growth approached $1x10^7$ cells/ml.

2.14.10. Freezing Sf9 cells for long term storage

$1x10^7$ cells were harvested at 500 *g* and resuspended in 1 ml fresh media supplemented with 1% (v/v) heat inactivated fetal bovine serum (FBS) (Thermo Fisher Scientific). Cells were slowly chilled from 4**°**C to -80**°**C before storage in the liquid N_2 vapour phase in a cell bank.

2.14.11. Generating and amplifying baculovirus

Baculoviruses non-pathogenic to mammals and therefore are a safe option for delivering genetic material to insect cell lines such as Sf9 for protein overexpression. Here, AcMNPV baculovirus was used. Recombinant baculovirus is generated by means of recombination *via* DNA sequences flanking the gene of interest that are homologous to sequences that are found downstream of the polyhedrin promoter. Recombination events occur following transfection of the bacmid into Sf9 cells. For transfection, Sf9 cells were seeded at $5x10^5$ cells/ ml in a 24 well plate and allowed to adhere for 40 minutes. A mix of 250 ng bacmid, 100-500 ng vector DNA 50 µl media and 1.5 µl FuGeneHD transfection reagent was incubated at RT for 30 minutes before addition to the cells without disrupting the cell monolayer. For this work, a modification of standard protocols was used for generating recombinant viral DNA, by utilizing a bacmid that had been linearized in order to improve the efficiency of

generating recombinant virus. In this scheme (https://www.helmholtz-muenchen.de/fileadmin/PEPF/Protocols/LIC-cloning.pdf) the creation of recombinant bacmid is favoured by the inability of the original parent bacmid to re-ligate to itself without insertion for the desired fragment, hence resulting in a higher ratio of recombinant to native viral DNA.

Cells were incubated at 27°C for 6-7 days before removal of viral supernatant, P0, for storage at 4°C. For stage one viral amplification, Sf9 cells were seeded at $1x10^6$ cells/ml in a 24 well plate and allowed to adhere for 40 minutes. 5 µl P0 virus stock was added to the cells and was incubated at 27°C for 6-7 days. Viral supernatant was removed for storage at 4°C. For viral amplification stage 2, cultures (50 ml) at $1x10^6$ were infected with 400 µl P1 virus, and incubated with shaking at 27°C for 6-7 days. Cells were harvested at 1,000 *g* using a Heraeus Biofuge primo R benchtop centrifuge and viral supernatant P2 was passed through a sterile 0.45 µm syringe filter (Sartorious Stedim). Heat inactivated FBS (Thermo Scientific) was added to 10% (v/v) before storage at 4°C.

2.14.12. Optimisation of Sf9 HelQ and Rad51 paralogue protein expression.

Following detection of positive soluble protein expression of His-SUMO-HelQ and His- tagged Rad51 paralogue proteins in Sf9/ baculovirus systems at OPPF-UK, over-expression of these proteins was optimized. This was achieved by titrating virus into cultures containing 600,000 cells

and expressing protein from 0-4 days. Virus was titrated 0.5 µl/6x10^5 cells to 10 µl/6x10^5 cells for His-SUMO-HelQ and 1 µl/6x10^5 cells to 40 µl /6x10^5 cells for Rad51 paralogue proteins.

2.14.13. General procedures for over-expression of His-SUMO-HelQ and His-tagged Rad51 paralogues

Pilot over-expression trials detected that soluble HelQ and Rad51 paralogue proteins could be generated using baculovirus insect cell based protein expression systems. This section describes the general procedures utilized to achieve this in each case, and should be read in conjunction with summarized information in Table 2.14. To overexpress N-terminally His-SUMO-HelQ, His-SUMO and His- tagged Rad51 paralogue proteins, Sf9 cells were cultured in Insect Xpress Protein Free Media (Lonza) supplemented with 100 units penicillin and 0.1 mg/ml streptomycin solution (Sigma), 1% (v/v) Pluronic F-68 solution (Gibco, Thermo Fisher Scientific) and 2.5 µg/ml Amphotericin B solution (Sigma). 400x10^7 Sf9 cells were infected with appropriate titre baculovirus, listed in Table 2.14, and incubated at 27°C with shaking at 170 rpm for a period of time listed in Table 2.14. Key parameters including virus titre and length of protein expression are detailed in Table 2.14. Cells were harvested by centrifugation at 3,000 *g* for 20 minutes. Biomass was resuspended 150 mM TRIS-HCl pH 8.0, 150 mM NaCl, 10% (v/v) glycerol, 20 mM imidazole, flash frozen in liquid N_2 and stored at -80°C.

Table 2.14: Conditions used for overexpression of listed proteins using Sf9 insect cell protein expression systems.

Protein	Virus/1x10^6 cells	Expression
His-SUMO-HelQ	8.3 µl	48 hours
His-SUMO-Rad51B	8.3 µl	48 hours
His-SUMO-Rad51C	8.3 µl	72 hours
His-Rad51D	16.6 µl	48 hours

2.15. Visualisation of protein expression

2.15.1. SDS-PAGE

SDS-PAGE gels were cast and electrophoresis was carried out at 180 V and 40 mA (per gel) in 1x SDS-PAGE running buffer for 1 hour using Bio-Rad Mini-PROTEAN apparatus. Proteins were visualized using GelCode Blue coomassie stain (Thermo Fisher Scientific) according to the manufacturer's protocol that involved heating the SDS-PAGE gel to 95°C in water followed by staining. Samples were prepared by addition of SDS-PAGE loading dye to 1x and samples were boiled at 95°C for 5 minutes.

In each instance described below, unless stated, Nu-PAGE pre-cast 10% acrylamide SDS-PAGE gels (Thermo Scientific) were used for high throughput analysis of samples at OPPF. Electrophoresis was carried out at 200 V in 1x MES running buffer for 40 minutes. Proteins were visualised using Instant Blue stain (Expedeon).

2.15.2. Immuno-Blotting for ELISA detection of proteins

Proteins were detected by western blotting using a general protocol with a HRP-conjugate secondary antibody as a reporter. In most cases the procedure was as follows: Proteins were transferred from acrylamide gels onto PVDF membrane (GE Healthcare) using a Geneflow electroblotting system at 100 V for 1 hour. The membrane was blocked overnight in blocking buffer (1x WBB containing 2.5% Marvel milk powder and 0.1% v/v Tween) then washed in 1x WBB containing anti-biotin antibody (Cell signalling technologies) (1:1,000), to visualize a biotinylated marker ladder, and a primary antibody (1:5,000) in blocking buffer; the identity of the primary antibody was dependent on the protein being analysed as described in Chapters 3-5. The membrane was washed thoroughly using 1x WBB containing a polyclonal secondary antibody (1:10,000) appropriate for binding to the primary antibody. The membrane was washed thoroughly and treated with Pierce ECL Western Blotting Substrate (Thermo Fisher Scientific) according to manufacturer's protocol. The chemiluminescent signal was visualised using a Fujifilm LAS 3000 imager.

2.16. Protein purification

All protein purifications were carried out at 4°C, unless stated otherwise, and all buffers and columns were pre-chilled to 4°C prior to use. Cell lysis buffers contained EDTA-free protease cocktail inhibitor

(Fisher Chemicals) that was also included in some subsequent buffers for column chromatography as detailed in the protocols below. All buffers used for protein purification were filtered and degassed.

2.16.1. Purification of His-HsaRad51

Frozen biomass containing over-expressed soluble His-HsaRad51 was thawed and resuspended in 20 mM TRIS-HCl pH 8.0, 20 mM NaCl and 20 mM imidazole[147] binding buffer and Protease Inhibitor Cocktail VII for His-tagged proteins (Fisher Chemicals) was added 30 µl per litre initial culture volume. Resuspended cells were lysed by sonication on ice using a Soniprep 150 ultrasonic disintegrator (MSE – Measuring and Scientific Equipment), and soluble proteins clarified by centrifugation at 35,000 *g* at 4°C using a Beckman Avanti J-26 XP centrifuge in a JA-20 rotor. Supernatant was filtered using a 0.2 µm sterile syringe filter. His-HsaRad51 was purified using an AKTA Prime system (Amersham pharmacia). The soluble protein fraction was loaded onto tandem 2x 5 ml His-Trap HP columns (GE-Healthcare) pre-equilibrated in binding buffer (20 mM TRIS-HCl pH 8.0, 20 mM NaCl and 20 mM imidazole). Proteins bound on the column were thoroughly washed in binding buffer containing 1 M NaCl before buffer exchange into binding buffer containing 20 mM NaCl. Proteins were eluted using a linear gradient of 20 mM TRIS-HCl pH 8.0, 20 mM NaCl and 1 M imidazole over 20 CVs. Fractions containing His-HsaRad51 were loaded onto a 5 ml Hi-Trap Heparin (GE Healthcare) pre-equilibrated in 20 mM TRIS-HCl pH 8.0, 20 mM NaCl,

2mM DTT and 10% v/v glycerol. Bound proteins were washed thoroughly before elution using a linear gradient of 20 mM TRIS-HCl pH 8.0, 1M NaCl, 2 mM DTT and 10% glycerol over 20 CVs. Fractions containing His-HsaRad51 were dialysed at 4^{o}C overnight into 50 mM TRIS-HCl pH 8.0, 50 mM NaCl, 5mM DTT and 10% v/v glycerol. For His-tag removal His-HsaRad51 was dialysed overnight into 500 mM TRIS-HCl pH 8.0, 2 mM $CaCl_2$, and 1% v/v TWEEN® 20 dialysis buffer. Enterokinase was added to 0.02 U/1 mg His-HsaRad51 and incubated at RT overnight. Sample was loaded onto a His-trap column and HsaRad51 was collected in the flow through before dialysis into 50 mM TRIS-HCl pH 8.0, 50 mM NaCl, 5mM DTT and 10% v/v glycerol at 4**°**C overnight. HsaRad51 was concentrated using 30,000 molecular weight cut off (MWCO) spin concentrators (Sartorius Stedim Biotech) at 8,000 *g* using a Beckman GS-15R benchtop centrifuge. Protein was aliquoted, flash frozen in liquid N_2 and stored at -80^{o}C. SDS-PAGE analysis was used to follow the purification process. HsaRad51 protein concentration was calculated using an extinction coefficient generated by the online ExPASy ProtParam tool 14900 and absorbance at 280 nM that was determined using an ND 1000 Spectrophotometer NanoDrop.

2.16.2. Purification of His-SUMO-HsaXRCC2

Frozen biomass containing over-expressed soluble His-SUMO-HsaXRCC2 was thawed and resuspended in 20 mM TRIS-HCl pH 8.0, 20

mM NaCl and 20 mM imidazole binding buffer and protease inhibitor cocktail VII for His-tagged proteins (Fisher Chemicals) was added 30 µl per litre initial culture volume. Resuspended cells were lysed by sonication on ice using a Soniprep 150 ultrasonic disintegrator (MSE – Measuring and Scientific Equipment), and soluble proteins clarified by centrifugation at 35,000 *g* at 4°C using a Beckman Avanti J-26 XP centrifuge in a JA-20 rotor. Soluble protein was passed through a 0.22 µm filter. His-SUMO-HsaXRCC2 was purified using an AKTA Prime system (GE Healthcare). The soluble protein fraction was loaded onto tandem 2x 5 ml His-Trap HP columns (GE-Healthcare) pre equilibrated in binding buffer 20 mM TRIS-HCl pH 8.0, 20 mM NaCl and 20 mM imidazole. Bound proteins were thoroughly washed in binding buffer containing 1 M NaCl before buffer exchange to binding buffer containing 20 mM NaCl. Bound proteins were eluted using a linear gradient of 20 mM TRIS-HCl pH 8.0, 20 mM NaCl and 1 M imidazole over 20 CVs. Fractions containing His-SUMO-XRCC2, determined by SDS-PAGE, were loaded onto a 5 ml Hi-Trap Q-sepharose column (GE Healthcare) pre-equilibrated in 20 mM TRIS-HCl pH 8.0, 20 mM NaCl, 2mM DTT and 10% v/v glycerol. Bound proteins were washed thoroughly before elution using a linear gradient of 20 mM TRIS-HCl pH 8.0, 1 M NaCl, 2 mM DTT and 10% v/v glycerol. Fractions containing His-SUMO-HsaXRCC2, as determined by SDS-PAGE, were dialysed at 4°C overnight into 20 mM TRIS-HCl pH 8.0, 50 mM NaCl, 5 mM DTT and 10% v/v glycerol. The protein sample was diluted dropwise with 20 mM TRIS-HCl pH 8.0, 2 mM DTT and 10% v/v glycerol to approximately 10 ms

conductivity. Protein was loaded onto a 5 ml heparin column (GE Healthcare) pre equilibrated in 20 mM TRIS-HCl pH 8.0, 20 mM NaCl, 2 mM DTT and 10% v/v glycerol. Bound proteins were washed thoroughly before elution using a linear gradient of 20 mM TRIS-HCl pH 8.0, 1 M NaCl, 2 mM DTT and 10% v/v glycerol. Fractions containing His-SUMO-HsaXRCC2, as determined by SDS-PAGE, were loaded onto a Superdex S75 24/60 (GE healthcare) pre equilibrated in 50 mM TRIS-HCl pH 8.0, 50 mM NaCl, 5 mM DTT and 10% v/v glycerol. Protein was aliquoted, flash frozen in liquid N_2 and stored at -80^{o}C. SDS-PAGE analysis was used to follow the purification process.

2.16.3. Purification of His-SUMO-HsaHelQ

Trial His-SUMO-HsaHelQ over-expression in Sf9 cells and micro-scale purification by OPPF-UK showed soluble protein could be obtained. Following optimization of Sf9 cell based His-SUMO-HsaHelQ protein expression, His-SUMO-HsaHelQ was purified. This involved several iterations of protein purification at Nottingham to optimize the purification protocol. The currently used purification method is described below, yielding helicase active human HelQ protein.

Frozen biomass containing over-expressed soluble His-SUMO-HsaHelQ was thawed and resuspended in 150 mM TRIS-HCl pH 8.0, 150 mM NaCl, 10% v/v glycerol, 20 mM imidazole. Protease inhibitor cocktail for His-tagged proteins (Fisher Chemicals) was added 30 µl per litre initial

culture volume. Cells were lysed gently by sonication using a Soniprep 150 ultrasonic disintegrator (MSE – Measuring and Scientific Equipment), and the soluble fraction clarified by centrifugation at 45,000 *g* at 4°C using a Beckman Avanti J26-XP centrifuge in a JA-20 rotor. Proteins within the soluble fraction were precipitated by addition of 0-50% w/v ammonium sulfate at 4°C. Protein was centrifuged at 9,000 *g* using a Heraeus Biofuge primo R benchtop centrifuge and carefully washed to remove excess ammonium sulphate solution before resuspension in cold buffer, 150 mM TRIS-HCl pH 8.0, 150 mM NaCl, 10% v/v glycerol, 20 mM imidazole. The resuspended protein pellet was passed through a 0.22 μm filter (Sartorius) and loaded onto a pre chilled 5 ml His-Trap column (GE Healthcare) pre equilibrated in chilled 150 mM TRIS-HCl pH 8.0, 150 mM NaCl, 10% v/v glycerol, 20 mM imidazole using an AKTA Prime system (GE Healthcare). Bound proteins were washed thoroughly with the same binding buffer containing 1 M NaCl, before buffer exchange back to 150 mM NaCl buffer before elution using an ascending linear gradient of chilled 150 mM TRIS-HCl pH 8.0, 150 mM NaCl, 10% v/v glycerol, 1 M imidazole over 20 CVs. Following dropwise dilution to approximately 9 ms conductivity using chilled 150 mM TRIS-HCl pH 8.0, 10% v/v glycerol buffer; protein was loaded onto a pre-chilled 1 ml Hi-Trap Q column (GE Healthcare) pre equilibrated in chilled 150 mM TRIS-HCl pH 8.0, 20 mM NaCl, 10% v/v glycerol, and eluted using an ascending linear gradient of chilled 150 mM TRIS-HCl pH 8.0, 1 M NaCl, 10% v/v glycerol over 20 CVs. Protein was loaded onto a pre-chilled Superdex S200 24/60 (GE Healthcare) pre

equilibrated in chilled 150 mM TRIS-HCl pH 8.0, 50 mM NaCl, 10% v/v glycerol. Eluted protein was aliquoted, flash frozen in liquid N_2 and stored at -80°C. All columns and superloops were kept at 4°C immediately prior to use and all buffers were pre chilled and kept in ice baths during protein purification. SDS-PAGE analysis was used to follow the purification process.

2.16.4. Purification of His-SUMO-HsaRad51B and His-SUMO-HsaRad51C

Frozen biomass containing over-expressed soluble His-SUMO-HsaRad51B and His-SUMO-HsaRad51C was thawed and resuspended in 20 mM TRIS-HCl pH 8.0, 20 mM NaCl independently. His-tagged protein inhibitor cocktail was added 30 µl per litre initial culture. Cells were lysed by sonication and the soluble fraction clarified by centrifugation at 45,000 *g* at 4°C using a Beckman Avanti J26 XP centrifuge in a JA-20 rotor. The soluble fraction was passed through a 0.22 µm filter (Sartorius) and loaded onto a 5ml His-Trap column (GE Healthcare) pre equilibrated in 50 mM TRIS-HCl pH 8.0, 50 mM NaCl, 10% v/v glycerol, 20 mM Imidazole, 0.5% v/v Triton X. Bound proteins were washed thoroughly with the same buffer containing 1 M NaCl, before buffer exchange back to 50 mM NaCl buffer before elution using an ascending linear gradient of 50 mM TRIS-HCl pH 8.0, 50 mM NaCl, 10% v/v glycerol, 1 M imidazole, 0.5% v/v Triton X over 20 CVs. Following dilution to approximately 10 ms conductivity using 50 mM TRIS-HCl pH 8.0, 5 mM DTT, 10% v/v glycerol, 0.5% v/v Triton X buffer; protein was loaded onto a 1 ml Hi-Trap Q column (GE

Healthcare) pre equilibrated in 50 mM TRIS-HCl pH 8.0, 20 mM NaCl, 10% Glycerol, 5 mM DTT, and eluted using an ascending linear gradient of 50 mM TRIS-HCl pH 8.0, 5 mM DTT, 1M NaCl, 10% v/v glycerol, 0.5% v/v Triton X over 20 CVs. Following dilution to approximately 10 ms conductivity as previously described, protein was loaded onto a 1 ml Hi-Trap Heparin column (GE Healthcare). Bound proteins were washed and eluted as described for Q column purification. Protein was spin concentrated to approximately 6 ml using 30,000 MWCO spin concentrators (Sartorius Stedim Biotech) at 8,000 *g* prior to loading onto a Superdex S75 24/60 (GE Healthcare) pre equilibrated in 50 mM TRIS-HCl pH 8.0, 50 mM NaCl, 10% v/v glycerol, 5 mM DTT. Eluted protein was aliquoted, flash frozen in liquid N_2 and stored at -80°C.

2.16.5. Purification of MthHel308 and HsaHelQ WHD Proteins

Isolated WHD proteins, and their mutants, from Hel308 of the archaeal species *Methanothermobacter thermautotrophicus* and from HelQ of human were purified in the same way. All were insoluble when expressed under a variety of different conditions. Harvested cells were suspended in WHD buffer (20 mM TRIS-HCl pH 8.0, 2.0 M NaCl, 10 mM DTT, 1.0 % v/v Tween 20, 10% v/v glycerol) containing a protease inhibitor cocktail (SIGMAFAST^{TM}) and freeze-thawed repeatedly at -80°C and room temperature prior to sonication and clarification at 35,000 *g* in an Avanti J26-XP centrifuge. The insoluble fraction (P1) was then washed repeatedly in buffer WHD and clarified as before, giving insoluble material (P2)

containing WHD protein. P2 was suspended in buffer WHD containing 6 M urea for 18 hours at 37°C and clarified as before. A substantial proportion of WHD protein was present in the soluble fraction (S3) and this was dialyzed extensively into buffer A (20 mM TRIS-HCl pH 8.0, 500 mM NaCl and 10% v/v glycerol) for loading a 5 mL Hi-Trap Ni-NTA column to bind to the $(His)_6$ tag of each protein. WHD proteins were eluted in buffer WHD, in an imidazole concentration gradient of 0.005-200 mM. WHD containing fractions were dialyzed into buffer B (20 mM TRIS-HCl pH 8.0, 100 mM NaCl, 10% v/v glycerol and 5 mM DTT) and loaded into a 1 mL Hi-Trap heparin column. Purified WHD proteins eluted from heparin at 300-500 mM NaCl and were then dialyzed into buffer B containing 40% v/v glycerol, for storage in aliquots at -80°C. Protein concentrations were estimated in two measurements using Bradford's reagent in kit form from Bio-Rad. First to determine wild type Hel308 and WHD protein concentration against a standard curve of bovine serum albumin protein concentrations, and second by using wild type Hel308/HelQ protein concentration to generate a standard curve for determination of relative concentrations of mutant Hel308 and WHD proteins. Purified WHD proteins were assessed for re-folding into a single species by size exclusion chromatography through a superose-12 column. Mutations were also made in the β-wing of Mth WHD, substituting Arg-477 and Arg-479 for glycine, but we were unable to detect any expression of this WHD protein.

2.17. Measurement of protein concentration

Rad51 protein concentration was measured as stated in section 2.35, Purification of His_6-HsaRad51. In addition, several different reagents and methods had to be used to assess the concentration of purified proteins because standard Bradford assay (Bio-Rad Protein Assay reagent) gave unreliable measurements for HelQ and Rad51 paralogues. The following methods were used to obtain reliable measurements.

2.17.1. Bicinchoninic acid (BCA) protein concentration assay

The bicinchoncinic acid assay is a colourmetric approach to measuring protein concentration, based on the Biuret assay, by formation of a purple colour that absorbs at 562 nm. The colour change that occurs by reduction of Cu^{2+} to Cu^{1+} is sensitive to cysteine, cystine, tyrosine and tryptophan residues in addition to peptide bonds. HelQ protein concentration was analysed using a BCA assay kit (Thermo Fisher Scientific), and the protocol was carried out according to the manufacturer's instructions.

2.17.2. Non-Interfering (NI) protein concentration assay

Different proteins have different buffer requirements to enable them to remain stable in solution. As a result, some of these compounds such as DTT for example, can affect the quantification of protein concentration, causing interference. For this colourmetric assay, protein is precipitated and removed from its buffer that may contain interfering

agents. Colourmetric change on addition of a copper solution can be detected at 480 nm and is sensitive to binding of copper to the polypeptide backbone. HsaHelQ concentration was determined using a Non-interfering protein assay kit (EMD Millipore), and the protocol was carried out according to the manufacturer's instructions.

2.18. Making labeled DNA substrates for assays *in vitro*

γ^{32}P labeling of oligonucleotides

DNA oligonucleotides were labeled with ^{32}P at 5' ends in 20 µl reactions that contained 300 ng oligonucleotide, 1x T4 PNK buffer, 1 µl T4 PNK and 1 µl γ^{32}P ATP (Perkin Elmer). Reactions were incubated at 37**°**C for 2 hours before heat inactivation at 65**°**C for 20 minutes. Volume of reactions was adjusted to 50 µl by adding SDW, and unincorporated ATP was removed using Micro Bio-Spin 6 spin columns (BioRad) according to manufacturer's protocol. Final eluted radiolabeled oligonucleotide was diluted to 200 µl final volume and DNA concentration calculated based on 90% recovery of DNA through BioSpin 6 columns (BioRad) of known starting material.

2.19. Incorporating radiolabeled DNA oligonucleotides into model DNA substrates.

Radiolabeled oligonucleotides (300 ng) were annealed with appropriate other unlabeled oligonucleotides (900 ng) in reaction mixtures of <100 µl containing 1x SSC buffer. Reactions were heated to

95°C for 10 minutes and allowed to reach room temperature overnight. Reactions were loaded onto 10% TBE gels and run at 200 V for 3 hours to isolate the correctly annealed substrate; the DNA molecule of the correct size as judged by electrophoretic mobility. Radiolabeled DNA substrates were detected by exposure to photographic film. Substrates were subsequently excised from the gel and eluted from the gel by diffusion in 200 µl 10 mM TRIS-HCl pH 7.5, 100 mM NaCl over 1-2 days.

2.20. Making fluorescently labeled synthetic DNA substrates

Fluorescent labels have long been used for applications in cell biology, however, more recently, fluorescent dyes have become a valuable tool for biophysical analysis as a means of detecting biological molecules covalently attached to them. Fluorescent dyes absorb and then emit energy and these can vary for each dye. Fluorescent dyes used here are Cy5 and Cy3 as they remain stable for up to two years at -20°C, and they have emission spectra that are compatible with our imager.

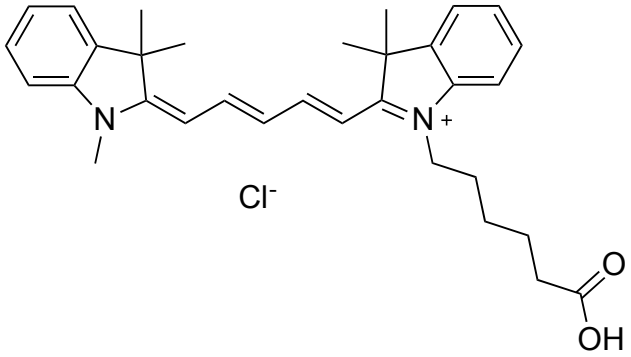


Figure 2.1: The chemical structure of Cy5 with carboxylic acid reactive group. Chemical structure was downloaded from ChemDraw online catalogue.

Cy3 and Cy5 fluorescent dyes are synthetic polymethine dyes that are used to label DNA and protein molecules for their visualization in *in vitro* and in cell biology assays, Figure 2.2. Cy3 and Cy5 fluorescent dyes are useful for FRET as they have overlapping absorption and emission wavelengths, therefore they can be used as FRET pairs. Absorption and emission wavelengths for Cy3 and Cy5 fluorescent dyes are detailed in Table 2.15. Oligonucleotides were ordered pre labeled with either Cy5 or Cy3 from Sigma.

Table 2.15: Details of Cy3 and Cy5 fluorophores

	Absorption/ nm	Emission/ nm	Extinction coefficient
Cy3	550	570	136,000
Cy5	649	670	250,000

Relevant oligonucleotides, listed in Table 2.9, were ordered labeled with Cy3 or Cy5 at the 5' end. Annealing reactions (50 µl) were assembled to contain 5 µM fluorescently labeled oligonucleotide, 6 µM

unlabeled oligonucleotide(s) and 1x annealing buffer comprising 10 mM Tris, pH 7.5, 50 mM NaCl and 100 mM sodium citrate. Reactions were heated to 95°C for 10 minutes and allowed to reach room temperature overnight. 10 µl loading dye (80% v/v glycerol, Orange G) was added to each reaction and were subsequently loaded onto 10% TBE gels and electrophoresed at 160 V for 3 hours. Substrates were subsequently excised from the gel and eluted from the gel by diffusion in 200 µl 10 mM TRIS-HCl pH7.5, 100 mM NaCl over 1-2 days. Concentration of substrates was determined using the Beer-lambert law (A=ECL) by obtaining an absorbance reading at 260 nm using a ND 1000 Spectrophotometer NanoDrop (Thermo Scientific) and generating an extinction coefficient for the substrate using OligoAnalyzer.

2.21. Assays for ATPase activity: Thin layer chromatography and malachite green methods

For thin layer chromatography measurements of ATPase activity, assays comprised reactions containing γ^{32}P ATP. Liberated γ^{32}P was separated from unhydrolysed γ^{32}P ATP by thin layer chromatography in which liberated γ^{32}P is carried by the mobile liquid phase faster than the large unhydrolysed ATP molecule. ATPase reactions of 10 µl were assembled to contain 250 nM ssDNA as indicated, 10 nM γ^{32}P ATP in ATPase buffer (20mM TRIS-HCl pH 7.5, 5 mM $MgCl_2$ and 10% v/v glycerol). Reactions were incubated for 1 hour at 37°C and were terminated by addition of 20 mM EDTA. 1.7 µl from each reaction was spotted onto PEI

cellulose TLC plates (Merck Millipore) that were developed in 0.375 M K_3PO_4. TLC plates were analysed by autoradiography using a BioRad molecular imager. Data was analysed using BioRad QuantityOne software.

ATPase activity was also assayed using malachite green as a reporter dye to measure liberation of free phosphate from ATP. In these assays all material and equipment was free of detergent, because of the sensitivity of malachite green for residual phosphate in detergents. For example assays avoided use of water from glassware that has been washed, and avoided pipette Tips from boxes because the boxes have been detergent washed. The following stock solutions were made:

- 0.045% aqueous solution of Malachite green hydrochloride in a dark bottle.
- 4.2% solution of Ammonium Molybdate in 4M HCl
- 34% aqueous solution of sodium citrate.
- Potassium dihydrogen phopshate as a standard solution (0.1M).
- Buffer (1xHB): 20 mM Tris.HCl pH 7.5, 100 ug/ml BSA.
- ATP (0.25 M), in ddH_20.

All of the above stocks are stable at room temperature, but ATP is stored in aliquots at -20, as is buffer.

Stock solutions were used to prepare assay solution immediately prior to assays as follows:

- Colour reagent was made by mixing 3:1 malachite green: Ammonium Mb in an allotted detergent-free beaker using plastic disposable pipette to measure out. Mix for 30 min. (flea must be detergent free also), and then pour through Whatman filter paper into a plastic disposable bottle. This is the colour reagent.

- A phosphate standard curve was generated: Use the potassium dihydrogen phosphate stock to generate the following standards in 1xHB to a final vol. of 100 uL: 0, 2, 4, 8, 12 & 16 nmols.

To the 100 uL phosphate add 800 uL of colour reagent and incubate for 2 min. Add 100 uL sodium citrate solution and incubate at RT for 20-40 min. Measure absorbance at 660 nM to give a linear calibration trace. Enzyme ATPase activity was measured in 100 uL reactions. Typically, reactions contained magnesium, ATP, 20 uL of 5 x HB and varying amounts of enzyme, made up to 100 uL with water. Reactions were stopped by adding 800 uL colour reagent for measurement as described above.

2.22. Electrophoretic mobility shift assays (EMSA)

Electrophoretic mobility shift assays (EMSAs) are biochemical 'in-gel' approaches for the study of protein-DNA interactions. The technique relies on stable protein-nucleic acid binding complexes forming that are detectable *via* labeling of the nucleic acid after gel electrophoresis. Complexes of higher molecular weight move more slowly through the gel, in comparison to unbound labeled DNA substrate.

Protein-DNA binding was assessed in 20 µl reactions containing 25 nM DNA substrate, 1x HB (20 mM Tris–HCl, pH 7.5, 100 µg/ml BSA), 0, 5 or 50 mM DTT as indicated and HsaHelQ between 1-160 nM range. Reactions were incubated for 10 minutes at 37**°**C. Protein-DNA complexes were resolved using 5% acrylamide TBE gels and were run at 100 V for 1 hour 45 minutes. Data was analysed using GelEval software.

2.23. Nucleic acid unwinding assays

DNA unwinding by HsaHelQ helicase activity was assessed in assays to detect dissociation of annealed DNA strands, detectable through ^{32}P or Cy5 labeling of one DNA strand. Ultimately, fluorescently labeled DNA substrates were used to assay unwinding by HsaHelQ, and it became apparent that an unlabeled strand 'trap' of the same sequence as the labeled strand was required to detect HsaHelQ unwinding activity, discussed in Chapter 4. Completed reactions are loaded onto TBE gels for electrophoresis to separate DNA strands by molecular weight. On dissociation of a duplex DNA by HsaHelQ, the product generated can be identified from the radio or fluorescent labeled strand. HsaHelQ unwinding was assayed by assembling 20 µl reactions to contain 5mM $MgCl_2$, 5mM ATP, 25 nM DNA, 1x HB (20 mM Tris–HCl, pH 7.5, 100 mg/ml BSA), 2.5 µM cold trap DNA, as specified, 50 mM DTT where indicated and HelQ between 0-160 nM range. Reactions were incubated at 37°C for 10 minutes and were terminated by addition of 1/5th volume stop buffer (100 mM TRIS-HCl pH 7.5, 100 mM $MgCl_2$, 3% w/v SDS, 10 mg/ml proteinase K). Reactions were loaded onto a 10% w/v acrylamide TBE gel in loading dye (80% v/v glycerol, Orange G), and products separated by electrophoresis on 10% acrylamide TBE gels at 160V for 1 hour. Gels were visualized using a FujiFilm FLA3000 machine at 633 nm and an R765 filter. TIFF images were analysed using GIMP and GelEval (FrogDance) software. For time courses, 20 µl was removed from a 200 µl reaction at indicated time intervals and stopped by addition of 1/5th volume stop buffer.

2.24. Native gel electrophoresis

Nu-Page Native gels were used for Blue Native page electrophoresis. Protocol used was as described in the manufacturers instructions. Protein loading dyes were chosen according to the pI of the proteins to be analysed. The methodology specific to the work reported in this thesis is detailed in Chapter 4.

2.25. Analytical gel filtration

Analytical gel filtration (AGF) is a form of size exclusion chromatography. AGF columns are densely packed and operate under high pressure to achieve high-resolution protein separation. This type of chromatography is sensitive to protein size and shape.

To assess protein size and shape after refolding, 3 µg protein sample was injected onto a Superose 12 10/30 column pre equilibrated in 20 mM Tris pH 8.0 containing 150 mM NaCl using an AKTA Purifier (GE Healthcare). BioRad Gel Filtration standards (151-1901) were used to make a marker trace for comparison.

2.26. Bioinformatics and molecular modeling of proteins

2.26.1 Uniprot

The Universal Protein Resource (UniProt) is a database of protein sequences and annotations. Protein sequences in FASTA format were obtained from Uniprot for the following protein accession numbers, Table 2.16. FASTA format represents protein sequences in single letter amino

acid code in a simple text based format that is compatible with most bioinformatics tools. Protein sequences were obtained for the major isoforms of proteins where relevant.

Table 2.16: UniProt accession numbers of proteins analyzed by bioinformatics and molecular modeling.

Protein	**Uniprot accession number**
Homo sapiens HelQ	Q8TDG4
Methanothermobacter thermautotrophicus Hel308	O26901
Homo sapiens Rad51	Q06609
Homo sapiens Rad51B	O15315
Homo sapiens Rad51C	O43502
Homo sapiens Rad51D	O75771
Homo sapiens XRCC2	O43543

2.26.2. Protein sequence alignments

Different amino acid alignment tools are available that vary in the type of alignment output, such as pairwise or multiple sequence alignments for example, and by the algorithms used to generate the alignments. For this work I used Clustal OMEGA or EMBOSS Needle for global alignments, to identify similarities between proteins of different species. EMBOSS Water local alignments were used to analyse regions of local homology. Protein sequences were entered in FASTA format.

2.26.3. Molecular modeling: Phyre 2, I-TASSER

Protein amino acid sequences were entered into the online molecular modeling programs Phyre 2[218] and I-TASSER[219] in FASTA format obtained from UNIPROT. Modeling by Phyre 2 was carried out under 'intensive' mode, which is discussed further where relevant in section 4.3. Results were downloaded as PDB files and rendered using MacPymol.

2.26.4. Modeling multimeric HsaHelQ structures: COTH and Galaxy Gemini

COTH and Galaxy Gemini model protein sequences were obtained from UniProt and were entered in to COTH[220] and Galaxy Gemini[221] in FASTA format. Both COTH and Galaxy Gemini are limited to protein sequences no longer than 1,000 amino acids in length. Results were downloaded as PDB files and rendered using MacPymol.

2.27. NanoTemper Microscale Thermophoresis

NanoTemper Microscale Thermophoresis is a method of detecting binding between a protein and its ligand or another protein[222]. This technique is primarily used to generate kinetic data including dissociation constants. A 40 µl reaction mix containing 1x reaction buffer (20 mM Tris–HCl, pH 7.5, 100 mg/ml BSA and 0.1% v/v Triton-X) and 250 µM HelQ was made. This was serially diluted 1:1 into sixteen 0.5 ml Eppendorf tubes. 20 µl 1 nM fluorescently labeled DNA substrate diluted in 1x reaction buffer

was added to each tube and incubated for 5 minutes. Reactions were centrifuged at 13,000 rpm for 5 minutes using a Genfuge 24D microfuge (Progen). Approximately 4 µl each sample was transferred into capillary tubes for analysis using the NanoTemper Monolith NT.115 Pico. An initial 'Capillary Scan' was carried out and the LED intensity adjusted until readings were between 2,500 and 25,000 units, and appropriately shaped peaks were obtained indicative of sample quality. Following this, microscale thermophoresis (MST) readings were taken at 20% and 40% LED power.

2.28. Immobilising DNA on quartz slides for TIRF

We aimed to investigate possible 'DNA reeling' by HsaHelQ by single molecule FRET (smFRET) using total internal reflection (TIRF) microscopy. Förster resonance energy transfer is the transfer of energy between an excited donor fluorophore to its acceptor 'partner'. This phenomenon is very sensitive to distance as described by:

$$E=(1+(R/R_0)^6)^{-1}$$

Where R_0 represents the Förster radius where E = 0.5, and R = inter dye distance[223]. As a result, conjugation of fluorescent FRET pairs to biological molecules for example, can enable the study of conformational change on single molecules using TIRF microscopy, where numerous single molecules can be visualized and populations of these molecules may or may not exhibit FRET. This involves immobilization of fluorescently labeled

biotinylated DNA molecules onto polyethylene glycol (PEG) and biotinylated quartz slides *via* a neutravidin linkage. By immobilizing fluorescently labeled DNA molecules to a surface, background scattering is reduced and smFRET can be used to observe protein mediated activities and conformational change of individual molecules. The PEG biotin coated quartz slides contained four channels for use with different samples. Chambers were flushed with 0.2 mg/ml neutravidin and 50 mM TRIS-HCl pH 7.5 and incubated for 5 minutes. Unbound neutravidin was flushed from the wells using 20 mM TRIS-HCl pH 7.5 before injection of 5 pM biotinylated fluorescently labeled DNA substrates. After 20 minutes incubation, unbound DNA molecules were flushed from the wells using 20 mM TRIS-HCl pH 7.5. Approximately 100 fluorescently labeled DNA molecules should be in the field of view, described below, for good data collection.

DNA substrates used for smFRET were 3'-OH tailed partial duplex molecules of differing overhang lengths that carried two fluorescent dyes: a Cy5 acceptor fluorophore at the 3'-OH end of the overhang, and a Cy3 donor fluorophore at the 5' end at the duplex-ssDNA junction. This allows detection of potential ssDNA reeling activity by HsaHelQ.

Specialised imaging buffers must be used for smFRET data acquisition by TIRF. This is because O_2 can cause photobleaching, or loss of fluorescence, of fluorescent dyes such as Cy3 and Cy5 thereby interfering with FRET. To avoid this, imaging buffers for observing smFRET

by TIRF microscopy contained O_2 scavengers to remove O_2 from the sample. Glucose oxidase and 6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid (TROLOX) are commonly used O_2 scavengers. Both of these utilize O_2 in the presence of their substrates. By addition of O_2 scavengers and their relevant substrates in imaging buffers, O_2 is removed from the samples and FRET can be observed.

2.28.1. TIRF for smFRET

smFRET data were acquired using an Olympus IX71 microscope adapted for prism-type total-internal reflection (TIRF). This involves direction of a laser (Crystalaser) at 562 nm at a prism at a specific angle known as the critical angle. An evanescent wave occurs at the interface of the quartz and the solution as a result of different refractive indices. This subsequently results in localized illumination to observe the immobilized labeled molecules with minimal background, therefore reducing noise[224]. Donor and acceptor images were recorded in two channels using an Ixon EMCCD camera (Andor, Belfast) and 1,000 frames were collected at different speeds. Images were processed using Interactive Data Language (IDL). Data was analysed using the 'Single Molecule FRET' program written by members of the St Andrews Single Molecule BioPhysics laboratory (Penedo group) using MatLab. FRET efficiency was calculated using:

$$E=I_A/(I_A+I_D)$$

where I_A and I_D represent intensity of the acceptor (Cy5) and donor (Cy3)

respectively[225].

Protein samples were injected into the chambers in 1x imaging buffer (6% (w/w) glucose (Fisher Scientific), 0.1 mg/ml glucose oxidase (Sigma) and 0.02 mg/ml glucose catalase (Sigma), 1 mM TROLOX (Sigma), 1 mM ATP and 1 mM $MgCl_2$) and incubated for 5 minutes prior to imaging. HsaHelQ concentrations in the range 100 pM-1 µM were tested.

2.29. Circular dichroism

Circular dichroism is a biophysical technique used to study secondary structure of proteins. WHD proteins for circular dichrosim were dialysed into 20 mM TRIS-HCl (pH 8.0) and 150 mM potassium acetate. 200 µl protein sample was inserted into a 1 mm Helma quartz cuvette. CD measurements were taken using a Chirascan CD Spectrometer (Applied PhotoPhysics) from 180-250 nm at 25°C. Results were analysed using CDNN software[226].

2.30. Genetic assay for evaluating *Methanothermobacter thermautotrophicus* Hel308 function

Assessment of archaeal Hel308 activity by genetic assay in *E. coli* strain *dnaE486* Δ*recQ* is described fully elsewhere[5,227]. Routine growth and handling of this strain is at 30°C, except for a 30 second, 42°C, heat shock step during the plasmid transformation protocol. For viability "spot"

tests, *E. coli* cells transformed by plasmid encoding Hel308 protein, or an empty plasmid control, were grown as overnight cultures in LB containing ampicillin (50 μg/ml). These were used to inoculate fresh LB for growth to O.D. 600 of 0.8, before serial dilution of the cultures into M9 minimal medium. Ten microliters of each dilution was spotted onto LB agar plates and the agar plates were incubated at 30°C or 37°C, as indicated. The strain becomes unstable for DNA replication (semi-permissive) at 37°C degrees, allowing the phenotype associated with expression of wild type Hel308 to be observed. In each assay a control agar plate was also incubated at 42°C, which is non-permissive for growth of *dnaE486* Δ*recQ*, and from which we would expect no colony growth unless suppressor mutations or contamination had occurred.

Chapter 3: Molecular cloning, over-expression and purification of human and archaeal proteins

3.1. Introduction and aims

Hel308 and HelQ are DNA translocases implicated in processing of stalled replication forks to facilitate their restart. Human HelQ is reported to interact with other HR proteins, including the Rad51 paralog complex BCDX2 that comprises a 1:1:1:1 ratio of Rad51B, Rad51C, Rad51D and XRCC2[185]. A first aim of this PhD project was to generate reproducible preparations of human proteins HelQ, Rad51 paralogues and Rad51 to investigate (a) biochemical properties of HelQ, (b) physical and functional interactions between HelQ, Rad51 and Rad51 paralogues, (c) to analyse the winged helix domains of human HelQ and its archaeal homologue Hel308, this last part being presented in chapter 5. Chapter 3 describes reproducible and reliable purification of active human HelQ and Rad51 proteins in high yield, and attempts to isolate Rad51 paralogue proteins.

3.2. Molecular Cloning for HelQ and other human proteins

For all the human proteins investigated here, gene codons were optimized for protein expression in *E. coli* during their design for synthesis using the GeneArt® service (Thermo Fisher Scientific), detailed in gene sequence and project information in appendices at the end of this thesis. Options for codon optimization of genes for protein expression in *E. coli* are readily available for genes synthesized *de novo*. Codon optimization

was necessary because although codons that encode amino acids are universal across the domains of life, some species exhibit "codon bias", a preference for one codon to encode for an amino acid in preference to others. This occurs because although multiple codons are universal to encode specific amino acids, in some organisms particular codons are favoured over others leading to common and rare tRNAs. As a result, some codons within human genes may correspond to rare tRNAs in *E. coli* leading to sub-optimal translation and therefore problems in generating significant protein over expression. Codon optimization of a gene for protein expression in a given species therefore minimises this problem, to improve protein expression.

GeneArt® (Thermo Fisher Scientific) synthesized *rad51* (sequence is given in an appendix to the thesis) was subcloned into pBADHisA to generate pEBPS2, encoding N-terminally hexahistidine tagged HsaRad51, herein referred to as HsaRad51, a construct that included an enterokinase cleavage site for removal of the His-tag if desired. pBADHisA is a protein expression vector suitable for protein expression in *E. coli* when transcriptionally activated ("induced") by the addition of L-arabinose to growth media. HsaRad51 ATPase-defective mutants were made using site-directed mutagenesis on pEBPS2, generating genes encoding HsaRad51 K133R, HsaRad51 D222A and a HsaRad51 K133R/D222A double mutant, described in section 2.10. Rad51 K133R, a Walker A motif (-G-xxxx-GK-TS) mutant, is defective in ATP binding because the arginine residue carries a guanido group that is delocalized in comparison to the more focused

formal positive charge of the lysine NH_3 group. HsaRad51 D222A, a Walker B motif (R/K-XXX-G-XXX-L-hhhhD) mutant, is defective in ATP hydrolysis as the aspartate is no longer able to coordinate Mg^{2+} that is required as a co-factor for catalysis.

High-throughput cloning of human genes *helq*, *rad52* and the genes encoding HsaRad51 paralogues Rad51B, Rad51C, Rad51D, XRCC2 and XRCC3 was carried out at the Oxford Protein Production Facility (OPPF-UK). Each gene was amplified by PCR from plasmid constructs made by GeneArt® custom gene synthesis (Thermo Fisher Scientific, see appendices for gene sequences), using primers kindly designed by Dr. L. Bird (Chapter 2) for ligation independent cloning (LIC) into a selection of vectors, primarily: pOPINE for untagged constructs, pOPINF for N-terminally His-tagged constructs and pOPINS3C for N-terminally His-SUMO-tagged constructs. These are described in more detail in section 2.13.1.

Human *helq*, *rad52* and Rad51 paralogue genes were amplified by PCR from GeneArt plasmid constructs also made by GeneArt® (Thermo Fisher Scientific, see appendices for gene sequences), using primers designed for InFusion™ cloning (ClonTech) into pOPIN vectors. This was carried out in a 96 well format in which 24 constructs were generated in quadruplicate. Human Rad51 paralogue genes were successfully amplified by PCR. Although PCR of Rad51 paralogues was straightforward, multiple attempts using different parameters were required to clone human *helq* by varying annealing temperatures and testing different DNA polymerases

before successful PCR amplification. The successful method is described in section 2.13.1. Overall, difficulties cloning *helq* was encountered at several stages throughout this project.

Following apparently successful cloning, detected using blue-white screening, PCR verification using a vector-specific forward primer and a gene-specific reverse primer confirmed the presence of the correct genes as visualized by size. PCR products containing amplified genes from new constructs can be seen in Figure 3.1, and PCR products for *helq* that were achieved at a later date, due to technical challenges during cloning, are presented in Figure 3.2. Gene sequences were subsequently verified by Sanger sequencing and no unexpected mutations or polymorphisms were detected. A complete table of constructs made at OPPF-UK is listed in Table 3.1.

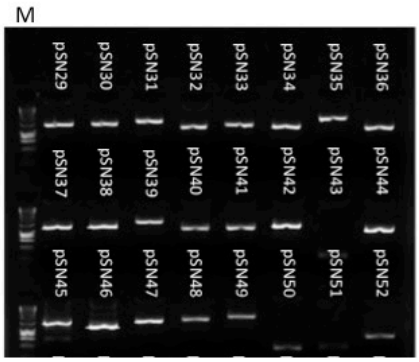


Figure 3.1: Cloning of human Rad51 paralogue genes at OPPF-UK.
1.0% agarose gel showing products from PCR amplification of Rad51 paralogue and Rad52 genes generated from newly made constructs at OPPF, using a vector-specific forward primer and gene-specific reverse primer. M indicates the molecular weight ladder. Further details of each pSN prefixed clone shown is given in Table 3.1.

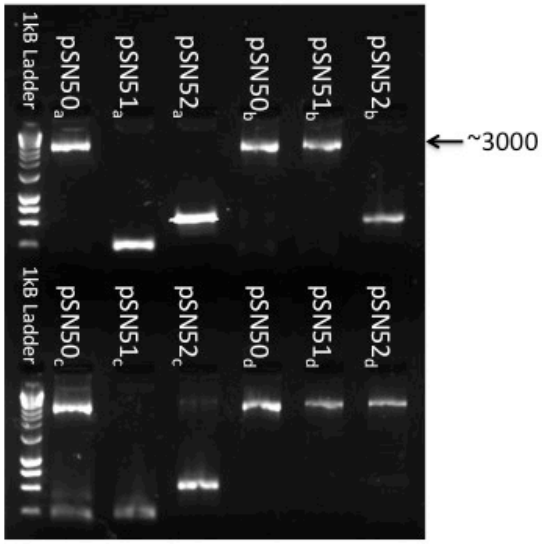


Figure 3.2: Successful cloning of human *helq* at OPPF-UK.

1.0% agarose gel showing products from PCR amplification of human *helq* genes from newly made constructs at OPPF, using a vector-specific forward primer and gene-specific reverse primer. Each putative clone was tested in quadruplicate (pSN50, 51 or 52, Table 3.1) and was expected to generate product of approximately 3,000 bp, consistent with full-length *helq*. However, note that some clones were substantially shorter than expected, reflecting recurring difficulties during *helq* cloning described in the main text.

Table 3.1: DNA constructs made at OPPF-UK. Details of constructs containing *helQ*, *rad52* and Rad51 paralogue genes generated at OPPF-UK. Genes are listed alongside parent expression vectors with details of affinity purification tags and antibiotic resistance markers.

Plasmid name	Gene	Parent Vector	Tag	Antibiotic resistance marker
pSN29	*rad51b*	pOPINE	Untagged	Carbenicillin
pSN30	*rad51b*	pOPINF	N-terminal His_8-3C	Carbenicillin
pSN31	*rad51b*	pOPINS3C	N-terminal His_6-SUMO-3C	Carbenicillin
pSN32	*rad51c*	pOPINRSE	C-terminal $KHis_6$	Kanamycin
pSN33	*rad51c*	pOPINE	Untagged	Carbenicillin
pSN34	*rad51c*	pOPINF	N-terminal His_8-3C	Carbenicillin
pSN35	*rad51c*	pOPINS3C	N-terminal His_6-SUMO-3C	Carbenicillin
pSN36	*rad51d*	pOPINCDE	C-terminal $KHis_6$	Spectinomycin
pSN37	*rad51d*	pOPINE	Untagged	Carbenicillin
pSN38	*rad51d*	pOPINF	N-terminal His_8-3C	Carbenicillin
pSN39	*rad51d*	pOPINS3C	N-terminal His_6-SUMO-3C	Carbenicillin
pSN40	*xrcc2*	pOPINE	Untagged	Carbenicillin
pSN41	*xrcc2*	pOPINF	N-terminal His_8-3C	Carbenicillin
pSN42	*xrcc2*	pOPINS3C	N-terminal His_6-SUMO-3C	Carbenicillin
pSN43	*xrcc3*	pOPINE	Untagged	Carbenicillin
pSN44	*xrcc3*	pOPINF	N-terminal His_8-3C	Carbenicillin
pSN45	*xrcc3*	pOPINS3C	N-terminal His_6-SUMO-3C	Carbenicillin
pSN46	*xrcc3*	pOPINRSE	C-terminal $KHis_6$	Kanamycin
pSN47	*rad52*	pOPINE	Untagged	Carbenicillin
pSN48	*rad52*	pOPINF	N-terminal His_8-3C	Carbenicillin
pSN49	*rad52*	pOPINS3C	N-terminal His_6-SUMO-3C	Carbenicillin
pSN50	*Helq*	pOPINE	Untagged	Carbenicillin
pSN51	*Helq*	pOPINF	N-terminal His_8-3C	Carbenicillin
pSN52	*Helq*	pOPINS3C	N-terminal His_6-SUMO-3C	Carbenicillin

HsaHelQ has a 300 amino acid long N-terminal region with no significant sequence homology to any other protein when compared in database searches. Therefore structural models of this part of HsaHelQ give no secondary structure, appearing as a 'long N-terminal tail' prior to the five predicted helicase domains. This region of HsaHelQ is hypothesized to be involved in interacting with HsaRad51 paralogues and therefore cloning of this region alone was attempted for use in subsequent pull down assays with purified HsaRad51 paralogs. Initially cloning was attempted into C-terminal MBP tagged vector pMAL-C2, however, PCR of the HsaHelQ N-terminal extension proved challenging and despite attempting different conditions, no PCR products could be obtained reproducibly from templates pSN50-52 carrying *helq*. As an alternative, the HsaHelQ N-terminal tail was successfully synthesized by GeneArt (Life Technologies) and subcloned into pEX-C-GST that encoded a C-terminal GST tag, to generate pSN53. The GST tag was placed at the C-terminus of the N-terminal tail, as this is where the polypeptide is naturally connected to the HsaHelQ helicase core, with the aim of potentially minimizing any occurrence of steric hindrance between MBP and HsaHelQ-N-terminus.

Additionally, cloning of the HsaHelQ helicase without the long N-terminal extension was carried out. PCR amplification of HsaHelQ core helicase to contain *Xho1* and *Kpn1* restriction sites was successful (Figure 3.3), described in section 2.11, but subcloning of this PCR fragment into pBADHisA could not be achieved. This was reminiscent of cloning

challenges encountered by GeneArt (Thermo Fisher Scientific), who were unable to sub-clone the newly synthesized intact *helq* into pBADHisA during early stages of the project. This is discussed further in Chapter 6. A HsaHelQ fragment of approximately 2800 bp consistent with the expected size for the HsaHelQ core helicase can be seen in Figure 3.3.

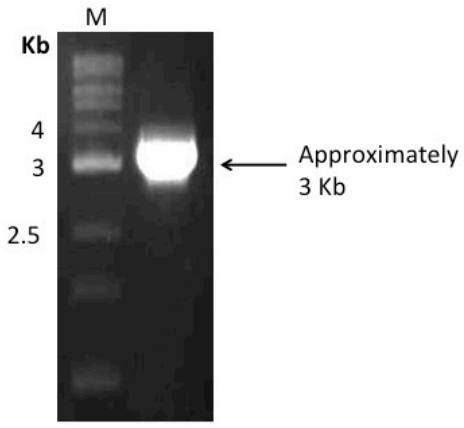


Figure 3.3: PCR amplification of HsaHelQ core helicase.
1.0% agarose gel showing the PCR amplification product for HsaHelQ core helicase containing Xho1 and Kpn1 restriction sites. This fragment can be observed at approximately 2800 bp, consistent with the expected size.

3.3. Protein over-expression in *E. coli*

There are several advantages to *E. coli* protein expression systems in comparison to other expression systems, including insect cell – baculovirus, mammalian, and *in vitro* cell free transcription-translation systems. *E. coli* protein over-expression is often most cost effective and least time consuming. However, there are disadvantages to over-expression of human proteins using *E. coli*. Post-translational modifications (PTMs) such as phosphorylation or glycosylation are often made to proteins in eukaryotic hosts, and these often impact on protein

function. It is unlikely that a PTM required for function of a eukaryotic protein will be successfully reconstituted within *E. coli* expression systems. However, due to the advantages of this system, over-expression of HsaHelQ, HsaRad51, HsaRad51 paralogue proteins was initially carried out using *E. coli*.

3.4. HsaRad51

Soluble expression of N-terminally hexahistidine tagged HsaRad51 was achieved in BL21 AI *E. coli* at 37°C after three hours post-induction, described in section 2.14.1, see Figure 3.4. HsaRad51 over-expression was heavily dependent on adequate aeration of *E. coli* cultures during growth and protein over-expression as described in Chapter 2, section 2.19.

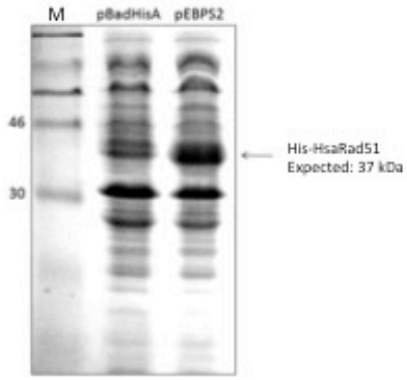


Figure 3.4: Over-expression of HsaRad51:
Coomassie blue stained 10% acrylamide SDS-PAGE analysis of pilot over-expression of HsaRad51, at 37 kDa, from pEBPS2 in BL21 AI *E. coli* following induction using 0.2% L-arabinose for 3 hours. Cells carrying empty vector pBADHisA after induction are shown as a negative control.

3.5. HsaRad51 paralogues

Protein over-expression trials of HsaRad51 paralogues in *E. coli* were carried out at OPPF-UK, described in section 2.14.4. Expression plasmids each carrying one Rad51 paralogue gene were transformed into chemically competent Lemo21 or Rosetta 2 *pLacI* (DE3) *E.coli* expression strains, both of which overcome "leaky" transcriptional control of gene expression from plasmids and therefore repress basal protein expression. Protein expression in each strain was tested using two different types of media: Power broth, and auto-induction media as described in Chapter 2. Following micro-scale protein purification using automated His-tag protein purification described in section 2.14.5, SDS-PAGE analysis showed unconvincing HsaRad51 paralogue protein over-expression, or in some cases low expression of soluble protein; see example in Figure 3.5 and summary in Table 3.2.

It was reasoned that this problem may be related to the propensity of HsaRad51 paralogues to exist in large complexes such as BCDX2, or at least smaller sub-complexes, such as Rad51B-Rad51C[228]. Therefore, protein co-expression was tested, for the potential to overcome the possibility that expression of individual proteins was sub-optimal because of their instability in the absence of a binding partner(s). Co-expression of HsaRad51 paralogues is also likely to stabilize protein structures and may improve solubility on protein expression generally. However, attempts at co-expression of HsaRad51C and HsaRad51B to obtain soluble protein was not successful in any strain of *E. coli* tested

(Table 2.2) using various permutations of growth temperature, growth media and buffer conditions for resuspending harvested biomass. Therefore additional cloning of HsaRad51 paralogues was carried out into specialized vectors designed for co-expression of proteins in *E. coli*, such as pETDuet, however this was also unsuccessful in yielding any soluble protein.

Table 3.2: Detectable expression of Rad51 paralogue proteins obtained at OPPF-UK. Summary of pilot experiments to test for protein expression in *E. coli* strains Lemo21(DE3) and Rosetta 2(DE3)pLacI, following micro-scale His-tag protein purification carried out at OPPF-UK. Untagged constructs are not shown here as they were not amenable for micro-scale purification due to lack of a hexahistidine affinity tag. *Helq* clones are listed for completeness, but no expression outcomes are shown here, due to delays in *helq* cloning at the time that meant that pilot HelQ expression tests were not carried out at OPPF.

Plasmid name	**Gene**	**Expression in *Lemo21 E. coli***		**Expression in Rosetta 2 (DE3) pLacI**	
		Power broth	**Overnight Express**	**Power broth**	**Overnight express**
pSN29	*rad51b*	-	Negative	-	Negative
pSN30	*rad51b*	Negative	Negative	Negative	
pSN31	*rad51b*	Negative	Positive	Positive	
pSN32	*rad51b*	-	-	-	-
pSN33	*rad51b*	-	Negative	-	Negative
pSN34	*rad51b*	Negative	Negative	Negative	
pSN35	*rad51b*	Positive	Negative	Negative	
pSN36	*rad51d*	-	-	-	-
pSN37	*rad51d*	-	Negative	-	Negative
pSN38	*rad51d*	Negative	Negative	Negative	
pSN39	*rad51d*	Negative	Negative	Negative	
pSN40	*xrcc2*	-	Negative	-	Negative
pSN41	*xrcc2*	Positive	Negative	Positive	
pSN42	*xrcc2*	Positive	Positive	Negative	
pSN43	*xrcc3*	-	Negative	-	Negative
pSN44	*xrcc3*	Negative	Negative	Negative	
pSN45	*xrcc3*	Positive	Positive	Positive	
pSN46	*xrcc3*	-	-	-	-
pSN47	*rad52*	-	Negative	-	Negative
pSN48	*rad52*	Positive	Negative	Positive	
pSN49	*rad52*	Negative	Negative	Negative	
pSN50	*helq*	-	-	-	-
pSN51	*helq*	-	-	-	-
pSN52	*helq*	-	-	-	-

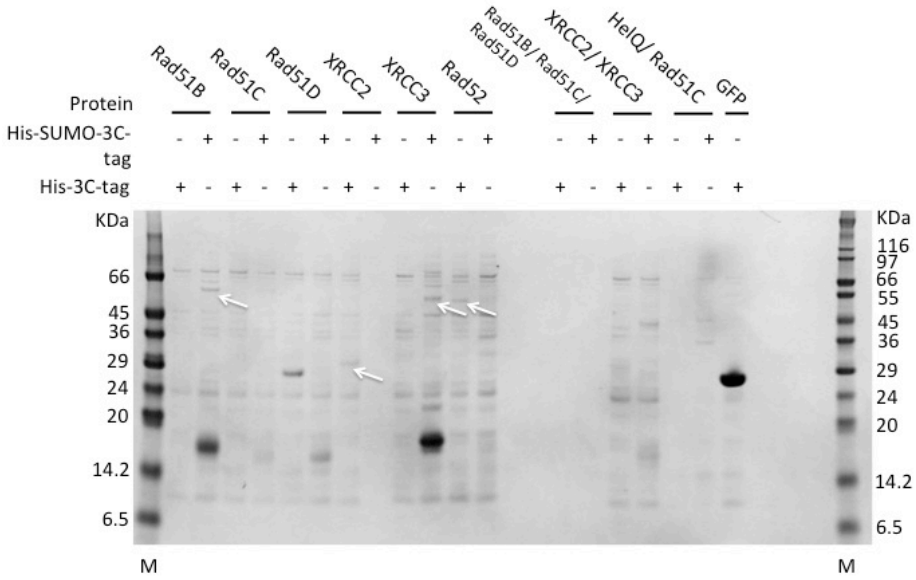


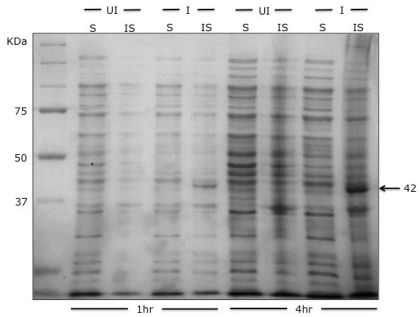
Figure 3.5: Over-expression of soluble Rad51 paralogue proteins using *E. coli* expression systems at OPPF-UK

Instant blue stained 10% acrylamide SDS-PAGE analysis of soluble protein over-expression and micro-scale purification of HsaRad51 paralogue proteins in strain Rosetta 2(DE3)pLacI cultured in Power Broth™ media. Arrows indicate soluble expression of desired proteins. GFP was used as a control. Co-expressions were carried out in where the first protein listed was tagged as indicated. M indicates molecular weight markers. This is one of four gels to assess soluble over-expression of these proteins in each case using different parameters of growth media as detailed in Table 2.5, Chapter 2. To save space, only one of these gels is shown.

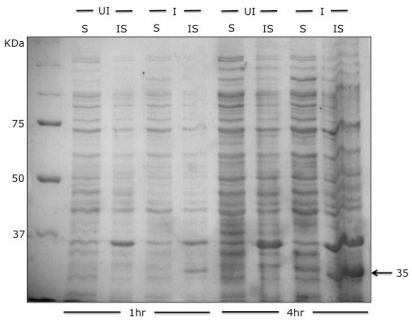
Optimisation of Rad51 paralogue protein over-expression was attempted using additional *E. coli* strains including BL21 AI, T7 Shuffle and Rosetta-gami 2; protein expression was tested at lower than standard temperatures (18°C) overnight and using different media such as Overnight Express Instant TB media, described in section 2.14.6. However, although protein expression was often very good, human Rad51B,

Rad51C, Rad51D and XRCC3 variants were all insoluble under conditions tested, see examples in Figure 3.6.

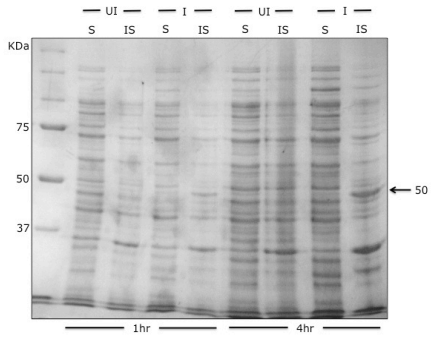
A: Untagged HsaRad51B



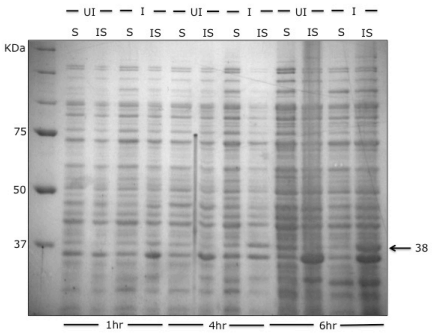
B: Untagged HsaRad51D



C: His_6-SUMO-HsaRad51D



D: Untagged HsaXRCC3



E: His-HsaXRCC3

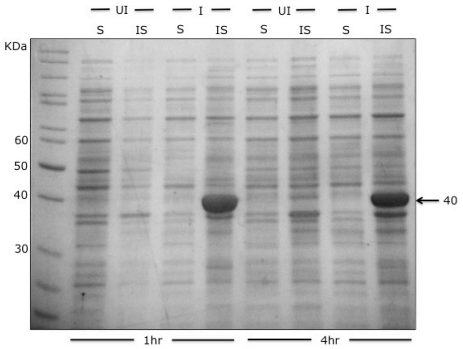


Figure 3.6: Insolubility of Rad51 paralogue proteins using *E. coli* expression systems.

GelCode Blue stained 12% acrylamide SDS-PAGE gels displaying proteins present in soluble (S) and insoluble (IS) fractions from total cell protein samples taken at indicated time points. Uninduced (UI) and induced (I) samples are shown as indicated. All of the desired paralogue proteins are

visible in the insoluble fraction. Arrows and expected kDa sizes indicate each over-expressed protein as follows; A: Untagged HsaRad51B; B: Untagged HsaRad51D; C:His-SUMO-HsaRad51D; D: Untagged HsaXRCC3; E: His-HsaXRCC3.

In contrast, His-SUMO-HsaXRCC2 could be expressed as soluble protein, using Lemo21(DE3) *E. coli* grown in Overnight Express Instant TB media described in section 2.14.7, see Figure 3.7. This was the only HasRad51 paralogue for which a substantial yield of soluble protein could be obtained for purification. This may be explained due to the smaller size of HsaXRCC2 in comparison to the other Rad51 paralogues. Additionally, molecular models of HsaXRCC2 predict that the N-terminal domain found in recombinases and Rad51 paralogues is missing in HsaXRCC2. This domain may be responsible for rendering the other Rad51 paralogues insoluble.

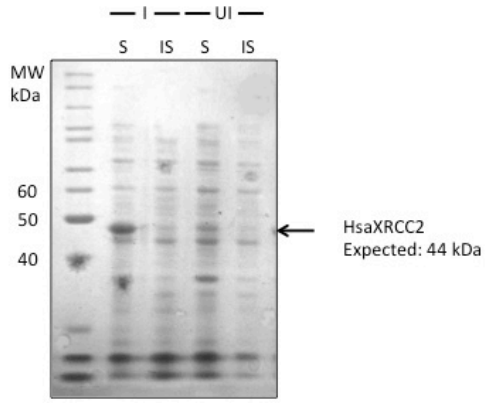


Figure 3.7: Soluble over-expression of HsaXRCC2 in *E. coli*. Coomassie stained 12% acrylamide SDS-PAGE gels displaying His-SUMO-HsaXRCC2 protein present in soluble (S) and insoluble (IS) fractions from total cell protein samples. Uninduced (UI) and induced (I) samples are shown as

indicated. A protein at approximately 44 kDa, consistent with His-SUMO-HsaXRCC2 is observed in the induced soluble lane.

In some cases insoluble protein can be denatured from the insoluble fraction and refolded. This is typically only suitable for small proteins up to approximately 20 kDa, as refolding of large proteins is likely to result in mis-folding and a high risk of inactive protein recovery. This was attempted with HsaXRCC3, by immobilizing the protein on a His-trap column and carrying out buffer exchange very slowly. However, no soluble XRCC3 could be obtained in this way. Other refolding options may have been more successful.

3.6. HsaHelQ over-expression in *E. coli*

Pilot HsaHelQ protein expression tests were not possible alongside Rad51 paralogues at OPPF-UK due to delays with *helq* cloning, and therefore these were carried out later. Analysis of HsaHelQ was essential to this project; therefore a great deal of effort was put in to obtain soluble protein expression in *E. coli*. In summary, HsaHelQ constructs were transformed into several *E. coli* strains: BLR, DE3, BL21 AI, Lemo21(DE3) and Rosetta-gami 2. Protein expression under different conditions was trialed by varying protein expression at lower than standard temperatures to allow for slower protein folding that may improve solubility. *E. coli* cultures were also grown in different media including LB broth, 2XYT media and Overnight Express Instant TB media, in addition to adding

glucose supplements to the media. Untagged HsaHelQ could be over-expressed in BLR *E. coli* using LB, however despite rigorous efforts, no soluble HsaHelQ protein was observed.

The insolubility of HsaHelQ in *E. coli* could be caused by mis-folding or a toxic effect of its expression triggering physiological responses leading to insolubility. To investigate the latter possibility, we attempted to generate helicase inactive mutants in *E. coli* plasmids. Insoluble HsaHelQ expression may have been due to toxic effects of HsaHelQ to *E. coli*. However, this could not be tested due to difficulties during site directed mutagenesis by PCR, as no products could be formed even under a range of conditions. In addition, the HsaHelQ N-terminal region is predicted to be disordered when analyzed using database searching. Removal of this region may improve expression of the core HsaHelQ helicase, however, this could not be tested, again because of cloning problems in which subcloning of the PCR fragment into the target vector could not be achieved. Following these various strategies used to try to obtain soluble HsaHelQ from *E. coli*, ultimately, HsaHelQ was expressed using baculovirus-Sf9 insect cell protein expression systems described in Chapter 2 and in section 3.8 below.

3.7. N-terminal HelQ (pSN53).

Pilot protein-overexpression studies for pSN53 encoding a C-terminally GST tagged HsaHelQ N-terminal tail was successful using BL21 AI and Codon Plus *E. coli*, shown in Figure 3.8, as described in Chapter 2.

This polypeptide was intended for use in HsaRad51 paralogue pull downs and biophysical structural analysis.

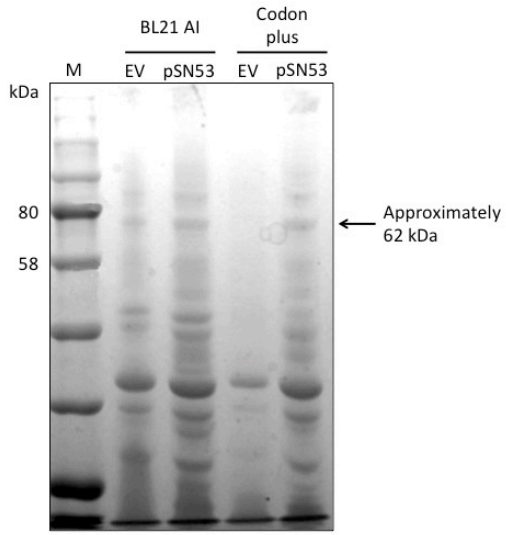


Figure 3.8: Over-expression of HsaHelQ N-terminal tail. Coomassie stained 10% acrylamide SDS-PAGE analysis of over-expression of HsaHelQ N-terminal tail protein from BL21 AI and Codon Plus *E. coli*, observed at approximately 62 kDa, the expected molecular weight for this protein.

3.8. Over-expression of HsaHelQ and HsaRad51 paralogues using baculovirus-Insect cell systems

Problems encountered in HsaHelQ production using *E. coli* led us to investigate if we could over-express soluble HsaHelQ using Baculovirus-Sf9 insect cell protein expression systems at OPPF-UK. Insect cell protein expression has several advantages over other protein expression systems, particularly for human proteins. Insect cells are more similar to vertebrate cells than yeast for example, and expressing human proteins in an organism as close to their natural host as possible is beneficial due to the natural occurrence of protein chaperones and post-translational

modifications. As a result, higher yields of soluble protein can be obtained due to correct folding. In addition, insect cells can carry out complex post translational modifications similar to those that occur in mammalian cells that cannot be carried out in bacterial or yeast protein expression systems. Protein expression in insect cells often produces a higher yield of protein in comparison to mammalian protein expression systems.

Baculovirus-Sf9 insect cell-based protein expression produced soluble HsaRad51B, HsaRad51B, HsaXRCC2 and HsaSUMO-HelQ proteins, shown in Figure 3.9 after micro-scale His-tag protein purification described in section 2.14.5. This was in contrast to insoluble proteins from expression of HsaRad51 paralogues when using *E. coli*. Once again, co-expression of Rad51 paralogue complexes was attempted in an effort to express stabilizing protein complexes, however, this proved unsuccessful using this expression system. Following encouraging observation that soluble HsaRad51 paralogue and HsaHelQ proteins could be obtained using baculovirus Sf9 insect cell based protein expression systems at OPPF-UK, optimization of protein expression was next carried out as described in sections 2.14.8 to 2.14.12. Recombinant baculovirus generated at OPPF-UK was used where baculovirus titre and time of protein expression was varied. Protein expression was not always detectable by eye following SDS-PAGE, therefore western blots were used, probing with monoclonal antibodies specific for hexahistidine-tags as described in sections 2.15.1 and 2.15.2.

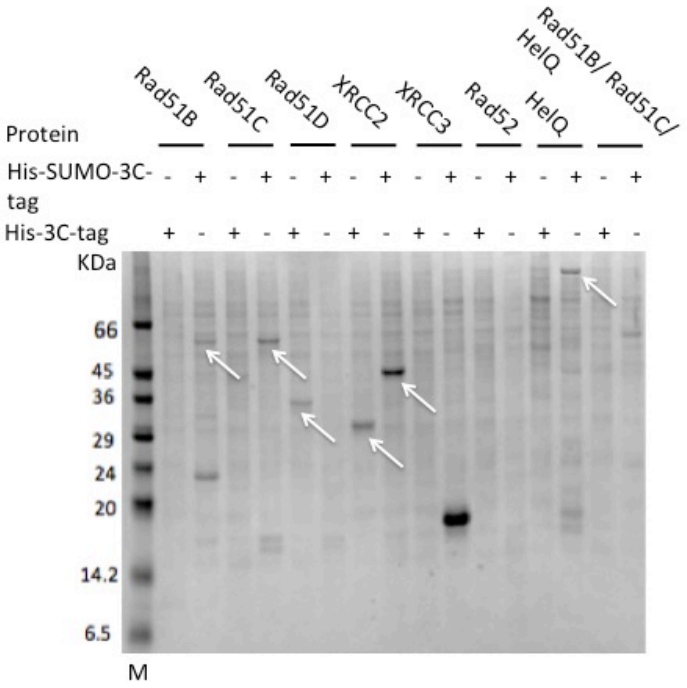


Figure 3.9: Over-expression of soluble Rad51 paralogue and HsaHelQ proteins using Sf9 insect cell based protein expression systems at OPPF. Instant blue stained 10% acrylamide SDS-PAGE analysis of soluble protein over-expression and micro-scale purification of HsaRad51 paralogue proteins and HsaHelQ using Sf9 insect cell based protein expression systems following micro-scale His-tag protein purification. Arrows indicate soluble expression of human Rad51B, Rad51C, Rad51D, XRCC2 and HelQ proteins.

3.9. Optimization of HsaHelQ protein expression using baculovirus Sf9 insect cell systems

Protein expression optimization showed expression of a protein at approximately 137 kDa, consistent with HsaHelQ. This was further confirmed by western blot, described in section 2.15.2 and shown in Figure 3.10. However, there are also multiple other bands of lower molecular weight, which could be HsaHelQ fragments as a result of

degradation, or other histidine rich proteins. Nevertheless, according to this data, optimal protein expression is achieved at 8.3 µl virus/ $1x10^6$ cells for 72 hours.

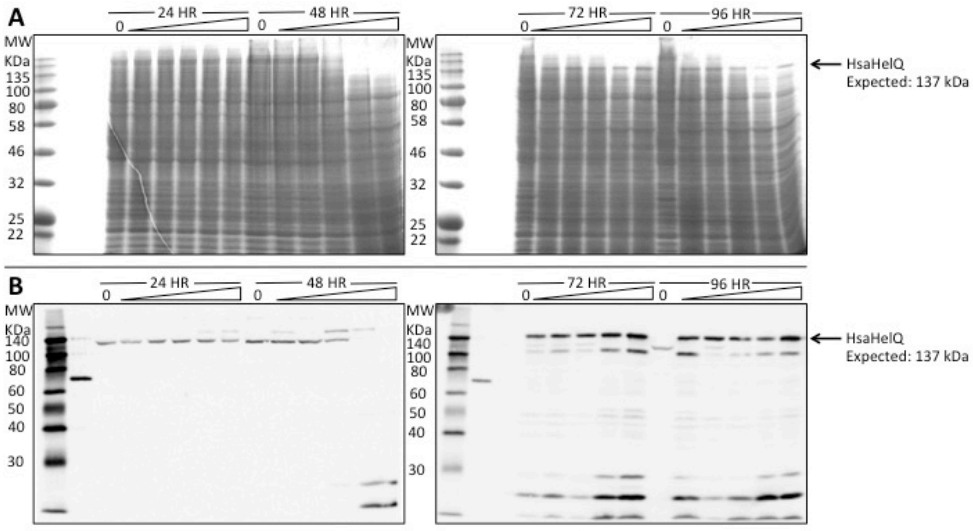


Figure 3.10: Detection by SDS-PAGE and Western blotting of optimised His-SUMO-HsaHelQ protein expression from baculovirus-Sf9 insect cells. His-SUMO-HsaHelQ protein expression using baculovirus Sf9 insect cell protein expression systems was optimized using recombinant baculovirus generated by OPPF-UK, by varying baculovirus titre and time of protein expression. Virus was titrated from 0, 0.835, 1.67, 3.34, 8.35, 16.7 µl virus/ $1x10^6$ cells and samples taken every 24 hours for four days. A protein of approximately 137 kDa indicated by the arrow, consistent with His-SUMO-HsaHelQ, could be observed by western blot after 48 hours. Good expression of His-SUMO-HelQ occurred after 72 hours with little variation in viral titre, and can be observed by GelCode Blue stained 10% acrylamide SDS-PAGE (A) and western blot (B) analysis. A hexahistidine tagged protein was used as a positive control for western blot probing using an anti-histidine antibody. His-SUMO-HsaHelQ degradation can be observed by western blot.

3.10. Optimization of HsaRad51B and HsaRad51C protein expression using baculovirus-Sf9 insect cells

Soluble Rad51 paralogues His-SUMO-HsaRad51B, and His-SUMO-HsaRad51D were also over-expressed using Baculovirus-Sf9 insect cell based protein expression at OPPF-UK. Protein expression optimization of HsaRad51B and HsaRad51C in Sf9 showed over-expression of proteins at approximately 45 and 46 kDa, consistent with His-SUMO-HsaRad51B and His-SUMO-HsaRad51C respectively. Again, these proteins were detected, and expression monitored by western blot using anti-histidine antibodies shown in Figure 3.11 and 3.12. Optimal expression of His-SUMO-HsaRad51B was obtained at 8.3 µl/ $1x10^6$ cells over 48 hours, and for His-SUMO-HsaRad51C 8.3 µl/ $1x10^6$ over 72 hours.

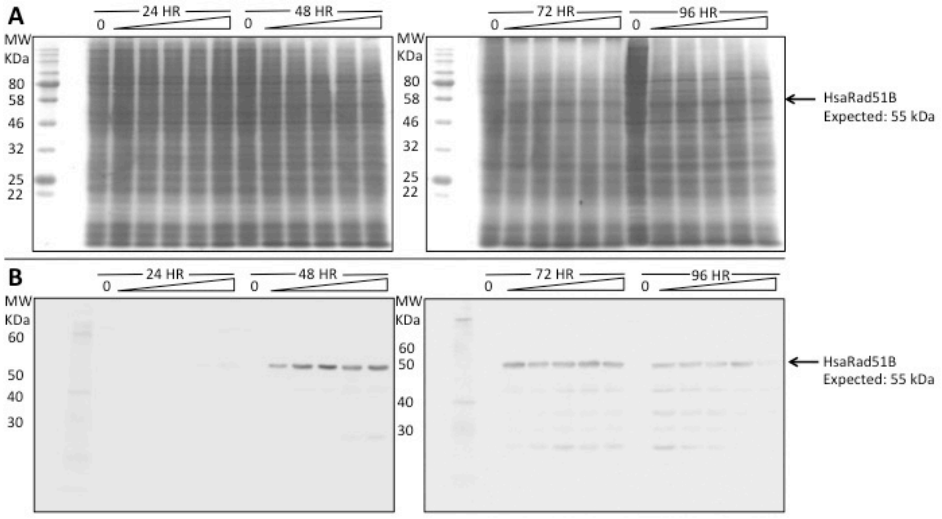


Figure 3.11: Detection by SDS-PAGE and Western blotting of optimised protein expression of His-SUMO-HsaRad51B from baculovirus-Sf9 insect cells. His-SUMO-HsaRad51B protein expression using baculovirus Sf9 insect cell protein expression systems was optimized using recombinant baculovirus generated by OPPF-UK, by varying baculovirus titre and time of protein expression. Virus was titrated from 0, 4.2, 8.3, 16.7, 33.35, 66.7 µl virus/ $1x10^6$ cells and samples taken every 24 hours for four days. A protein of approximately 55 kDa indicated by arrows, consistent with His-SUMO-HsaRad51B, could be observed by western blot after 48 hours with little variation in viral titre, and can be observed by GelCode Blue stained 10% acrylamide SDS-PAGE (A) and western blot (B) analysis. A hexahistidine tagged protein was used as a positive control for western blot probing using an anti-histidine antibody.

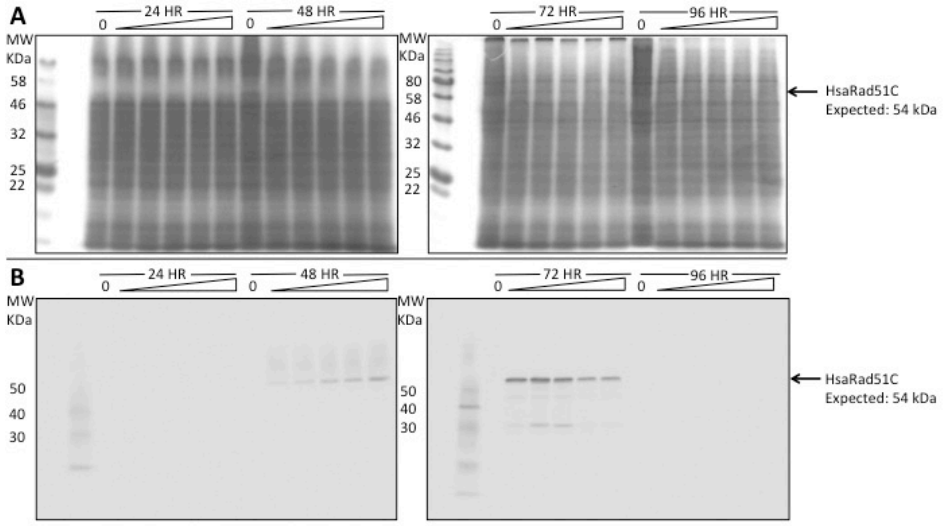


Figure 3.12: Detection by SDS-PAGE and Western blotting of optimised protein expression of His-SUMO-HsaRad51C from baculovirus-Sf9 insect cells. His-SUMO-HsaRad51C protein expression using baculovirus Sf9 insect cell protein expression systems was optimized using recombinant baculovirus generated by OPPF-UK, by varying baculovirus titre and time of protein expression. Virus was titrated from 0, 4.2, 8.3, 16.7, 33.35, 66.7 µl virus/ $1x10^6$ cells and samples taken every 24 hours for four days. A protein of approximately 54 kDa indicated by arrows, consistent with His-SUMO-Rad51C, could be observed by western blot after 48 hours. Good expression of His-SUMO-HsaRad51C occurred after 72 hours with little variation in viral titre, and can be observed by GelCode Blue stained 10% acrylamide SDS-PAGE (A) and western blot (B) analysis. We do not know why detectable expression of His-SUMO-HsaRad51C is lost at 96 hours, but this may be because of instability of the protein after prolonged expression. A hexahistidine tagged protein was used as a positive control for western blot probing using an anti-histidine antibody.

3.11. Protein purification

As detailed in Chapter 2, various FPLC column procedures were exploited to try to purify HsaHelQ, HsaRad51 and each human Rad51 paralogue. The next sections detail purification attempts and activity assays for each protein in turn, and are summarized in Table 3.3. Protein molecular weight and pI details used to design purification protocols are listed in Table 3.4.

Table 3.3: Summary data about the human proteins over-expressed in Sf9 or *E. coli* cells studied in this work, including the affinity tag utilized during purification of each protein, and the outcome.

Recombinant protein	Molecular weight/ kDa	Affinity tag	Molecular weight + tag/ kDa	Purification
HsaHelQ (Sf9)	124	His-SUMO	137	Successful, Figure 3.14
HsaRad51 (*E. coli*)	37	His	42	Successful, Figure 3.12
HsaRad51B (Sf9)	42	His-SUMO	55	Partially-successful, Figure 3.17
HsaRad51C (Sf9)	42	His-SUMO	54	Partially-successful, Figure 3.17
HsaRad51D (Sf9)	35	His	38	Partially-successful, Figure 3.17
HsaXRCC2 (*E. coli*)	32	His-SUMO	44	Successful, Figure 3.16

Table 3.4: Theoretical isoelectric points (pI) for affinity tagged HsaHelQ and human homologous recombination proteins that were studied in this work. Isoelectric points were used to identify a significant charge at a particular pH for loading onto ionic exchange columns during optimization of protein purification. All proteins gave strongest charges at pH 8.0, therefore all purification buffers were TRIS based at pH 8.0. Nevertheless, some of the predicted charges are weak, which may explain sometimes weak binding observed when utilizing ion-exchange purification methods.

Protein	Estimated pI	Charge at pH 8.0
His-Rad51	6.41	-8.6
His-SUMO-HelQ	6.38	-25.2
His-SUMO-Rad51B	5.69	-20.2
His-SUMO-Rad51C	6.33	-14.7
His-Rad51D	6.29	-9.6
His-SUMO-XRCC2	5.9	-20.1

3.12. Purification of archaeal Hel308

MthHel308 was purified to homogeneity *via* a combination of salting out, and FPLC-based chromatography using hydrophobic, ion-exchange and size-exclusion procedures, as described in Guy and Bolt, 2005[5], with an example of the resulting purified protein shown in Figure 3.13.

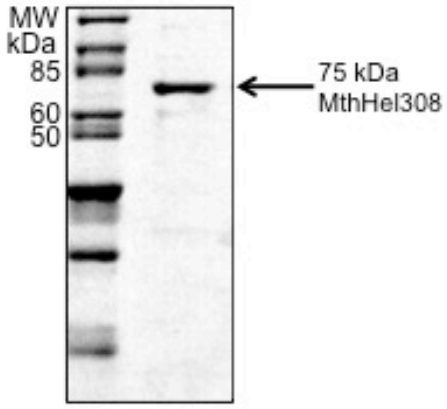


Figure 3.13: Coomassie-blue stained 8% acrylamide SDS-PAGE gel showing purified Hel308 enzyme from the euryarchaeon *Methanothermobacter thermautrophicus.*

3.13 Protein purification: HsaRad51

N-terminally hexahistidine tagged HsaRad51 was purified to homogeneity by affinity chromatography through Ni^{2+}-NTA, an immobilized metal affinity (IMAC) resin, and heparin resin as described in section 2.16.1. For His-HsaRad51, protein loaded onto the Ni^{2+}-NTA column was washed with 20 mM imidazole before elution over an ascending imidazole gradient from 50 mM-1M over 20-column volumes (CVs). Peak fractions containing His-HsaRad51 were pooled and dialysed prior to using a heparin column. His-HsaRad51 was purified to near homogeneity, by washing with 50 mM NaCl while immobilized on heparin, then eluted using an ascending NaCl gradient from 50 mM- 1 M over 20 CVs. His-HsaRad51 purification analysis by SDS-PAGE is shown in Figure 3.13. To note, His-HsaRad51 runs slightly larger than expected by SDS-PAGE.

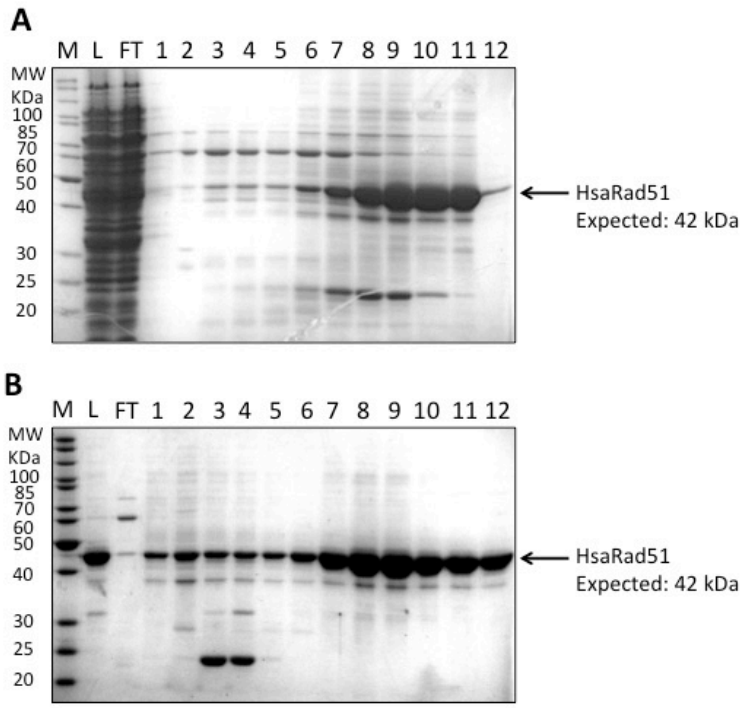


Figure 3.14: His-HsaRad51 protein purification.

Coomassie stained 12% SDS-PAGE analysis of His-HsaRad51 purification using affinity chromatography techniques. (A) His-HsaRad51 was initially separated by Ni^{2+}-NTA affinity chromatography. Fractions 3-11 were pooled for loading onto heparin. (B) Heparin affinity chromatography purified His-HsaRad51 to near homogeneity. Fractions 5-12 were pooled for dialysis into storage buffer; the final purified Rad51 protein is shown in Figure 3.19.

Subsequent His-HsaRad51 mutant proteins, described later, were purified using the same protocol for WT His-HsaRad51 described in section 2.16.1, and behaved similarly to WT His-HsaRad51 throughout the purification process. Pure K133R His-HsaRad51 can be seen in Figure 3.15.

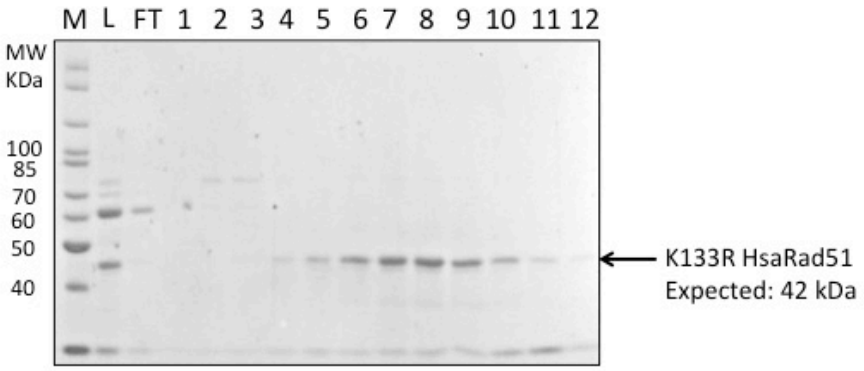


Figure 3.15: Heparin purification of HsaRad51 ATPase defective mutant K133R

8% acrylamide coomassie stained SDS-PAGE. K133R HsaRad51 could be purified to homogeneity using the same method as for WT HsaRasd51 as detailed in section 2.35. Heparin elution fractions are shown and fractions 5-11 were pooled for dialysis into storage buffer.

The N-terminal hexahistidine tag could be removed from His-HsaRad51 using enterokinase cleavage of the Asp-Asp-Asp-Lys-X sequence, where cleavage of the peptide bond is located immediately after the lysine residue. Tagged and untagged HsaRad51 proteins were tested in downstream assays, and no difference was observed, therefore uncleaved N-terminal hexahistidine tagged HsaRad51 was used for the majority of this work and is stated as and when relevant. Purified HsaRad51 protein concentration was determined by measuring A_{280} and A_{280}:A_{260} ratios in a ND 1000 Spectrophotometer Nanodrop. During purification attempts it was shown that HsaRad51 is amenable to spin concentration, if necessary. Typical His-HsaRad51 protein yields were approximately 2.5 mg/L.

3.14. Protein purification: HsaHelQ

Purification of His-SUMO-HsaHelQ was based on three steps: salting out using 0-50% ammonium sulphate precipitation, followed by Ni^{2+}-NTA and size exclusion chromatography as described in section 2.16.3. Ammonium sulphate precipitation of proteins is a useful method for purification of some proteins, by removing a significant amount of contaminants, and protein can also often be stored stably in this way. Ammonium sulphate precipitation can also remove DNA from samples, which is often essential when studying DNA binding proteins.

His-SUMO-HsaHelQ was loaded onto Ni^{2+}-NTA, and proteins bound to HsaHelQ *via* metal unspecific electrostatic interactions were removed using binding buffer containing 1 M NaCl. Following buffer exchange into low salt conditions required for the remaining chromatography steps, His-SUMO-HsaHelQ was eluted using an ascending gradient from 20 mM to 1 M imidazole over 20 CVs. However, much His-SUMO-HsaHelQ protein was lost in the flow-through, possibly due to partial occlusion of the N-terminal hexahistidine tag. This could be potentially improved by tagging HsaHelQ at the C-terminus, although tagging the C-terminus of MthHel308 renders it inactive. His-SUMO-HsaHelQ purification was followed by SDS-PAGE analysis, Figure 3.16.

His-SUMO-HsaHelQ was then subjected to purification by ion exchange chromatography. His-SUMO-HsaHelQ has a theoretical pI of 6.38 and a charge of -25.2 at pH 8.0, and therefore was purified using a Q column in a TRIS based buffer at pH 8.0. Protein was eluted using an

ascending gradient from 50 mM to 1 M NaCl, over 20 CVs. Again, on all attempts at His-SUMO-HsaHelQ purification, a large amount of protein was lost during loading, perhaps indicative of sup-optimal buffer conditions. This is also evident in that small amounts of His-SUMO-HsaHelQ elutes from the column throughout the entire gradient. It is apparent from His-SUMO-HsaHelQ elution from the Q-column, shown in Figure 3.16B, which this stage did little to further separate His-SUMO-HsaHelQ from the many contaminants that were loaded following Ni^{2+}-NTA chromatography. A previous attempt at His-SUMO-HsaHelQ purification used heparin chromatography following Q-sepharose. Although this did not significantly improve the purity of His-SUMO-HsaHelQ it did result in more concentrated protein as it eluted from heparin in a sharp peak at approximately 400 mM NaCl. Although heparin chromatography was not used to generate active His-SUMO-HsaHelQ protein used in this project, this may be considered in the future to ensure removal of any nuclease contaminants that may interfere with any downstream assays.

For a final step, His-SUMO-HsaHelQ was purified to near homogeneity by size exclusion chromatography, a technique that separates proteins according to shape and size where large proteins or protein complexes elute from the size exclusion column earlier than smaller proteins. This was important following His-SUMO-HsaHelQ elution from the Q-sepharose column because of remaining protein contaminants spanning a range of molecular weights. His-SUMO-HsaHelQ eluted from a

Superdex S200 column as two peaks, consistent with His-SUMO-HsaHelQ forming a higher order oligomeric state, or protein aggregates. The oligomeric state of His-SUMO-HsaHelQ is unclear from existing literature, however, through previous analytical gel filtration reported in Marini *et. al.*[4], the authors suggest that HsaHelQ may adopt a higher order state as opposed to a monomeric form. New data about this potentially interesting feature of HsaHelQ is presented in Chapter 4.

Typical His-SUMO-HsaHelQ protein yields following optimisation of the purification protocol were approximately 0.2 mg/L.

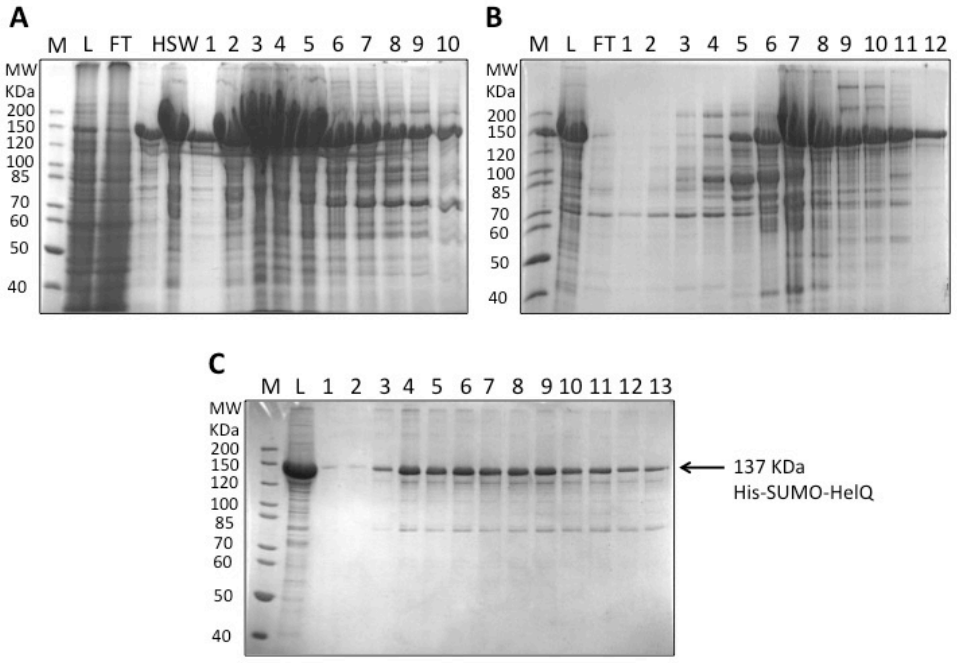


Figure 3.16: Purification of His-SUMO-HsaHelQ protein.

Coomassie stained 10% SDS-PAGE analysis of His-SUMO-HsaHelQ purification using affinity chromatography techniques. His-SUMO-HsaHelQ is indicated by an arrow at 137 kDa on each gel. See also Figure 3.17 for a summary gel showing purified HelQ protein. (A) His-SUMO-HsaHelQ purification by Ni^{2+}-NTA affinity chromatography. Fractions 1-10 were pooled for loading onto Q-sepharose. (B) Q-sepharose affinity chromatography did little to separate His-SUMO-HsaHelQ from contaminants that remained after Ni-NTA. Fractions 3-12 were pooled for protein separation by size exclusion chromatography. (C) His-SUMO-HsaHelQ was purified to near homogeneity by size exclusion chromatography; other detectable proteins in this gel are degradation products of His-SUMO-HsaHelQ, whose abundance follows the same peak intensity as the intact His-SUMO-HsaHelQ protein, are readily detectable in western blots as being fragments of His-SUMO-HsaHelQ, for example in Figure 3.10, and some of which have had their identities confirmed by mass spectrometry.

3.15 General considerations on refinement of HelQ protein purification after several iterations

We encountered several problems and inconsistencies in His-SUMO-HsaHelQ purification during different iterations of HsaHelQ purification using several purification attempts. These are informative in understanding more about HelQ function and behavior for future work, and so are detailed here with reference to further experiments detailed in Chapter 4.

3.16. Instability of purified HsaHelQ protein

Initial attempts at purification of His-SUMO-HsaHelQ involved multiple dialysis steps, but following a final size exclusion step multiple bands were visible after coomassie blue staining of SDS-PAGE gels. Following protein identification by trypsin digestion and mass spectrometry carried out by L. Manzi (Oldham Lab, School of Chemistry, University of Nottingham), to rule out protein contaminants, all proteins present in the HelQ sample were revealed to be HsaHelQ protein fragments. We concluded that HsaHelQ is prone to degradation, as shown in Figure 3.17 A. In order to prevent degradation, it became apparent that His-SUMO-HsaHelQ purification must be carried out rapidly and at low temperature as far as possible. All subsequent His-SUMO-HsaHelQ purifications were carried out using pre-chilled equipment (i.e. all purification columns and superloops) by chilling at 4^{o}C overnight prior to use. In addition, all buffers were made the day prior to His-SUMO-

HsaHelQ purification and chilled at 4^{o}C overnight. During purification, all buffers were kept in ice baths.

HsaHelQ degradation initially observed during protein purification was also a problem on repeatedly freeze-thawing of the purified enzyme, or if the enzyme was used at room temperature, in which case degradation of HsaHelQ was severe, and manifested as compromised activity in assays with DNA substrates, exemplified in Figure 3.17 B. DNA binding activities of HsaHelQ are detailed further in results chapter 4.

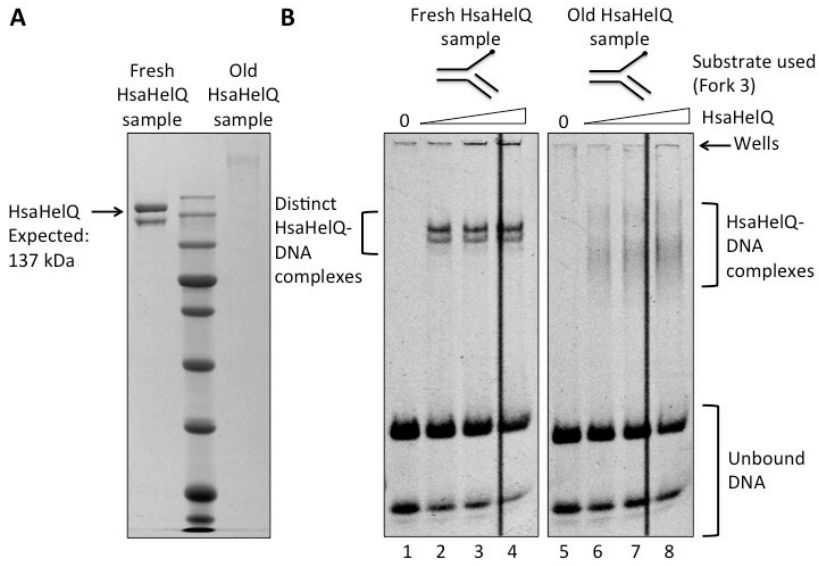


Figure 3.17: Handling of HsaHelQ to retain integrity and activity *in vitro*. (A). Coomassie stained 10% acrylamide SDS-PAGE analysis of freshly prepared HsaHelQ compared to "old" HsaHelQ that had undergone repeated freeze-thawing cycles. (B). 5% acrylamide TBE gel showing EMSA analysis of HsaHelQ binding to a model forked DNA when freshly purified (left panel) or having undergone repeated cycles of freeze-thawing (righ panel). Further detail of DNA binding by HsaHelQ is given in Chapter 4.

3.17. Effect of reducing agents on purified His-SUMO-HsaHelQ

For initial His-SUMO-HsaHelQ purifications, all buffers contained 5 mM DTT. Degradation following protein purification was also observed using these buffers. HsaHelQ contains 24 cysteine residues; a relatively large number for a protein of this size, suggesting that some HsaHelQ cysteines may form cystine pairs *via* disulphide bridges that would be expected to influence structure and function of the protein. We rationalized that by removing DTT from subsequent HsaHelQ purifications, HsaHelQ may be less inclined to degrade to obtain a higher yield of functional protein. Subsequently, we determined that although avoidance of excessive DTT from purification of HsaHelQ was advantageous, DTT was required as an assay additive to achieve DNA helicase activity, as described in Chapter 4.

3.18. Quantification of HsaHelQ protein concentration

Purified His-SUMO-HsaHelQ concentration was measured using NanoDrop, Bradford's reagent and bicinchoninic acid assay methods as standard, but these proved unreliable for this protein, when calculated protein concentrations from these methods were compared with one another, and with the presence of His-SUMO-HsaHelQ apparent in coomassie stained SDS-PAGE gels. These methods are described in section 2.17. A "non-interfering" protein concentration assay provided reproducible concentration values for His-SUMO-HsaHelQ. Determining

an accurate protein concentration is also a common issue for measuring the concentration of MthHel308 helicase.

3.19. Protein purification: HsaHelQ-N-Term-GST

HsaHelQ possesses an N-terminal region of approximately 300 amino acids that precedes RecA-like motor domains and shares no homology with other known proteins. Although absent from archaeal Hel308, this region of HsaHelQ can be found in eukaryotic Hel308 homologues such as proteins from *Mus musculus* and *Adineta vaga*. C-terminally GST tagged HsaHelQ N-terminal tail was over-expressed from pSN53. Soluble protein expression of HsaHelQ-N-Term-GST was achieved using BL21 AI *E. coli*, but during attempted purification this protein was unstable and degraded rapidly.

3.20. Protein purification: His-SUMO-HsaXRCC2

His-SUMO-HsaXRCC2 was purified using affinity chromatography by Ni^{2+}-NTA, Q and heparin, using ascending gradients of imidazole and NaCl from 50 mM to 1 M over 20 CVs as appropriate for each column before S75 size exclusion, described in section 2.16.2. Buffer pH was pH 8.0 given His-SUMO-HsaXRCC2 pI of 5.9 to give a charge of -20.1 for anionic exchange chromatography. A summary of His-SUMO-HsaXRCC2 protein purification by SDS-PAGE analysis can be seen in Figure 3.18. His-SUMO-HsaXRCC2 eluted in two peaks following size exclusion, indicating

protein aggregates or higher order protein complexes. Protein identity was verified by mass spectrometry.

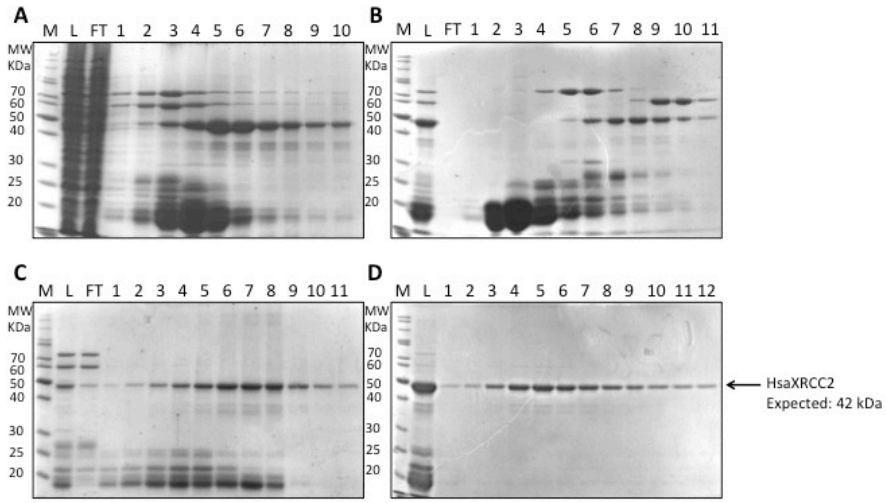


Figure 3.18: HsaXRCC2 protein purification:

Coomassie stained 12% SDS-PAGE analysis of His-SUMO-HsaXRCC2 purification using affinity chromatography. Purified His-SUMO-HsaXRCC2, is indicated by an arrow in panel D, and is consistent with the expected 42 kDa size of His-SUMO-HsaXRCC2. (A) His-SUMO-HsaXRCC2 was initially separated by Ni^{2+}-NTA affinity chromatography. Fractions 2-10 were pooled for loading onto Q-sepharose. (B) Q-sepharose chromatography did little to separate His-SUMO-HsaXRCC2 from contaminants. Fractions 5-11 were pooled for protein separation by heparin chromatography. (C) Heparin chromatography removed high molecular weight contaminants prior to separating proteins in fractions 4-8 by size exclusion chromatography. (D) His-SUMO-HsaXRCC2 was purified to homogeneity by size exclusion chromatography.

3.21 Protein purification: Rad51 paralogues

Rad51 paralogue proteins were much more challenging to purify than Rad51 and persistent contamination by other proteins was encountered. Rad51 paralogues are known to form 'insoluble aggregates',

a feature that makes them difficult to study[203]. This may be because of their propensity to form BCDX2 or CX3 complexes, and as such their attempted purification individually exposes surfaces for aggregation. His-SUMO-HsaRad51B and His-SUMO-HsaRad51C were initially purified from Sf9 cells following loading of clarified Sf9 cell lysate onto a His-trap column, and was eluted from the column using a 20 mM to 1 M imidazole gradient over 20 column volumes.

His-SUMO-HsaRad51D was partially purified however, in very small quantities. A summary of the purified Rad51 paralogues is shown in Figure 3.19. HsaRad51 paralogue identities were verified by mass spectrometry following trypsin digestion carried out by L. Manzi.

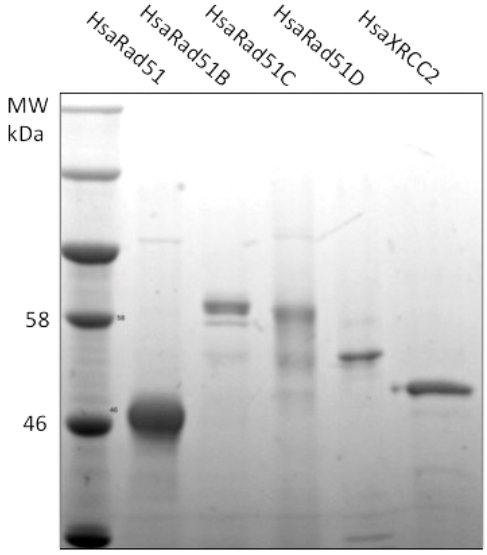


Figure 3.19: Purified His-SUMO tagged HsaRad51 and HsaRad51 paralogue proteins

Coomassie stained 12% acrylamide SDS-PAGE analysis displaying purified HsaRad51 and His-SUMO tagged HsaRad51 paralogue proteins. Proteins are observed at: His-HsaRad51 at 46 kDa (expected 42 kDa), His-SUMO-HsaRad51B at 55 kDa, His-SUMO-HsaRad51C at 54 kDa, His-SUMO-HsaRad51C at 55 kDa, His-SUMO-HsaRad51D at 38 kDa and His-SUMO-HsaXRCC2 at 44 kDa.

3.22. Verification of enzyme activity from purified proteins

Purified HsaHelQ, HsaRad51 and HsaRad51 paralogue proteins were initially subjected to qualitative assessment for minimal catalytic activities after purification, to verify their activity prior to more detailed analysis described in Chapters 4 and 5.

3.23. HsaRad51

In vivo, HsaRad51 polymerises on ssDNA to form a nucleoprotein filament (NPF). The filament adopts an active form for strand invasion

when in complex with ATP. However, unless stabilized by other proteins, HsaRad51 rapidly hydrolyses ATP into ADP, switching the NPF to an inactive form that quickly disassembles. This is part of the dynamic nature of the HsaRad51 NPF. HsaRad51 ATPase activity is stimulated in the presence of ssDNA[147], therefore HsaRad51 purified from *E. coli* as described in section 3.5.2 was tested for ATPase and DNA binding activities. ATPase activity of His-HsaRad51 was analyzed in ATPase reactions using radiolabelled ATP (^{32}P), and reaction products were separated in TLC plates as described in section 2.21. As shown in Figure 3.20, low ATPase activity of HsaRad51 in the presence of duplex DNA was stimulated in the presence of ssDNA, identified by increasing phosphate liberated from ATP. This is in agreement with expected behavior of HsaRad51, and as expected, contrasts with RecA, which shows very low ATPase activity, even in the presence of ssDNA.

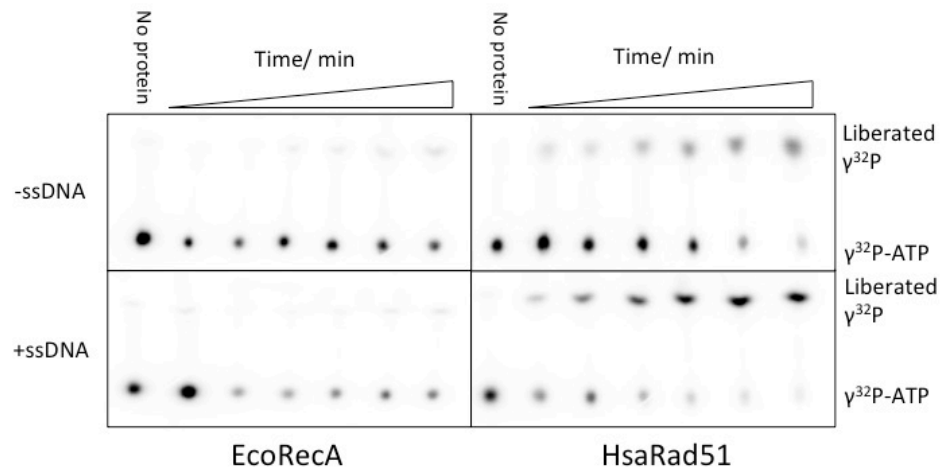


Figure 3.20: HsaRad51 ATPase activity is stimulated by ssDNA.
Separation of ATPase reaction products by TLC. ATPase reactions of 10 ul contained 10 nM Υ^{32}P-ATP in 1x ATPase buffer (20 mM TRIS-HCl pH 7.5, 5 mM $MgCl_2$ and 10% v/v glycerol) 4 µM protein as indicated and 250 nM ssDNA as indicated at 37^{O}C. Samples were taken at 5, 10, 15, 30, 60 and 120 minutes and stopped by addition of 20 mM EDTA. ATP hydrolysis, indicated by liberated Υ^{32}P, occurs in the absence of ssDNA, however, is stimulated in the presence of ssDNA.

Different divalent cations have been reported to influence Rad51 ATPase activity. Ca^{2+} has previously been shown to inhibit Rad51 ATPase activity and subsequently enhance Rad51 filament formation by keeping it locked in an ATP active state by inhibiting ATP hydrolysis[229]. Mg^{2+} and Ca^{2+} were tested on HsaRad51, and this behaved as expected, shown in Figure 3.21. No ATP hydrolysis was observed by HsaRad51 in the presence of Ca^{2+}, even when incubated for 1 hour, in comparison to Mg^{2+}, an ion essential for physiological ATP hydrolysis, where HsaRad51 could hydrolyse ATP over a period of 1 hour where 8-10 mM Mg^{2+} was most

efficient. Rad51 ATPase activity was still efficient in the presence of both Ca^{2+} and Mg^{2+}.

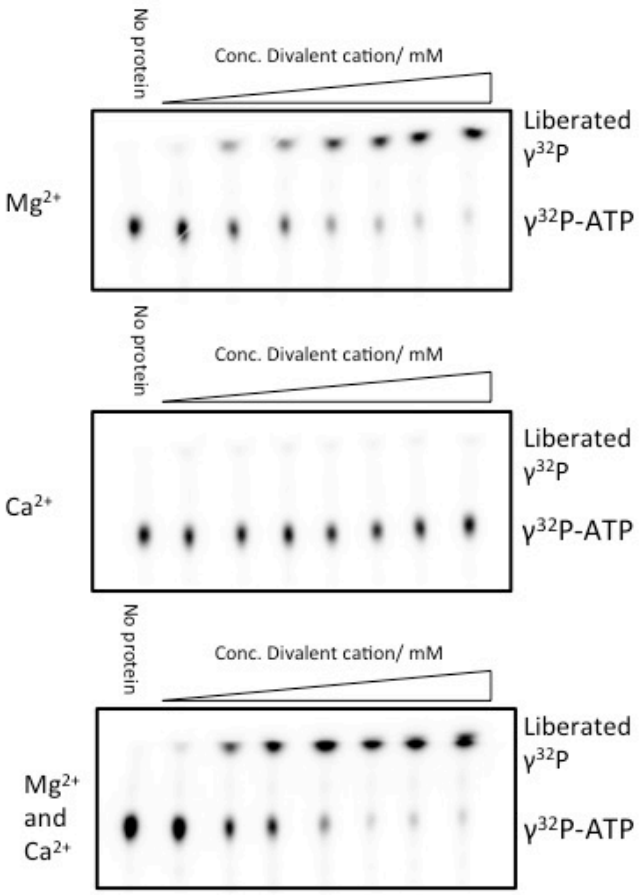


Figure 3.21: HsaRad51 ATPase activity is inhibited by Ca^{2+}.

Separation of ATPase reaction products by TLC. ATPase reactions of 10 ul contained 10 nM Υ^{32}P-ATP in 1x ATPase buffer without Mg^{2+} (20 mM TRIS-HCl pH 7.5, and 10% v/v glycerol) 250 nM ssDNA, 4 µM HsaRad51 and divalent cations were titrated from 0-10 mM as indicated at 37^{O}C. Reactions were stopped by addition of 20 mM EDTA after 60 minutes. ATP hydrolysis occurs in the absence of ssDNA, however, is stimulated in the presence of ssDNA.

HsaRad51 mutant proteins with amino acid substitutions in the Walker A motif (K133R), or the Walker B motif (D222A) or both (K133R/D222A) were also generated, and were expected to be defective

in ATP binding and/or ATP hydrolysis. ATPase assays confirmed that the Walker A-B HsaRad51 double mutant protein (K133R/D222A) could not hydrolyse ATP, even in the presence of ssDNA (Figure 3.22). A positive control, the purified helicase HsaHelQ hydrolyses ATP well, described more later. However, unexpectedly, both individual mutants of HsaRad51 (D222A and K133R) hydrolysed ATP.

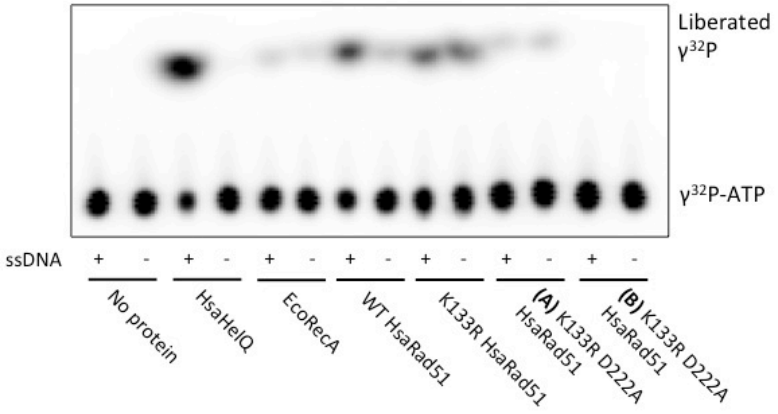


Figure 3.22: HsaRad51 mutant K133R D222A does not hydrolyse ATP
Separation of ATPase reaction products by TLC. ATPase reactions of 10 ul contained 10 nM Υ^{32}P-ATP in 1x ATPase buffer (20 mM TRIS-HCl pH 7.5, 5 mM $MgCl_2$ and 10% v/v glycerol) 4 µM protein and 250 nM ssDNA as indicated at 37^{O}C. Reactions were stopped by addition of 20 mM EDTA after 60 minutes. HsaHelQ was used as a control. ATP hydrolysis by HsaRad51 proteins occurs to varying extents across all mutant proteins, apart from (B) K133R D222A HsaRad51, where (A) and (B) refer to independent protein preparations.

HsaRad51 assembles into filaments along ssDNA[147], and we therefore tested HsaRad51 for its ability to bind DNA as a means of verifying HsaRad51 activity. HsaRad51 ability to bind simple DNA substrates were assessed using Electrophoretic Mobility Shift Assays

(EMSA) as described in section 2.22. This is a biochemical approach that relies upon a protein forming stable complexes with labeled DNA acting as a reporter for the position of shifted DNA-protein complexes; Complexes of higher molecular weight move more slowly through the gel, in comparison to labeled unbound DNA substrate.

EMSAs to verify that HsaRad51 could bind to DNA substrates showed HsaRad51-DNA complex formation with ssDNA and partial duplex, as expected, but no binding was observed between HsaRad51 and linear duplex DNA, which in contrast to other studies. This may be indicative of a problem with the protein, or with the assay conditions used here.

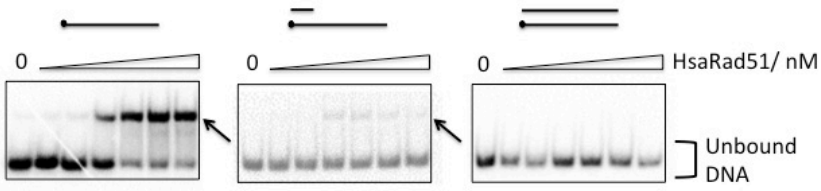


Figure 3.23: EMSA analysis of Rad51-DNA interactions. 5% acrylamide TBE gels show HsaRad51 forming stable complex with ssDNA and partial duplex, but not linear duplex. Reactions were for 37^{o}C for 20 minutes and contained 5mM $MgCl_2$, 5mM EDTA and 0.2nM DNA. HsaRad51 was titrated at 0, 1, 5, 25, 62.5, 125 and 250nM against each substrate. Arrows indicate protein-DNA complexes. Filled dots indicate 5' γ^{32}P labeled ends.

3.24. Verification of minimal HelQ activity

Purified HsaHelQ was assayed to verify that HsaHelQ was catalytically active as a ssDNA-stimulated ATPase, Figure 3.24, as expected from previous work [4]. This was carried out using TLC based ATPase assays as described in section 2.21.

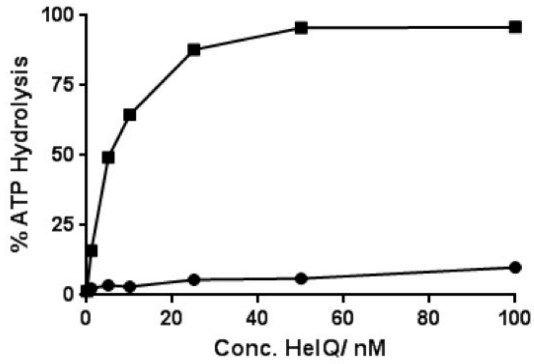


Figure 3.24: HsaHelQ ATPase activity is strongly stimulated in the presence of ssDNA. Reactions were at 37^{o}C for 1 hour and contained 5 mM $MgCl_2$ and 10 nM γ^{32}P labeled ATP, and ATPase activity was assessed from TLC plates. Reactions were with 250 nM ssDNA (filled squares) and without ssDNA (filled circles). His-SUMO-HsaHelQ was titrated at 0, 5, 10, 25, 50 and 100 nM.

The minimal expected activity of purified HsaHelQ was also assessed in its ability to bind simple DNA substrates in EMSAs as described in section 2.22. As demonstrated in Figure 3.25, HsaHelQ forms stable complexes with ssDNA and linear duplex DNA.

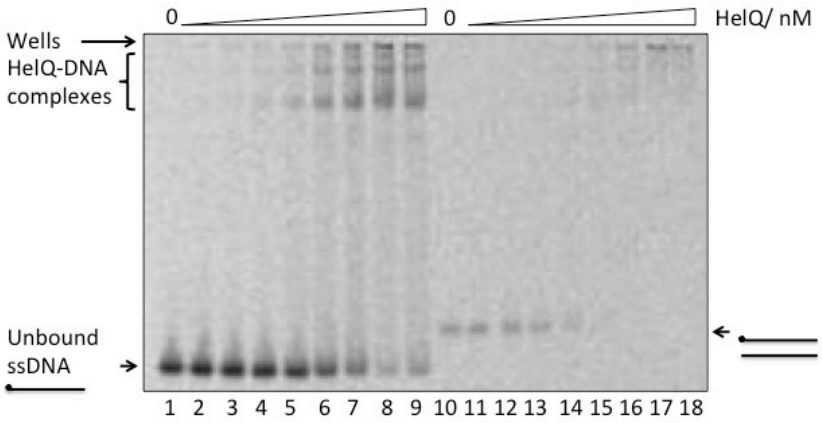


Figure 3.25: HsaHelQ DNA binding. 5% TBE gel EMSA showing binding of HsaHelQ to ssDNA forming distinct stable protein-DNA complexes. Some binding to duplex DNA is also observed. Reactions were at RT for 10 minutes and contained 1x HB, and 25 nM DNA. HsaHelQ was titrated at 0, 5, 10, 20, 40, 50, 80, 100 and 160nM against each substrate. Protein-DNA complexes are highlighted. Filled dots indicate 5' Cy5 labeled ends. HsaHelQ appears to form more stable complexes with ssDNA than with linear duplex DNA, as shown in lanes 6-9.

3.25. Summary

At the start of the project I set out to purify Human proteins HelQ, Rad51 and the five Rad51 paralogues. The aim was to use these to facilitate analysis of HsaHelQ helicase activity in isolation and how this activity may be affected on substrates for Rad51 and Rad51 paralogues. HsaHelQ and HsaRad51 proteins were successfully purified, and their minimal activities verified. The functionality of these proteins is explored in more detail in Chapter 4. Importantly for ongoing studies on HsaHelQ, its purification through several iterations has created a new reliable procedure with improved understanding of how to handle this protein during purification and experimentation. Due to difficulties purifying

HsaRad51 paralogues to homogeneity reproducibly at yields sufficient for biochemistry, adequate reproducible activity assays were not possible in the time of this project, whilst it was important to focus in particular on studies of HsaHelQ and HsaRad51.

Chapter 4: Biochemical analysis of human HelQ protein

4.1. Introduction

Human HelQ protein, was first isolated and determined to be an active DNA helicase in 2001[4], but has been difficult to generate in microgram or milligram quantities, therefore hindering biochemical and structural studies. In 2005 an archaeal homologue of human HelQ, then called Hel308a[5] or Hjm[230], was identified as part of a general enlightenment that archaeal DNA metabolism may be a good model system for more complex homologous processes in eukaryotes. This led to atomic resolution crystal structures of archaeal Hel308 proteins, and further characterization of Hel308 activities by biochemical and genetic analyses. From these comparisons have been made with HsaHelQ that are mainly reliant on conserved protein features identified at amino acid sequence level.

I generated milligram quantities of catalytically active HsaHelQ (Chapter 3) that has facilitated trials for structural analysis of HsaHelQ, work that is ongoing and beyond the current scope of this thesis. In Chapter 4 and Chapter 5 biochemical characterisation of purified HelQ is described using several techniques. The aim of the work presented in Chapter 4 was to analyse HelQ biochemistry as a helicase enzyme, combined with experimental and *in silico* analysis of its oligomeric state, to begin to ascertain more information about the mechanism of action. This generated information for comparison of the biochemical activities of HsaHelQ with the much better characterized archaeal Hel308.

4.1.1. A note about Hel308/HelQ nomenclature

Metazoan Hel308 was named in 2001 in reference to its homology to *mus308,* which at that time was a predicted helicase-polymerase identified by genetics in *Drosophila*, and which is now known to encode PolQ protein. In line with this, Hel308 was renamed HelQ in 2010. Archaeal Hel308 (Hel308a) was so named in 2005, using the historical precedent of human Hel308. Contemporaneously, archaeal Hel308 was also named Hjm, for "Holliday junction migration"[231], although it is now generally accepted that archaeal Hel308 is not capable of "branch migration" in the way analogous to RuvAB or RecG, but is a ssDNA stimulated ATPase-helicase with a preference for targeting forked DNA with ssDNA gaps. Therefore, HelQ is used to mean the human homologue of archaeal Hel308.

4.2. Hel308 atomic resolution crystal structures

To date there is no atomic resolution structure of HsaHelQ. There are regions of HsaHelQ (e.g. helicase motifs) that show strong sequence conservation with archaeal Hel308 and other helicases and with HsaPolQ, facilitating structural modeling, that gives some insight into HelQ function, see especially Chapter 5. However, there are also significant differences and non-homologous regions between HsaHelQ and archaeal Hel308 sequences, suggesting additional roles for HsaHelQ.

Crystal structures of archaeal Hel308 are available from archaeal species *Sulfolobus solfataricus* (Sso, PDB: 2VA8)[190], *Archaeoglobus fulgidus* (Afu, PDB: 2P6R)[191] and *Pyrococcus furiosus* (Pfu, PDB: 2ZJ8)[232]. These revealed a five-domain architecture, which for AfuHel308 was solved in complex with partial duplex DNA, the minimal substrate for Hel308 helicase activity, although this structure lacks nucleotide co-factor[191]. This provided useful information with regard to Hel308-DNA interactions; however, Hel308 targets forked structures as its preferred substrate *in vitro*, where it is capable of unwinding 85% of a lagging strand model replication fork in comparison to unwinding only 30% of a partial duplex with a 3'-OH overhang[5].

PfuHel308 crystal structures were solved with and without (apo-Hel308) bound non-hydrolysable ATP analogue[232]. Comparison of such structures can identify conformational movement induced in the co-factor bound state, but it is evident from PfuHel308 structures that there is little conformational change caused by AMP-PNP binding. However, this may be different in the presence of DNA and a nucleotide cofactor such as AMP-PNP, or if Hel308 is bound to a preferred substrate such as a fork.

The AfuHel308 atomic resolution structure in complex with partial duplex shows the extensive contacts made between the Hel308 and the DNA, where the ssDNA of the partial duplex lies across the top of the RecA like motor domain and domain four, the helicase ratchet, and then threads through the central channel[191]. Numerous protein-DNA contacts are also made in domain five, a 'molecular brake' located at the duplex

region of the partial duplex[191]. Altogether, this atomic resolution structure reveals approximately sixty contacts made between Hel308 and the partial duplex DNA molecule. This provides a useful basis for interactions between HsaHelQ and DNA in the absence of a bona fide HsaHelQ atomic resolution structure in complex with DNA.

Further to studies of archaeal Hel308 proteins, the crystal structure of the helicase domain from human TLS polymerase PolQ (Polθ) was recently solved at atomic resolution[189], referred to as HsaPolQHD from now onwards. HsaPolQHD is homologous to HsaHelQ, however it lacks helicase activity. HsaPolQHD structures were in complex with DNA, and with a non-hydrolysable ATP analogue, in addition to apo structures. These identified a five-domain architecture similar to archaeal Hel308, but also gave interesting information about the oligomeric state of HsaPolQHD, factors utilized in attempts at structural modeling of human HelQ.

4.3. Structural modeling of HsaHelQ

Several online protein structural modeling tools are available for instances where a query protein has no structural data available. These include SWISS-PROT, Phyre 2 and I-TASSER among others. Phyre 2 (Protein Homology/ analogY Recognition Engine V 2.0), a protein fold recognition server[218] was used to generate a molecular model for HsaHelQ because, alongside I-TASSER, (Iterative Threading ASSEmbly Refinement)[219] it is ranked one of the most accurate online molecular modeling tools

available. Phyre 2, and the other modeling programs listed above, are template based modeling (TBM) tools. Other protein modeling methods include assembly of modeled structures using small regions of homology for which there is known structure, and a method that applies energy calculations to a given amino acid sequence to reach an optimal predicted conformation for the sequence, known as simulated folding. TBM is regarded as a reliable modeling tool among the bioinformatics and molecular modeling communities. An attractive feature of TBM is that through sequence alignments against the query, evolutionary patterns are detected and accounted for in model development. Phyre 2 gives the user the option to model a query sequence solely using TBM in 'normal' mode, or using 'intensive' mode, where simple mathematical based simulated folding by Phyre 2 'Poing' is incorporated into the standard Phyre 2 modeling process[218].

Phyre 2 was used to generate a molecular model for HsaHelQ, using the HsaHelQ amino acid sequence obtained from UniProt accession number Q8TDG4, under the intensive modeling mode as described in sections 2.26.1 and 2.26.3. In summary, Phyre 2 generates multiple sequence alignments through profile-profile matching through a process called HHblits[218]. Using the multiple sequence alignaments, secondary structure prediction is generated through PSIPRED. Phyre 2 combines the results from HHblits and PSIPRED to generate a hidden markov model (HMM). This HMM is queried *via* HMM-HMM matching (HHSearch) against HMMs of proteins with known structure in the Phyre 2 fold library.

Insertions and deletions of the query sequence are accounted for *via* loop modeling in which the sequence is queried against another library containing fragments of protein sequences with known structure. At this point a general topology for the protein backbone has been generated. Finally the side chains of the amino acid residues must be modeled. This is done using a side chain rotamer library to obtain the most accurate positions for the side chains, whilst also avoiding steric clashes. The Phyre 2 modeling process is shown in Figure 4.1.

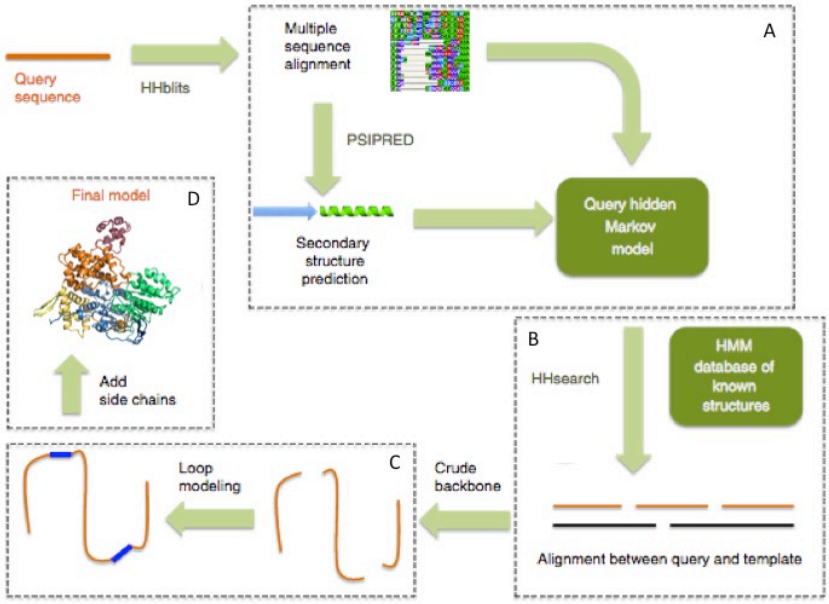


Figure 4.1: Phyre 2 molecular modeling flow diagram.

(A) HHblits uses the query sequence to generate multiple sequence alignments. These are used to generate secondary structure predictions using Psipred that combined with the multiple sequence alignment generate a query hidden markov model. (B) The HMM is queried against a Phyre database of known structures. This generates a crude backbone from which a general topology is generated by loop modeling. (D) Finally, amino acid side chains are added and refined to avoid potential steric clash to generate a final model. Adapted from Kelley *et. al.*[218].

Phyre 2 successfully generated a molecular model for HsaHelQ under the intensive modeling mode (Figure 4.2). The five top hits for sequence homology of proteins of known structure used to generate the multiple sequence alignment following the HHblits search are shown in Table 4.1. The top hit for the structural model for HsaHelQ generated by Phyre 2 was the HsaPolQHD, residues 67-894 of human DNA polymerase θ (PDB number 5AGA, deposited by J. A. Newman *et. al.*[189]) in complex with AMP-PNP. This claims 100% confidence and 37% sequence similarity to the HsaHelQ query. Other top hits include the Ski-2 RNA helicase Brr2, to which HsaHelQ and Hel308 proteins are related. Hel308 structures appear on the list of PDB templates used to generate the model. However, they featured lower down the list than HsaPolQHD and Ski2 helicases due to lower overall % identity in comparison to other proteins. I-TASSER generated the same structural homology hits as Phyre 2, increasing confidence in the structural model generated.

Table 4.1: Templates used for HsaHelQ modeling by Phyre 2

Top five templates used for HsaHelQ modeling by Phyre 2 ranked by % amino acid identity, following entry of HsaHelQ amino acid sequence obtained from UniProt accession numbers Q8TDG4. The value for confidence is a percentage, describing how complete each stage of the Phyre 2 modeling process was, including amino acid sequence alignments and acquisition of secondary structure data. Templates of 90% confidence or more are considered useful.

Rank	**PDB template**	**Confidence**	**% amino acid sequence identity**	**Corresponding protein structure**
1	5AGA	100	37	Human POLQHD in complex with AMP-PNP
2	4F92	100	22	Brr2 helicase region
3	4BGD	100	23	Brr2 helicase
4	2ZJ8	100	28	DNA helicase Hjm apo state
5	2VA8	100	26	SsoHel308

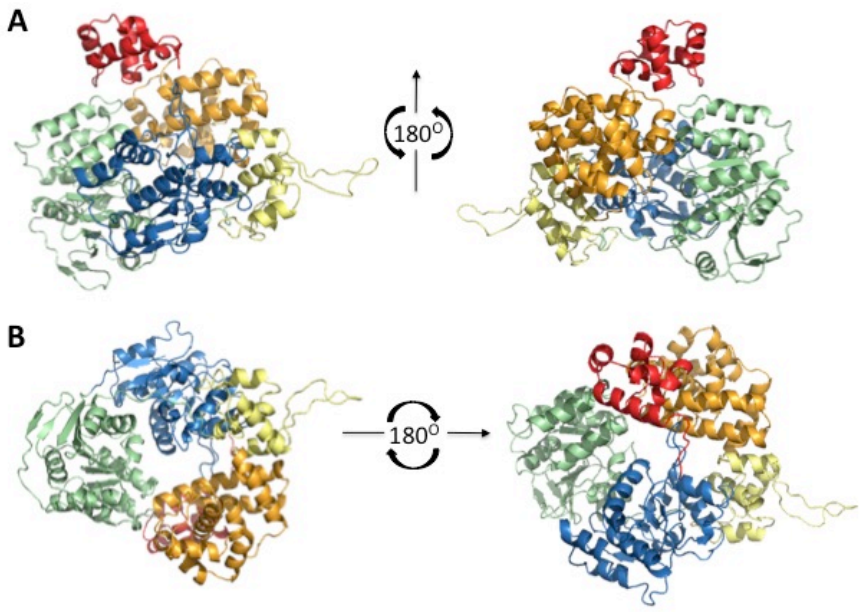


Figure 4.2: Phyre 2 HelQ model.

A Phyre 2 HsaHelQ model was generated using using FASTA amino acid sequences for HsaHelQ obtained from UniProt accession number Q8TDG4. Domains are coloured as follows for Figures 2A and 2B: domain 1 (RecA like): blue; domain 2 (RecA like): green; domain 3 (winged helix domain): yellow; domain 4 (helicase ratchet) orange and domain 5 (molecular brake) red. Models were rendered using MacPymol.

(A) HsaHelQ model displayed 'forwards' and rotated 180^{O} on a vertical axis to show the 'reverse' face. These orientations show domain 5, the molecular brake 'hovering' above the main structure where domans 1-4 are clustered.

(B) HsaHelQ model displayed from 'underneath' and rotated 180^{O} on a horizontal axis. The central channel is visible through the 'ring' formed by domains 1-4. Domains 1 and 2 (blue and green) that form the RecA like motor domain can be seen adjacent to one another and domain 3, the WHD (yellow), packs closely to domain 1 (blue). All orientations show domain 3 forms an extended hairpin structure that extends from the protein 'core'.

One interesting outcome of the Phyre 2 structural model for HsaHelQ was that no sequence alignment was possible for the first 300 residues of HelQ from the PDB library. This region corresponds to the N-terminal tail and could not be modeled against other known protein folds. Therefore Phyre 2 displayed this part of HsaHelQ is as an unfolded rod of amino acids, which for clarity is omitted from the structural models of HsaHelQ shown. Further information on this region of HsaHelQ, and an attempt to study it biochemically, are described in sections 3.7, 4.5, 4.7, 4.8 and 4.11. Another structural HsaHelQ model was generated using I-TASSER, to investigate the differences between models using different structural predictors. Comparison between HsaHelQ models generated by Phyre 2 and I-TASSER is shown in Figure 3.3. Overall they adopt a similar domain organization, although the Phyre2 model gave a more compacted protein than I-TASSER, and the presence of a significant protein loop emanating from the helicase ratchet winged-helix domain 3 in the Phyre2 model was not present from I-TASSER.

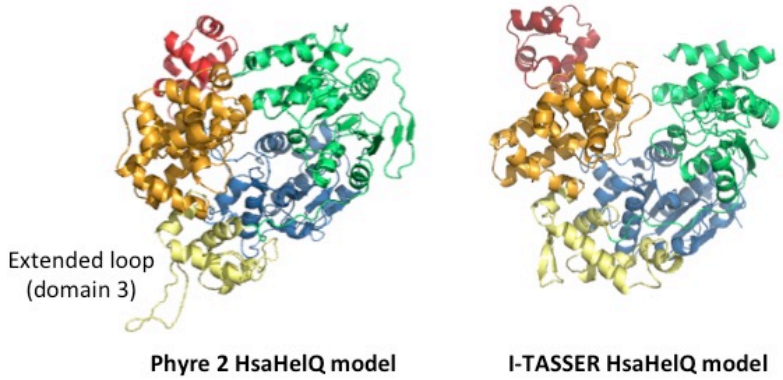


Figure 4.3: HsaHelQ models generated by Phyre 2 and I-TASSER are similar. Models were generated using Phyre 2 and I-TASSER using FASTA amino acid sequences for HsaHelQ obtained from UniProt accession number Q8TDG4. Domains are coloured as follows for Figures 2A and 2B: domain 1 (RecA like): blue; domain 2 (RecA like): green; domain 3 (winged helix domain): yellow; domain 4 (helicase ratchet) orange and domain 5 (molecular brake) red. Models were rendered using MacPymol.

Much biochemical analysis of archaeal Hel308 has utilized the enzyme from the euryarchaeon *Methanothermobacter thermautotrophicus*. There is no atomic resolution crystal structure for this protein, therefore a structural model was generated using Phyre 2, shown below in Figure 3.3. The top five highest scoring hits are shown in table 3.1B. Many of these structural templates, like for HelQ analysis, are Ski-2 like RNA helicases, reflecting Hel308 similarities to these types of helicases. It should be noted that although not shown in Table 3.1B, PDB template 5AGA for HsaPolQHD in complex with AMP-PNP was ranked 6th to generate the MthHel308 model.

Table 4.2: Templates that were used for Phyre 2 modeling of MthHel308

Rank	PDB template	Confidence	% amino acid sequence identity	Corresponding protein structure
1	5DCA	100	15	Brr2 in complex with Prp8 Jab12 domain
2	4F92	100	21	Brr2 helicase region
3	4BGD	100	15	Brr2 in complex with Prp8 Jab12 domain
4	3JCR	100	23	Full length Brr2
5	2VA8	100	28	SsoHel308

Top five templates used for MthHel308 modeling by Phyre 2 ranked by % amino acid identity, following entry of MthHel308 amino acid sequence obtained from UniProt accession numbers O26901. The confidence value is explained in Table 4.1.

4.4. Comparison of solved AfuHel308 and HsaPolQHD atomic resolution structures and MthHel308 and HsaHelQ structural models

Comparative analysis between atomic AfuHel308 and HsaPolQHD atomic resolution structures and MthHel308 HsaHelQ structural models was made, summarized in Figure 3.4, in which structures were orientated in the same manner using AfuHel308 in complex with partial duplex DNA as standard. All structures show very similar overall topologies, and each domain is located similarly relative to one another. Initially, HsaPolQHD and HsaHelQ model structures in Figure 4.4 may seem more compact, however, this is due to the positioning in order to make both domain 5 and the central channel more visible. A central channel within each structure is formed from a ring of the core helicase domains one (RecA-like), two (RecA-like) and four (helicase ratchet). Although the channels for HsaPolQ and HsaHelQ may initially seem smaller, this is also due to the

positioning of these molecules as not to obstruct domain five, shown in red, from the images. The partial duplex could be accommodated readily in the central channel in these models, analogous to the AfuHel308-DNA complex structures, where DNA contacts the RecA like motor domain (blue and green) and domain four, the helicase ratchet (orange).

Notably, there is strong structural conservation of some features that are mechanistically important to Hel308 or HsaHelQ function, or are predicted to be. These include tight-packing of the WHD (yellow) against RecA-like domain 1 (blue), which is explored further in Chapter 5. In addition, the helicase ratchet helix is located in the same position in the ratchet domain in all model structures, consistent with its crucial role in DNA translocation. Accessory domain 5, a 'molecular brake' can be observed to protrude from the compacted 'ring' comprised of domains 1-4. A major difference between structures is the presence of a beta wing located in the WHD in the HsaHelQ model that is not observed in the other structures. Overall, the structures are very similar, and additional elements of secondary structure in some domains may be attributed to the differences in size of these proteins, for example, HsaHelQ is larger than MthHel308 at 124 kDa and 74 kDa respectively. Additional differences may be explained that some of these structures are theoretical models and therefore are not entirely accurate.

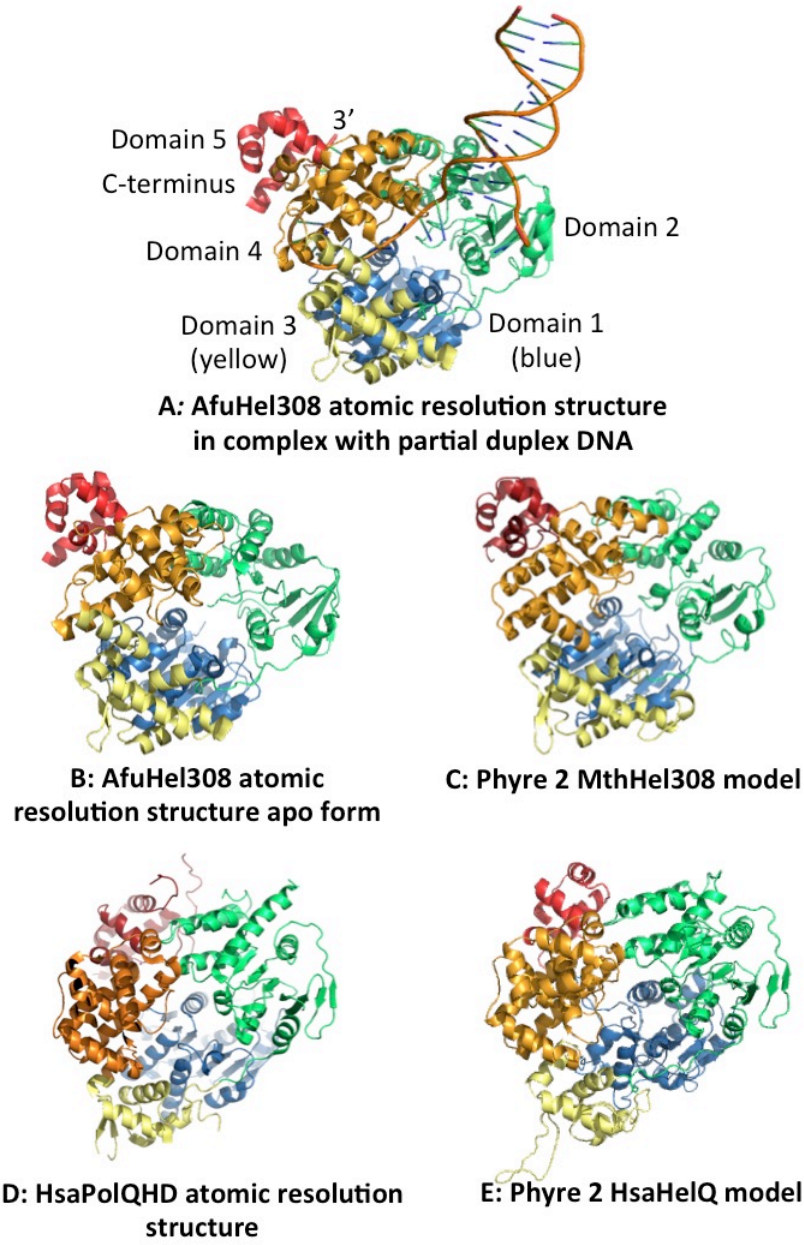


Figure 4.4: Comparison of archaeal Hel308 and HsaHelQ atomic resolution structures and Phyre 2 structural models.

AfuHel308 atomic resolution structure in complex with partial duplex DNA (and corresponding apo form) in comparison with Phyre 2 MthHel308 and HsaHelQ model structures and HsaPolQHD atomic resolution structures. All structures are orientated in the same position as AfuHel308 as if in complex with partial duplex DNA. Domains are coloured as follows: domain 1 (RecA-like): blue; domain 2 (RecA like): green; domain 3 (WHD): yellow; domain 4 (helicase ratchet): orange; domain 5 (molecular brake): red. HsaPolQHD and HsaHelQ structures are orientated to make the

central channel more visible that affects visibility of domain 5 (red). Coordinates for (A), (B) and (D) were downloaded from PDB using codes 2P6R, 2P6U and 5AGA respectively. Models (C) and (E) were generated using Phyre 2 from FASTA amino acid sequences obtained from UniProt accession numbers O26901 and Q8TDG4 respectively. For clarity, amino acids 1-300 are omitted from (E) as these could not be modeled by Phyre 2 and were shown as a rigid rod. Structures and models were rendered using MacPymol.

4.5. Oligomeric state of HelQ: Biochemical analysis

4.5.1. Analytical gel filtration

Archaeal Hel308 is thought to function as a monomer, and crystallises as such, but the oligomeric state of human HelQ is less clear. Initial studies by Marini *et. al.*[4] suggested that HelQ may form higher order structures based on gel filtration data, although without defining the extent or characteristics of the oligomerisation. During final stages of HelQ purification (described in sections 2.16.3 and 3.14) by size exclusion chromatography, two peaks of HelQ were eluted from a Superdex S200 column, indicative of protein aggregates or complexes of differing molecular weight. To investigate this further, analytical gel filtration (AGF) using a Superdex S200 10/30 column was used to determine approximate molecular weights of the protein species contributing to each elution peak on the UV trace, as described in section 2.25. AGF works on the same principle as large scale size exclusion chromatography, however, AGF columns are packed more tightly and operate under high pressure, therefore offering higher resolution for determining sizes of proteins or

protein complexes. However, the shape of the protein or complex must also be considered using this technique because it influences protein migration through the column matrix and can account for unexpected protein sizes. AGF of HsaHelQ confirmed the presence of two distinct peaks, one major peak and one minor peak. Comparison of the AGF trace with an AGF calibration curve showed the two HelQ peaks had approximate molecular weights, one of between 670 kDa and 156 kDa, and another at 44 kDa. His-SUMO-HelQ has a molecular weight of 137 kDa, including the 13 kDa His-SUMO tag. The major, early HsaHelQ peak may correspond to oligomeric HelQ, but an explanation for the 44 kDa peak is more difficult to propose, but could be due to proteolysis or if a population of HelQ molecules were to bind non-specifically to the AGF resin this might increase the protein retention time manifest as a lower molecular mass than expected.

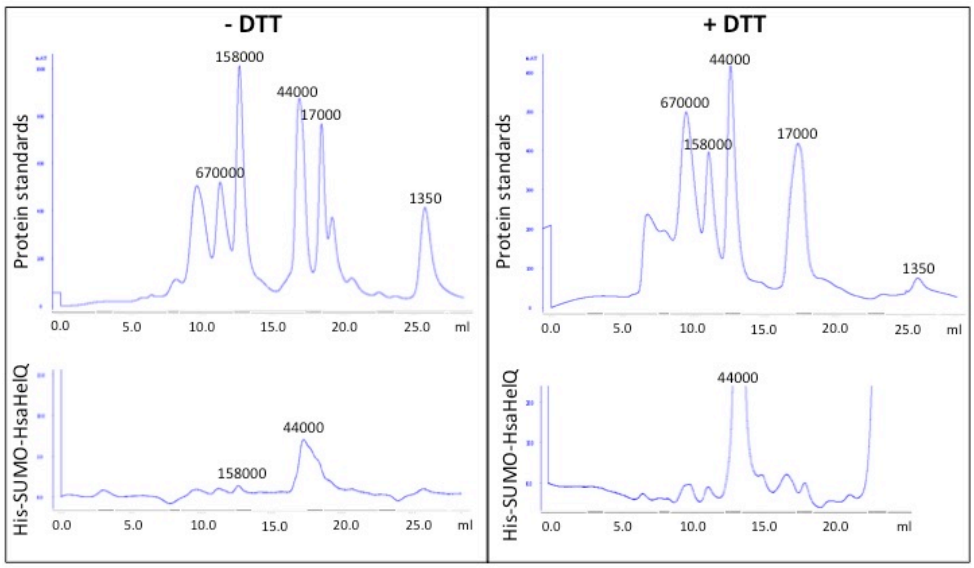


Figure 4.5. Analytical gel filtration of purified His-SUMO-HelQ protein in buffer containing either zero or 25 mM DTT. Molecular mass size markers are lablled as indicated. In both instances peaks of His-SUMO-HelQ were determined at high molecular mass >400 kDa and at 44 kDa, detailed further in the main text.

4.5.2. Native polyacrylamide gel electrophoresis

Blue native PAGE (BN-PAGE) was used to further investigate the potential oligomerisation of HelQ revealed by AGF. This technique is a biochemical gel based method that can be used for qualitative detection of protein oligomers and protein-protein interactions, described in section 2.24. This method omits SDS from the electrophoresis process and relies upon non-specific binding of Coomassie blue G-250, that carries a negative charge, to the protein therefore enabling electrophoresis, as SDS is not available to provide a negative charge, thus enabling protein electrophoresis in their native state. Gradient gels are used; therefore

proteins and protein complexes spanning a variety of molecular weights can be studied.

As previously shown in Chapter 3, section 3.14, under denaturing SDS-PAGE conditions, His-SUMO-HsaHelQ migrates consistently with its predicted molecular mass of 137 kDa. Purified His-SUMO-HsaHelQ loaded onto BN-PAGE gels showed two distinct bands migrating either somewhat more slowly than 720 kDa, or between the 720 kDa and 480 kDa size markers (Figure 4.5A). Boiling of the protein sample showed no protein on the blue native page gel, indicating that the protein had been destroyed. Normally, boiling of the protein would denature any higher order structures and the protein should migrate as a monomer at 137 kDa. It is likely that glycerol, a component of the storage buffer at 30% w/v, in the protein storage buffer has contributed to this artifact.

There was significant alteration to the migration properties of His-SUMO-HsaHelQ in native gels in response to DTT (Figure 4.5B), which had been omitted from His-SUMO-HsaHelQ purifications (section 3.17) as a rationale to generate much larger quantities of protein than previously, in response to the 24 cysteine residues present in HsaHelQ; the unusual abundance of Cys- residues suggested that they may be required for protein stability *via* disulphide bond formation. On addition of DTT, only a single His-SUMO-HsaHelQ species was observed on native gels, migrating at approximately 500 kDa, Figure 3.5B. This version of His-SUMO-HsaHelQ therefore migrated similarly, but not identically to one His-SUMO-HsaHelQ species in the absence of DTT. This is unusual and may be

explained by a change in protein conformation to a more open state, therefore causing it to migrate more slowly through the gel, rather than a gain in molecular weight *via* addition of another monomer. The migration property of this single HelQ species observed in the presence of DTT is consistent with a tetramer, which is the same oligomeric state of HsaPolQHD when solved as a crystal structure. It is interesting to note that in data presented later in this chapter, DTT was found to greatly stimulate HsaHelQ helicase activity.

SDS was also added to His-SUMO-HsaHelQ prior to native PAGE to determine if apparent oligomerisation could be disrupted, Figure 4.5C. As the concentration of SDS was increased, there was a corresponding decrease in tetramer His-SUMO-HsaHelQ co-incident with the appearance of a strong band migrating at about 400 kDa and some protein of similar size to the tetramer but migrating slightly more slowly. The 400 kDa species may be consistent with dimer or trimer forms of His-SUMO-HsaHelQ, The other form may be an intermediate state of HsaHelQ induced by SDS to adopt an open conformation before transition to the predominant lower oligomeric state shown in Figure 3.5C. Interestingly, no His-SUMO-HsaHelQ monomers were observed on addition of SDS, although they may have run off the bottom of the gel.

Natural substrates and ligands of HsaHelQ, ATP and Mg^{2+} and ssDNA, were added to His-SUMO-HsaHelQ prior to native PAGE to determine if they had any effect on the oligomeric form of HsaHelQ that might give evidence for the functionally active form of the enzyme,

without the addition of exogenous DTT (Figure 3.5D). However, no change was observed and His-SUMO-HsaHelQ remained at sizes consistent with possible hexamers and tetramers.

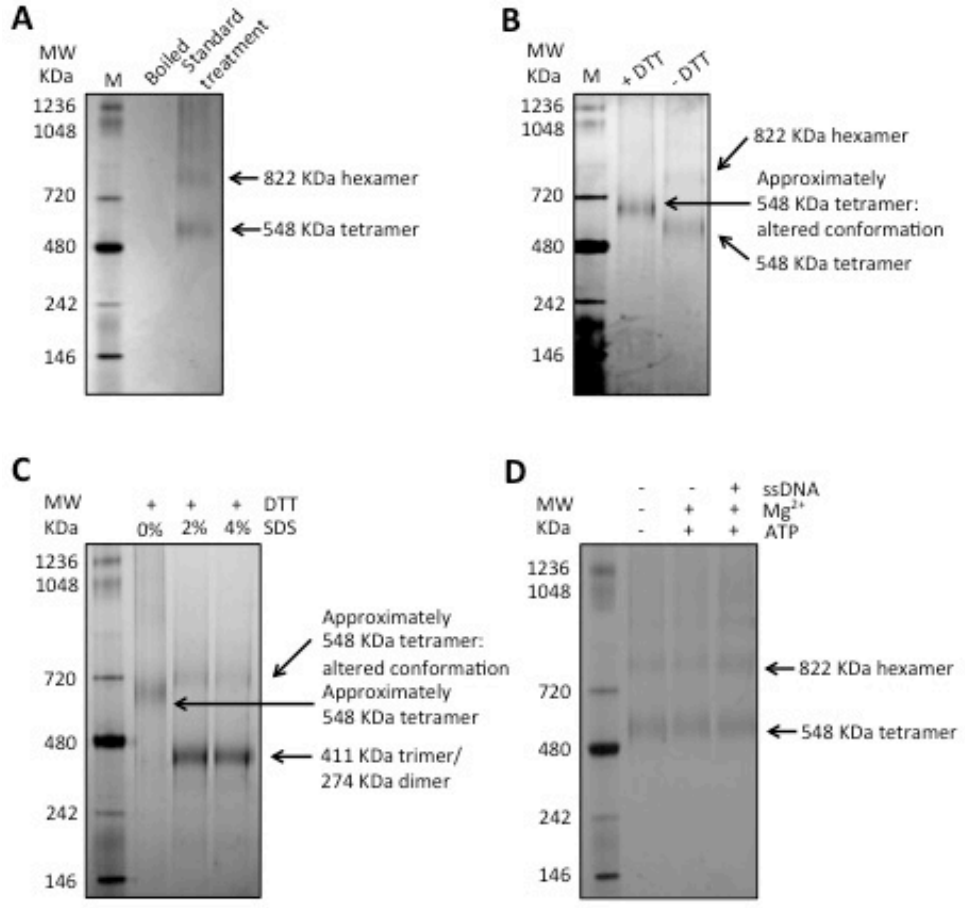


Figure 4.6: HsaHelQ forms oligomers. BN-PAGE was used to investigate the oligomeric state of HelQ. Pure HsaHelQ was loaded onto BN-PAGE.

(A) Boiling of HsaHelQ reveals total loss of protein in comparison to HsaHelQ under standard conditions in the absence of DTT.

(B) In the absence of DTT two bands could be observed at approximately 822 kDa and 548 kDa. On addition of DTT only one band could be observed at molecular weight of 548 kDa.

(C) SDS was titrated against HsaHelQ in the presence of DTT. This shows a loss of a single band at approximately 548 kDa into two bands at approximately 548 kDa and 441 kDa/ 274 kDa.

(D) HsaHelQ oligomers were tested in the presence of components required for HsaHelQ ATPase activity: ssDNA, Mg^{2+} and ATP in various combinations as indicated and in the absence of DTT. These showed no effect on HsaHelQ oligomeric state and two bands remain at approximately 822 kDa and 548 kDa.

4.5.3 Molecular modeling HsaHelQ oligomeric state by comparison with PolQ

Recently, the atomic resolution crystal structure for the helicase domain of PolQ (HsaPolQHD) has been solved, reporting that this protein is a tetramer[189]. The biological assembly for the HsaPolQHD can be seen in Figure 4.7. How this functions with regard to PolQ function in not clear, although the authors suggest that the helicase domain can act to melt and re-associate DNA ends for alternative forms of non-homologous end joining (alt-EJ) such as microhomology mediated end joining (MMEJ) or in synthesis dependent strand annealing (SDSA)[189]. This crystal structure and its proposed function is interesting, in the context of our observations from AGF and BN-PAGE that HsaHelQ may oligomerize into tetramers.

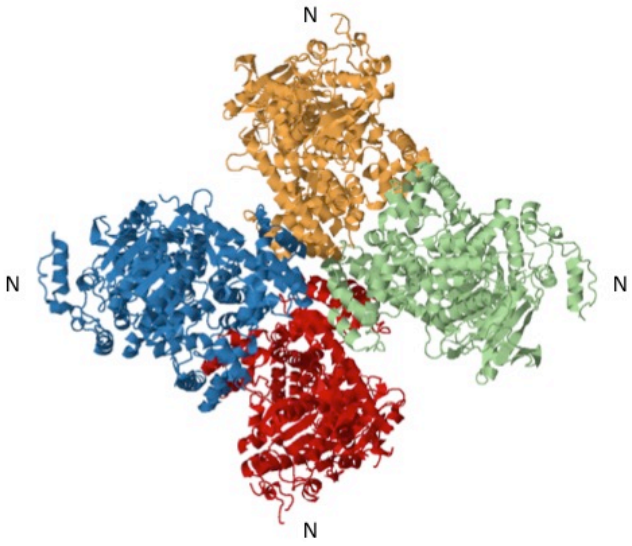


Figure 4.7: HsaPolQHD biological assembly.

The atomic resolution structure obtained for HsaPolQ (PDB 5A9F Newman *et. al.*[189]) was tetrameric. In this figure each monomer is highlighted in a different colour, and 'N' indicates the N-terminus of each monomer. Each HsaPolQHD monomer interacts with each other monomer *via* three interaction sites, detailed further below.

I therefore analyzed the tetrameric HsaPolQHD structure for comparison with amino acids present in HelQ. The authors of the HsaPolQHD structure presented an extensive study of HsaPolQHD interaction interfaces that occur between monomers. Three independent interfaces were identified, in which each contacts a different monomer to form an overall tetramer, shown in Figure 4.9. Amino acid sequences of HsaPolQ and HsaHelQ were compared overall and in short local alignments, to identify residues in HsaHelQ that are conserved with interface residues from HsaPolQHD. This was successful and these residues were then mapped onto the HelQ Phyre2 model, as presented below in Figure 4.8.

Overall alignment of HsaHelQ and full-length HsaPolQ amino acid sequences revealed 10.8% identity and 16.8% similarity using Emboss Needle alignments (Table 4.3). Emboss water local alignments, in which regions of homology between two protein sequences are identified, were made to assess conservation of proposed interaction surfaces, and revealed 34.0% identity and 54% similarity between HsaHelQ and HsaPolQHD, as described in section 2.26.2.

Table 4.3: Analysis of similarity and identity between HsaHelQ: HsaPolQ and HsaHelQ: MthHel308.

Alignment tools used to analyse % identity and similarity between human HelQ and other related proteins. Note, that Clustal Omega only provides a percentage identity matrix, not similarity.

Alignment	**Uniprot accession numbers**	**Tool**	**Alignment Type**	**Amino acid alignment outcome**
HsaHelQ against HsaPolQ	HsaHelQ: Q8TDG4 HsaPolQ: O75417	Emboss Water	Pairwise local alignment	34.0% identity; 52.8% similarity
		Emboss Needle	Pairwise global alignment	10.8% identity; 16.8% identity
		Clustal Omega	Multiple sequence alignment	35.1% identity
MthHel308 against HsaHelQ	MthHel308: O26901 HsaHelQ: Q8TDG4	Emboss Water	Pairwise local alignment	23.4% identity; 42.2% similarity
		Emboss Needle	Pairwise global alignment	16.1% identity; 29.1% similarity
		Clustal Omega	Multiple sequence alignment	23.8% identity

Pairwise local alignments identified some conservation of residues in the HsaPolQ interaction interface for tetramerisation with residues in HsaHelQ. Accurate positioning of conserved residues is not critical in this case as protein-protein interactions are often mediated by positive/ negative/ hydrophobic patches as opposed to individual residues. The positions of the conserved residues identified from comparison with HsaPolQ were next mapped on to the HsaHelQ Phyre 2 model, revealing potential HsaHelQ interaction interfaces primarily located on a solvent exposed surface of the helicase ratchet domain 4, similarly to in HsaPolQHD, and located on domain 3, the WHD. Comparison of HsaHelQ and HsaPolQHD structures with the interaction interfaces highlighted can be seen in Figure 4.8 and Figure 4.9 where the proposed interaction interfaces are located on domain 4, the helicase ratchet. This suggests that a possible interaction between HsaHelQ monomers may occur between these interfaces in a similar way observed by HsaPolQHD. Analysis of these interfaces, Figure 4.10, are shown to be a mix of negatively and positively charged, hydrophobic and polar residues that don't seem to form uniform patches for monomer interactions.

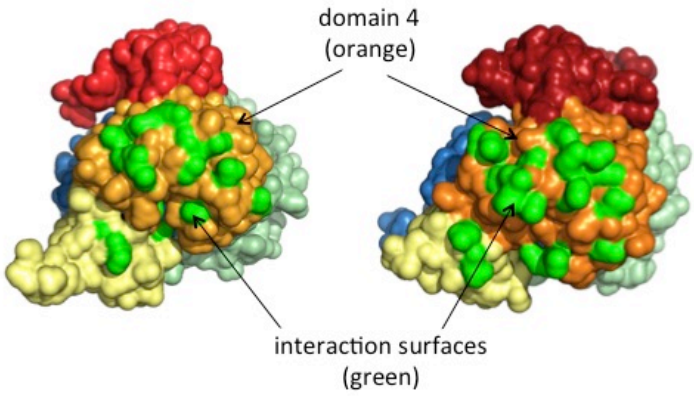


Figure 4.8: Oligomer interaction interfaces of HsaPolQHD and HsaHelQ are primarily located on domain 4 (shown in orange), the helicase ratchet.

A structural model for HsaHelQ (A), and the atomic resolution structure of HsaPolQHD (B) are shown, with domains coloured as follows: RecA-like domain 1, blue; RecA-like domain 2, tea-green ; WHD domain 3, yellow; helicase ratchet domain 4, orange; accessory domain 5, red. Predicted interaction interfaces that were obtained from the structure of HsaPolQHD and Emboss Water local alignments between HsaPolQHD and HsaHelQ are highlighted in lime green. HsaHelQ model was generated using Phyre 2 from FASTA amino acid sequence obtained from UniProt accession number Q8TDG4. For clarity, amino acids 1-300 are omitted from HsaHelQ as these could not be modeled by Phyre 2 and were shown as a rigid rod. Coordinates for HsaPolQHD were downloaded from PDB using code 5AGA. Structures and models were rendered using MacPymol.

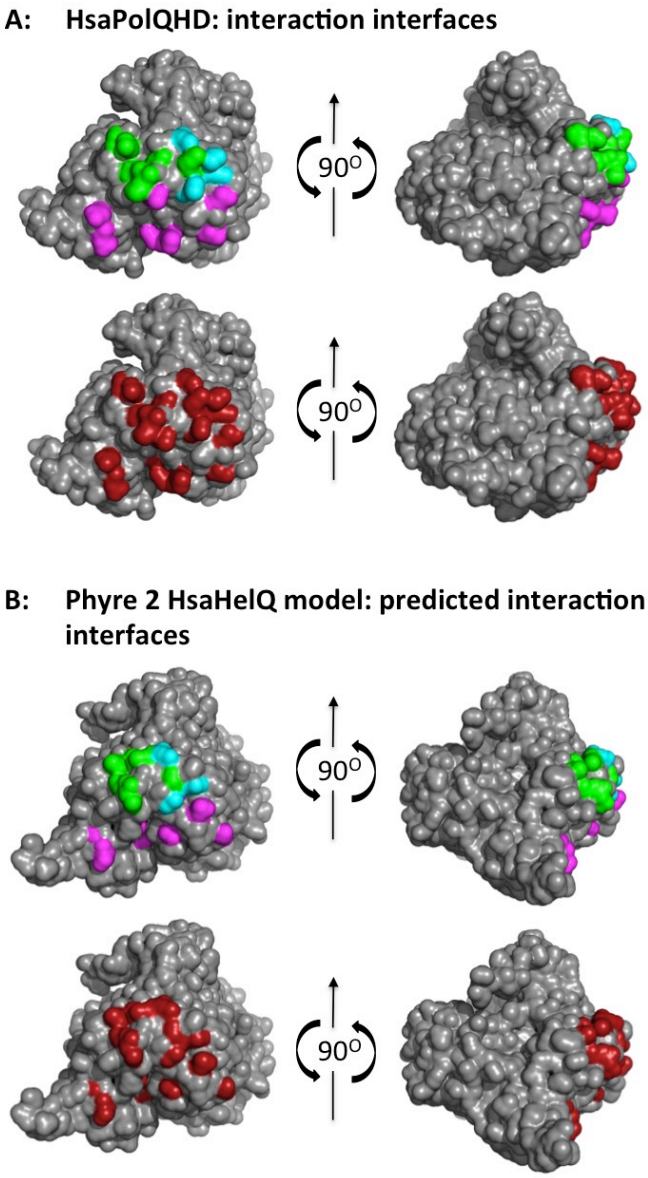


Figure 4.9: Oligomer interaction interfaces of HsaPolQHD and HsaHelQ are found in three clustered patches.

The orientation of HsaHelQ presented above is the same as described in Figure 4.8. Three interaction interfaces were identified for HsaPolQ and these were mapped onto a Phyre 2 HsaHelQ model. These are represented on HsaPolQHD and HsaHelQ monomers highlighted in green, cyan and magenta shown in (A) and (B) respectively, and in red to represent the entire interaction interfaces. Interaction interfaces cluster to one side of the protein that is represented in a front orientation and a 90^{O} rotation on a vertical axis.

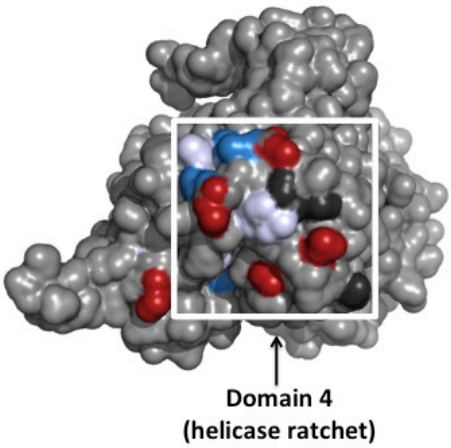


Figure 4.10: Proposed interaction surfaces in modeled HsaHelQ. Please also refer to Figure 4.8 for the relative domain orientations. The position of domain 4, which contains the major proposed interaction interfaces is highlighted. Proposed interaction surfaces from negatively charged amino acids (acidic) are in red, positively charged (basic) amino acids in blue, hydrophobic residues in white and polar residues in black. The model was rendered using MacPymol.

4.5.4. Modeling HsaHelQ dimers

Having identified a potential basis for oligomerisation of HsaHelQ, through comparison to the known oligomerisation interfaces of HsaPolQHD, modeling of HsaHelQ oligomers was attempted using COTH (CO-THreader)[220], which generates structural models of proteins for assembly into a dimer as described in section 2.26.4. This was used as a starting point for analyzing potential higher order oligomerisation regions in HsaHelQ. COTH has two limitations; it can only generate dimeric oligomers and can only accept FASTA sequences of 1,000 amino acids or less. HsaHelQ consists of 1,101 amino acids and so two HsaHelQ sequences were each used on separate analyses where (a) 101 amino

acids were removed from the N-terminal tail of HsaHelQ and (b) 101 amino acids were removed from the C-terminal region of HsaHelQ. This is rational adjustment because existing structural data from HsaPolQHD indicates the interaction interfaces are not predicted to be in either of these removed regions. COTH generated two different dimers using these inputs, Figure 4.11. The COTH generated HsaHelQ dimer for the HsaHelQ N-terminal truncation, Figure 4.11A, showed that the predicted interaction interfaces were very similar to those within the HsaPolQHD tetramer, *via* interactions between domain four, the helicase ratchet. The dimer generated for HsaHelQ truncated at the C-terminus is very different, Figure 4.11B, and does not dimerise *via* the helicase ratchet. It is likely that the C-terminal truncation required for the sequence adjustment has altered the alignments obtained during the modeling process so significantly that the predicted structure has produced such a different model B in comparison to that shown in Figure 4.11A. For example, the top homology modeling hit providing for the basis of the structure shown in Figure 4.11B is the atomic structure of the Rice dwarf virus (PDB: 1UF2), rather than a Ski2 helicase or PolQ. Based on the atomic resolution data for HsaPolQHD it is likely that the dimer shown in Figure 4.11A is most reliable.

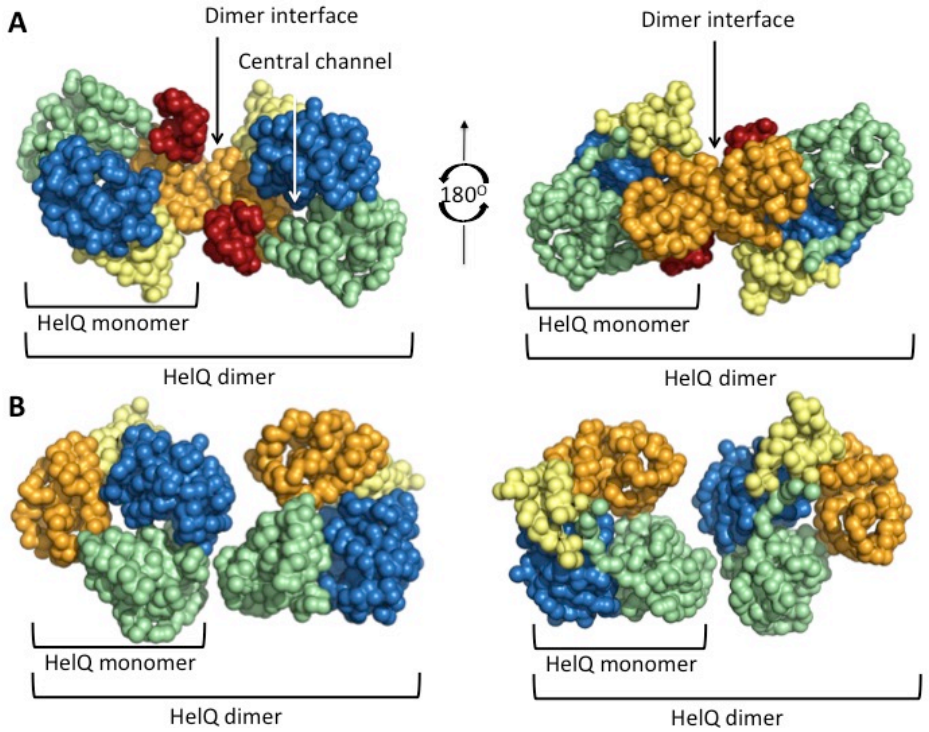


Figure 4.11: COTH generated HsaHelQ dimer models

Domains are coloured as follows: domain 1 (RecA like): blue; domain 2 (RecA like): green; domain 3 (WHD): yellow; domain 4 (helicase ratchet): orange; domain 5 (molecular brake): red. Dimerisation occurs *via* domain 4, the helicase ratchet.

(A) HsaHelQ primary sequence for residues 101-1,101, obtained from UniProt accession number Q8TDG4, was entered into COTH.

(B) HsaHelQ primary sequence for residues 1-1,000, obtained from UniProt accession number Q8TDG4, was entered into COTH.

Models were rendered using MacPymol. The similarity between model A and the known oligomerization of HsaPolQHD, and inconsistencies in the sequence alignment generated from HsaHelQ and HsaPolQHD to generate model B, suggest that model A is most likely.

4.5.5. Modeling HsaHelQ tetramers

COTH established potential dimerization interaction regions of HsaHelQ and this analysis was extended using Galaxy Gemini[221] to generate higher order HsaHelQ oligomers that may be consistent with BN-PAGE data. As for COTH, Galaxy Gemini is limited to accept FASTA sequences of 1,000 amino acids or less. Based on the outcomes from COTH, HsaHelQ primary sequence residues 101-1,101 were entered for analysis by Galaxy Gemini as described in section 2.26.4. This produced a tetrameric organization for HsaHelQ, Figures 4.12A, 4.12B that was strikingly similar to HsaPolQ in Figure 4.7. In agreement with atomic resolution data for HsaPolQHD and a COTH generated dimeric HsaHelQ model, HsaHelQ is predicted to oligomerize *via* interaction interfaces located on domain 4, the helicase ratchet *via* conserved interaction interfaces, Figures 4.12B and 4.12C.

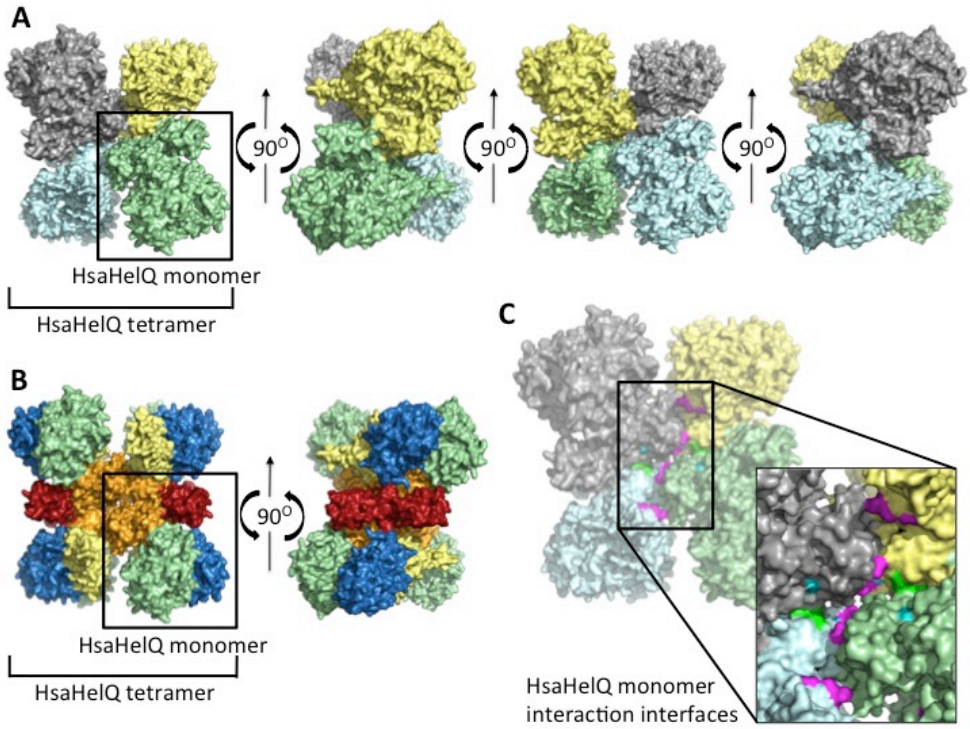


Figure 4.12: Galaxy Gemini generated HsaHelQ tetramer model.

HsaHelQ tetramer model generated using Galaxy Gemini using HsaHelQ primary sequence residues 101-1101 obtained from UniProt accession number Q8TDG4. Models were rendered using MacPymol.

(A) HsaHelQ tetramer: individual monomers are highlighted in different colours. The HsaHelQ tetramer is presented in four orientations, each rotated 90^{O} on a vertical axis.

(B) HsaHelQ tetramer: each domain on each HsaHelQ monomer is highlighted as follows: domain 1, blue; domain 2, green; domain 3, yellow; domain 4, orange; domain 5, red. A HsaHelQ monomer is highlighted in a box. HsaHelQ monomers appear to oligomerise *via* domain 4, the helicase ratchet (orange) to for a tetramer.

(C) HsaHelQ predicted interaction interfaces: Individual HsaHelQ monomers are highlighted in different colours. The three different predicted HsaHelQ monomer interaction interfaces, identified by conservation between HsaHelQ and HsaPolQHD, are highlighted magenta, green and cyan on each HsaHelQ monomer. These interfaces cluster at the interaction interfaces and each mostly contacts a different HsaHelQ monomer.

4.5.6. Cysteine residues and HsaHelQ oligomerisation

BN-PAGE, and to some extent AGF, showed that HsaHelQ has multimeric quaternary structure that although sensitive to addition of DTT, remains stable as a protein consistent with a tetramer. DTT would limit the ability of cysteine residues to form disulphide bridges, and therefore the apparent tetramers in the presence of DTT suggests disulphide bridges are unlikely to have any role in this oligomerization. It is also notable that there are no cysteine residues present in the conserved regions of HsaPolQHD and HsaHelQ that form the most likely surfaces for oligomerization. However, it is plausible that internal disulphide bridges within HsaHelQ may hold the protein in conformational shape optimal for HsaHelQ multimerisation, and on addition of reducing agents, HsaHelQ conformation becomes altered and forms a different oligomeric state. An explanation for high cysteine content in Rad51 paralogs has been proposed considering their relatively small size, but there are no atomic resolution structures available for the Rad51 paralogues, possibly due to the difficulties encountered when working with these proteins (see Chapter 3). Molecular models of human XRCC3 have been generated based on structures of yeast Rad51. In these, the majority of cysteine residues were found to be solvent exposed on HsaXRCC3 models. The authors suggest that these cysteines become reduced during reducing conditions in the cell, thereby activating HsaXRCC3 to participate in DNA protection and/ or DNA repair activities in order to prevent DNA damage induced by such reducing conditions[233].

Therefore these cysteines in HsaXRCC3 are proposed to function in signaling for DNA repair.

To investigate the nature of HsaHelQ cysteines the Phyre 2 HsaHelQ model was used to highlight positions of each of the 24 cysteine residues, shown in Figure 3.13. Very few cysteines appear to be solvent exposed on the HsaHelQ model, with the majority located internally. However, the possibility of more solvent exposed cysteines cannot be completely discounted because this is only a model of HsaHelQ.

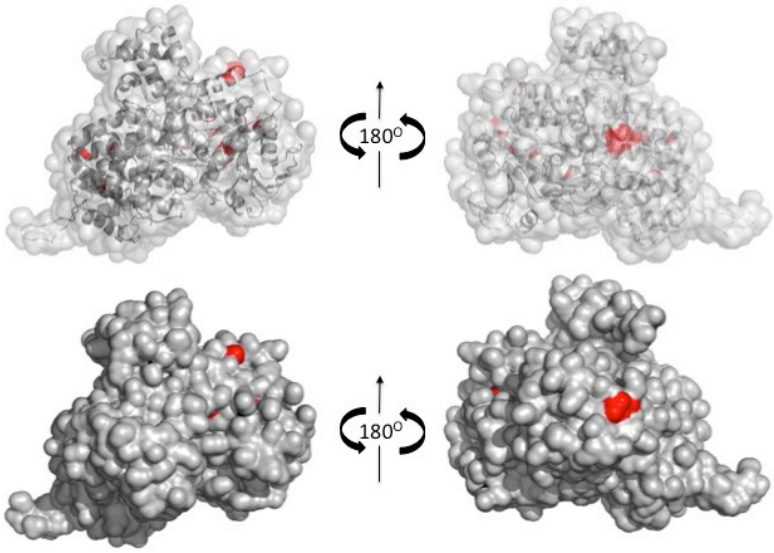


Figure 4.13: HsaHelQ cysteine residues are predicted to be mostly internal.

Phyre 2 HsaHelQ model with cysteine residues highlighted in red. Internal cysteine residues can be seen in the semi-transparent figures. Very few cysteines are solvent exposed on the solid surface models. Each model structure is shown in two orientations rotated 180° on a vertical axis.

Based on the presumption that the positioning of these cysteines in a model of HsaHelQ is accurate or similar to their positions in the native enzyme, it may be possible to speculate about why addition of DTT alters the oligomeric state of HelQ and stimulates helicase activity (later in this chapter).

4.6. The HsaHelQ N-terminal tail

Sequence alignments of HsaHelQ versus archaeal Hel308, highlights that a 300 amino acid N-terminal region of HelQ that precedes the RecA-like motor domains has no homology to Hel308 or to any other protein. HelQ proteins in *Adineta vaga* and some mammals also show this N-terminal extension although with very limited sequence conservation. During HsaHelQ modeling using Phyre 2, this 300 amino acid N-terminal region could not be modeled, due to the lack of sequence similarity with other proteins of known structure. This interesting feature of HsaHelQ was pursued further using online structural prediction analysis to obtain any information about why this N-terminal extension of HsaHelQ has been conserved.

IUPred, a predictor of protein intrinsic disorder[234] was used to determine whether the N-terminal region of HsaHelQ is likely to adopt a stable structure or not. IUPred considers the ability of amino acid residues on a polypeptide chain to form stable inter-residue contacts to maintain a stable structure, and therefore is an indicator of which regions of a protein may adopt energetically stable conformation or not. The HsaHelQ

amino acid sequence obtained from UniProt accession number Q8TDG4 was entered into IUPred which revealed that HsaHelQ amino acids 1-300, corresponding to the HelQ N-terminal tail, showed high tendency towards disorder in addition to a small region at the extreme C-terminal, Figure 4.14. This prediction is very similar to that observed for other DNA helicases such as yeast Sgs1 and its human homologue BLM which have disordered N-terminal regions of up to 500 amino acids[235].

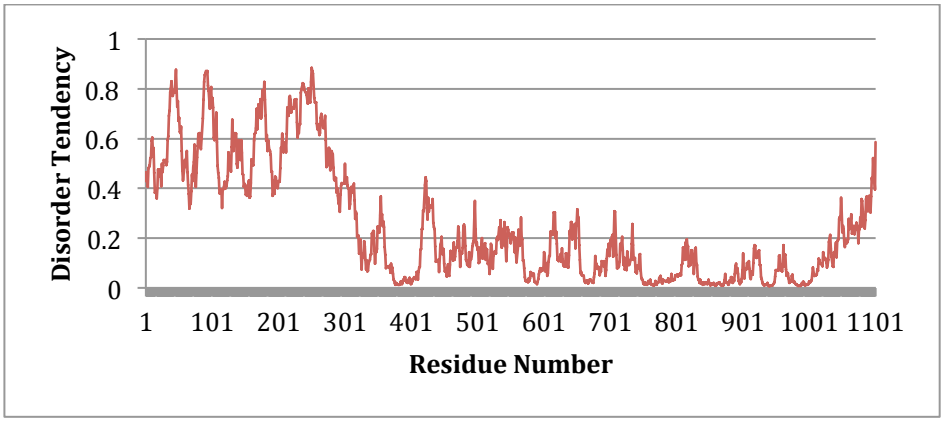


Figure 4.14: HsaHelQ N-terminal region is predicted to be intrinsically disordered:
IUPred [234] predicts HsaHelQ residues 1-300 and residues at the extreme C-terminus to have high tendency towards disorder.

Phyre 2 analysis showed the N-terminal tail of HsaHelQ to be unstructured with an inability to predict any secondary structure. However, molecular modeling using I-TASSER predicts this region to possess some elements of secondary structure, described in section 2.26.3. Here, the N-terminal tail is predicted to be largely solvent exposed, consistent with regions of disorder. However, some parts of this region are predicted to adopt small regions of secondary structure, although

these predictions have been generated with fairly low confidence. Models generated for the N-terminal tail of HsaHelQ by I-TASSER are shown in Figure 4.15.

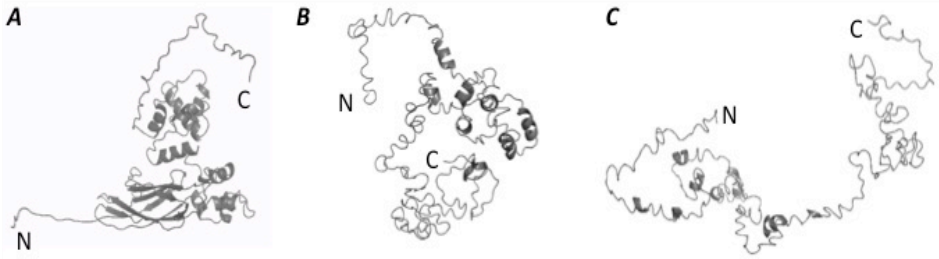


Figure 4.15: Structural prediction of HelQ N-terminal tail:
Predictions generated using I-TASSER with HsaHelQ amino acid sequence shows regions of random coil and small elements of secondary structure in the HsaHelQ N-terminal tail.

Substantial unstructured N-terminal tails have been described for other DNA helicases including DNA repair helicases of the RecQ family, Sgs1, BLM and WRN[235]. It has been proposed that these protein regions may have emerged during evolution of complexity in eukaryotes, especially metazoans. These tails appear to serve specific functions in the helicases listed above, and mediate protein-protein interactions that modulate helicase activity and may also be involved in helicase multimerisation[236]. Regions of disorder are often involved in protein-protein interactions, and upon binding will adopt elements of secondary structure. This is true for Sgs1 anti-recombinase in yeast, in which the disordered N-terminal tail forms transient α-helices essential for interactions with other proteins that modulate protein activity[235]. Predicted disorder and possible small elements of secondary structure in

the HsaHelQ N-terminal tail bear resemblance to those described in Sgs1. As previously mentioned, BLM helicase of the RecQ family, also possesses a long N-terminal tail. The oligomeric state of BLM helicase is not confirmed, however, there is one study proposing that proposed helical bundles in the BLM N-terminal tail may be involved in BLM oligomerisation[236]. My analysis of HsaHelQ presented in previous sections suggests that this is less likely to be true of HsaHelQ. However, it is possible to speculate, based on the precedents of other DNA repair helicases, that the N-terminal region of HsaHelQ is essential for interactions with Rad51 paralogs which modulate HsaHelQ function and *vice versa*.

Initial plans involved expression and purification of the isolated HsaHelQ N-terminal region for structural analysis by circular dichroism (CD), to determine experimentally some more information about this region of HsaHelQ. However, as described in section 3.7, this could not be achieved due to problems with protein stability.

4.7. Analysis of DNA binding by HelQ using EMSAs

Previous analyses of DNA binding and unwinding by an archaeal Hel308 from *Methanothermobacter thermautotrophicus* (MthHel308) showed a preference for targeting ssDNA and branched structures. *In vitro* biochemistry was used to investigate DNA binding and unwinding activity of purified His-SUMO-HsaHelQ, as described in section 2.22, in

comparison with existing information available for archaeal and metazoan Hel308 enzymes. Details of DNA substrates used are found in Table 2.9.

Most EMSA-based analyses of His-SUMO-HsaHelQ binding to model substrates were carried out prior to recognizing the importance of adding exogenous DTT to stimulate His-SUMO-HsaHelQ activity. Most EMSAs resulted in extensive smearing in lanes containing high concentrations of His-SUMO-HsaHelQ, summarized in Figure 4.16, indicative of highly unstable protein-DNA interaction. In addition, it was common for His-SUMO-HsaHelQ to provoke aggregation of DNA substrates into the wells of the gels. This was encountered for all DNA substrates that His-SUMO-HsaHelQ was titrated against, and was also the case when exogenous DTT was added into EMSA reactions. However, despite these caveats, we did observe with some substrates the presence of two defined protein-DNA complexes. This is consistent with stable binding of His-SUMO-HsaHelQ to DNA into complexes containing two different protein sizes, which is also consistent with two protein species observed for His-SUMO-HsaHelQ during native PAGE, in the absence of DTT. In the presence of DTT, HelQ binding to substrates was enhanced evidenced by HelQ binding to substrate at lower concentrations, although non-specific or unstable binding continued to be observed. However, both HelQ-DNA complexes remained in the absence and presence of exogenous DTT. It may be that the differing oligomeric states of HelQ observed in either the presence or absence of DTT suggest that the active helicase form of HelQ is a tetramer but that this form is not required for

DNA binding in EMSAs, which could occur by HelQ in monomeric or any other oligomeric form.

Quantification of these EMSAs was inconsistent because of the extensive smearing and aggregation in the wells of shifted DNA mixed with His-SUMO-HsaHelQ. This was reinforced by the fact there was persistent variation between the replicate gels for the same substrate. The observation of small amounts of stable and distinct protein-DNA complexes when His-SUMO-HsaHelQ was titrated into substrates allowed some quantification of DNA binding by HsaHelQ, albeit it limited using this method (Figure 4.16). For example, there appears to be very little unbound DNA substrate for fork 2 after 16 nM HelQ as from lane 5-9 Figure 4.16A in comparison to a subtle change in shifting of fork 5 by HelQ after 32 nM HelQ as from lanes 33-36, Figure 4.16A. Following attempts at quantification, unexpectedly, Fork 5, the fully base paired model replication fork is bound better than Fork 9, that is the same as Fork 5 but has ssDNA ends at each of the flayed arms.

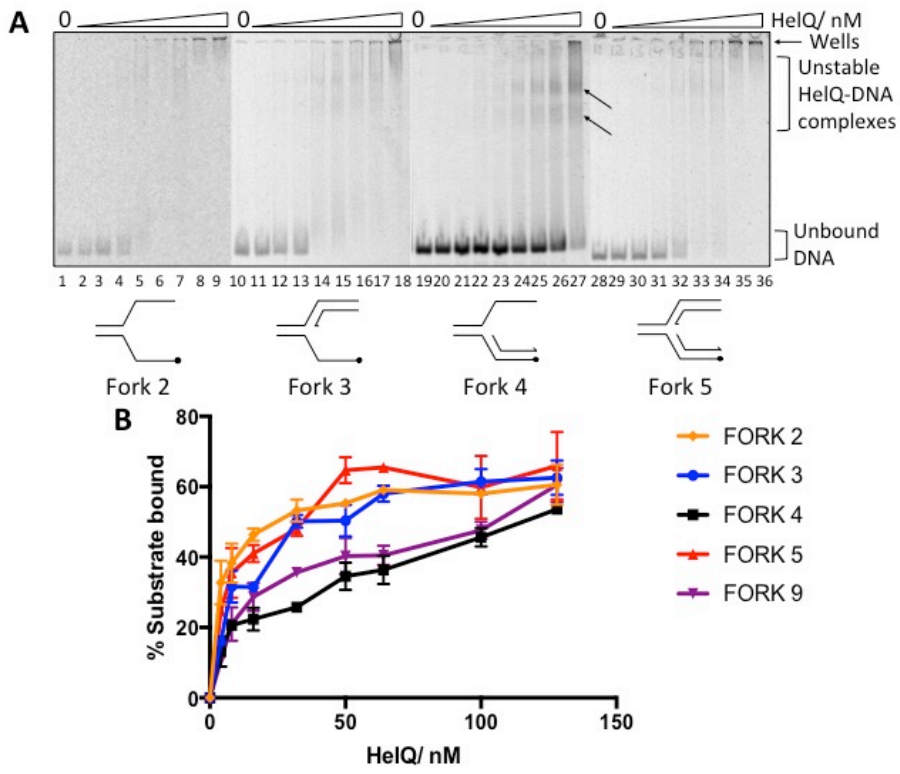


Figure 4.16: His-SUMO-HsaHelQ EMSAs. Reactions were at RT for 10 minutes and contained 1x HB, and 25 nM DNA. HsaHelQ was titrated at 0, 4, 8, 16, 32, 50, 64, 100 and 128 nM against each substrate. Protein-DNA complexes are highlighted. Filled dots indicate 5' Cy5 labeled ends.

4.8. Analysis of HsaHelQ helicase unwinding

Helicase unwinding by HsaHelQ was determined in assays measuring dissociation of annealed DNA strands, detectable by one strand being radio- or fluorescently labeled, as described in section 2.23. A "trap" ssDNA strand was also present in assay reaction mixtures, of the same sequence as the labeled strand within the substrate. The function of the trap is to base pair with unlabeled dissociated strand product of helicase activity, to prevent it re-annealing to re-form substrate. Details of DNA substrates used are found in Table 2.9. Unwinding reactions were loaded

onto TBE gels for electrophoresis to separate DNA strands by molecular weight. A boiled substrate sample was loaded as a control, to locate the position of the denatured labeled oligonucleotide that is used as a reference point of fully dissociated labeled ssDNA. This is useful, as often unwinding of more complex DNA substrates results in more than one product.

HsaHelQ, similarly to MthHel308, is an ATP-dependent 3' to 5' DNA translocase that possesses helicase activity, as reported by Marini *et al.*[4]. In light of the apparent differences in oligomeric state of HsaHelQ in the absence and presence of reducing agent DTT, we sought to test effects of DTT on His-SUMO-HsaHelQ unwinding activity of forked DNA, a preferred substrate for archaeal Hel308, compared to a fully base-paired linear duplex DNA which is not effectively unwound by Hel308. In the absence of DTT His-SUMO-HsaHelQ fork unwinding was weak, whereas in the presence of DTT fork unwinding by His-SUMO-HsaHelQ is greatly enhanced, summarised in Figure 4.17. Unwinding of fork 4, a model replication fork with a lagging strand only/ leading strand gap produces two products, shown alongside fork 2 in Figure 4.17. Unwinding of fork 4 by His-SUMO-HsaHelQ is consistent with it being a 3' to 5' DNA translocase that first unwinds the fork to a partial duplex with a 3'-OH ssDNA tail before further acting on that as a substrate for conversion to ssDNA. Figure 4.18 summarises this activity, and shows that His-SUMO-HsaHelQ can also unwind fork 3 that has a leading strand, but cannot unwind a fully base paired duplex. In all cases the addition of DTT to

reactions greatly enhanced His-SUMO-HsaHelQ unwinding activity. This helicase activity of His-SUMO-HsaHelQ on forks compared to duplex DNA is in agreement with previous data for MthHel308.

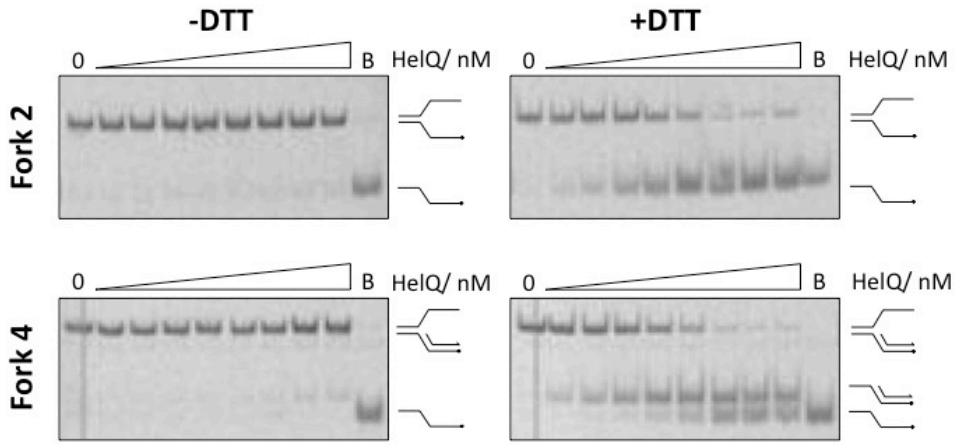


Figure 4.17: HsaHelQ unwinding activity is enhanced in the presence of DTT.

His-SUMO-HsaHelQ unwinding activity on fork 2 (flayed duplex) and fork 4 (fork with a lagging strand only). Unwinding reactions were assembled to contain 1x HB buffer, 5 mM ATP, 5 mM $MgCl_2$, 25 nM DNA as indicated, 0, 4, 8, 16, 32, 50, 64, 100 and 128 nM HsaHelQ and 50 mM DTT as indicated, and were incubated at 37°C. Reaction products were separated by electrophoresis on 10% acrylamide TBE gels at 160 V for one hour. HsaHelQ unwinding activity is similar for both substrates, and begins to plateau after approximately 50% unwinding.

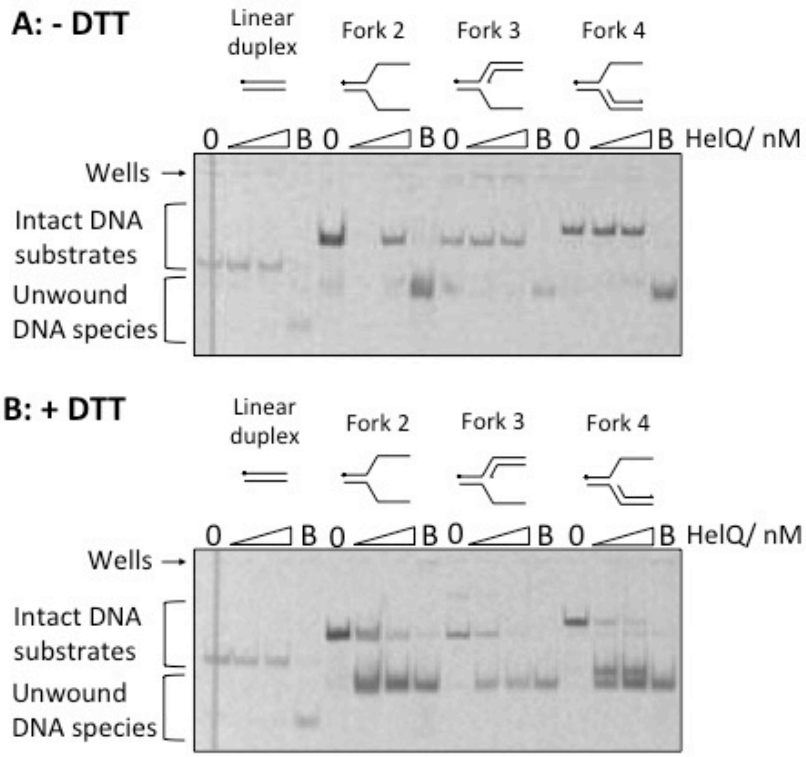


Figure 4.18: HsaHelQ unwinding activity is enhanced in the presence of DTT:

HsaHelQ unwinding activity on linear duplex, fork 2 (flayed duplex), fork 3 (fork with a leading strand only) and fork 4 (fork with a lagging strand only). Unwinding reactions were assembled to contain 1x HB buffer, 5 mM ATP, 5 mM $MgCl_2$, 25 nM DNA as indicated 10, 32, 64, 128 nM HsaHelQ and 50 mM DTT as indicated, and were incubated at 37°C. Reaction products were separated by electrophoresis on 10% acrylamide TBE gels at 160 V for one hour. HsaHelQ unwinding activity is efficient for branched substrates rather than linear duplex.

His-SUMO-HsaHelQ unwinding of fork 2 and fork 4, when measured as a function of time in the presence of DTT, showed no significant difference in unwinding efficiency (Figure 4.19), with about 50% of substrate being detectably unwound in these assays after ten minutes.

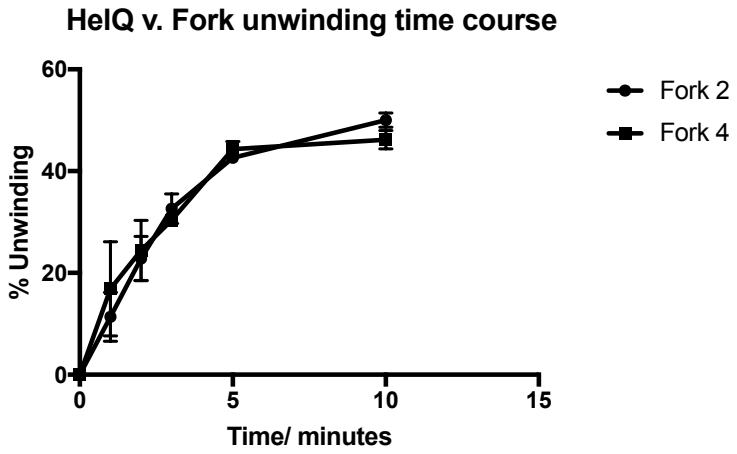


Figure 4.19: HsaHelQ unwinding activity on forked structures:
His-SUMO-HsaHelQ unwinding activity as a function of time on fork 2 (flayed duplex) and fork 4 (fork with a lagging strand only). Unwinding reactions were assembled in 200 µl to contain 1x HB buffer, 5 mM ATP, 5 mM $MgCl_2$, 25 nM DNA as indicated, 128 nM His-SUMO-HsaHelQ and 50 mM DTT, and were incubated at 37°C. 20 µl was removed from the reaction pool at indicated time intervals and stopped by addition of $1/5^{th}$ stop buffer. Reaction products were separated by electrophoresis on 10% acrylamide TBE gels at 160 V for one hour. HsaHelQ unwinding activity is similar for both substrates, and begins to plateau after approximately 50% unwinding. For each substrate the assay was done three times and error bars represent the standard deviation from the mean of each data point, calculated using GraphPad Prism software.

We also tested the ability of the reducing agent β-mercaptoethanol (BME) to stimulate unwinding by His-SUMO-HsaHelQ in comparison to DTT. This is because biophysical experiments initiated with HsaHelQ (see sections below) required use of the fluorescently labeled DNA substrates that are photo-bleached by DTT that may be problematic to experimental design of these assays, but BME does not cause photo-

bleaching. As observed in Figure 4.2, enhanced His-SUMO-HsaHelQ unwinding is observed on addition of BME, similarly as for DTT.

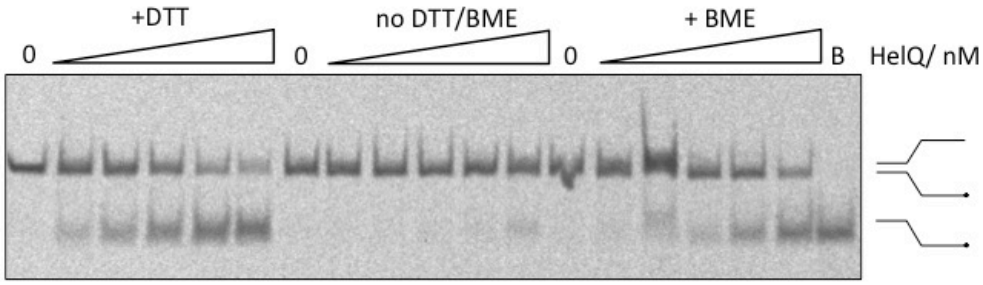


Figure 4.20: HsaHelQ unwinding activity is also stimulated in the presence of β-mercaptoethanol

His-SUMO-HsaHelQ unwinding activity on fork 2 (flayed duplex) in the presence or absence of reducing agents DTT or BME (50 mM in each case). Unwinding reactions (20 µl) contained 5 mM ATP, 5 mM $MgCl_2$, 25 nM DNA in 1x HB buffer and HsaHelQ was added to 5, 10, 25, 50 or 100 nM as indicated. Reactions were incubated at 37°C for 20 minutes before deproteinisation. Reaction products were separated by electrophoresis on 10% acrylamide TBE gels at 160 V for one hour. HsaHelQ was active as a helicase in both DTT and BME, compared with very weak or undectable activity in the absence of reductant. Lanes are marked for no protein controls (0) and a fully dissociated boiled fork (B).

4.9. Analysis of DNA binding by HsaHelQ using NanoTemper Microscale Thermophoresis.

After attempting gel based EMSAs to investigate His-SUMO-HsaHelQ-DNA binding, a biophysical approach was next used to try to refine the analysis and to obtain kinetic data for His-SUMO-HsaHelQ-DNA binding. The aim was to assess whether His-SUMO-HsaHelQ was

amenable to microscale thermophoresis (MST) measurements, and if so did it yield reproducible kinetic binding data. MST is sensitive to the size, hydration and charge of molecules defined by:

$$S_T = A/kT\,(-\Delta s_{hyd}(T)+(\beta\sigma_{eff}^2/4\varepsilon\varepsilon_0 T)x\lambda_{DH})$$

Where A = size, $-\Delta s_{hyd}(T)$ = hydration shell and σ_{eff}^2 = charge[222]. This technique detects fluorescent molecules, in this case Cy5 labeled ligand DNA, at a constant concentration against a titration of unlabeled molecule, in this case His-SUMO-HsaHelQ. Details of fluorescently labeled DNA substrates used are found in Table 2.9. MST is a solution-based technique, so proteins can be studied in their native state, and has the advantage of requiring only small quantities of protein, usually in the low nanomolar range. The principle of MST[222] is based on infrared (IR) heating of a small local area of the sample tube that causes molecules to diffuse away from the localized area of heat at a measurable rate. The localized region of heat is switched off and the molecules measurably diffuse back. Whether or not a protein is bound to a ligand influences its rate of diffusion, which is used as the basis to determine if binding has occurred. The output from this method is illustrated in Figure 4.22, which is an annotated MST trace for His-SUMO-HsaHelQ binding to fork 2. MST traces obtained for each concentration of unlabeled ligand are plotted as sigmoidal binding curves from which kinetic data (e.g. K_d; dissociation constants) can be derived. A total of 16 samples can be measured in a

single analysis, giving measurements over a broad range of protein or ligand concentrations. This method is described in section 2.27.

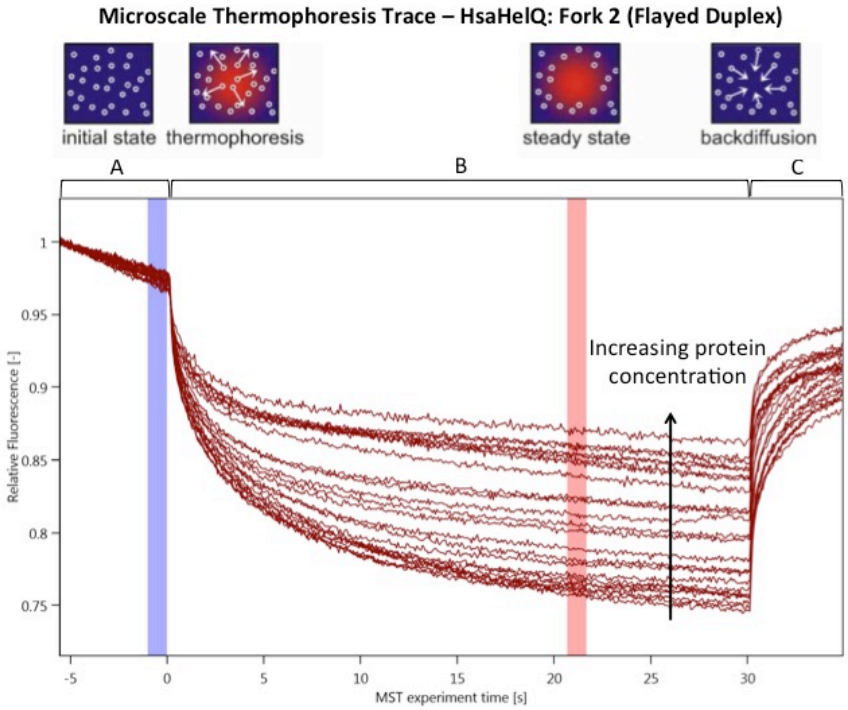


Figure 4.21: Microscale thermophoresis trace for His-SUMO-HsaHelQ titratrion against fork 2 (flayed duplex).

His-SUMO-HsaHelQ was titrated from 7.6 pM to 250 nM against 1 nM Cy5 labeled fork 2. This trace is typical of most traces obtained for His-SUMO-HsaHelQ DNA binding assays using MST. For detailed analysis of the curves refer to the main text.

(A) Initial fluorescence is measured, and initial diffusion rate is measured at the highlighted blue region (F_{COLD}).

(B) IR heating of a localized region commences and molecules diffuse away from this region *via* thermophoresis. This phenomenon reaches a steady state at which diffusion rate is measured (F_{HOT}).

(C) IR heat is switched off and molecules diffuse back. As fluorescently labeled DNA molecules are bound by protein, they diffuse more slowly indicated by the shift of the MST traces with increasing protein concentration denoted by the arrow.

Before MST measurements are commenced, a quality scan of each of 16 samples is used to assess quality of the reagents being used, particularly in protein-DNA binding assays to identify problems with protein samples. This is detected by a fluorescence scan in which an ideal readout is a smooth single peak. Peaks with 'shoulders' are indicative of protein aggregation or proteins adhering to the sample tube. Figure 4.22 shows a typical peak obtained for HelQ-DNA binding assay quality scans.

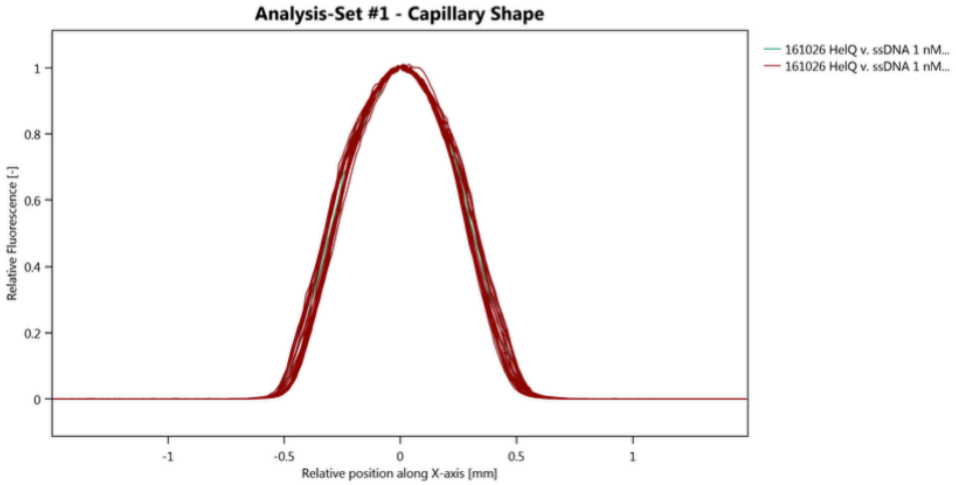


Figure 4.22: HsaHelQ - ssDNA binding assay capillary scan:
Capillary scans are carried out prior to MST measurements. This trace shows capillary scan data of all 16 capillaries each at 20% and 40% LED power, and is typical of all scans taken for all readings for each HsaHelQ DNA binding assay. The smooth peak indicates a good quality sample.

After satisfactory outcome from testing the quality of reagents (Figure 4.22) His-SUMO-HsaHelQ-DNA binding data was obtained using the same fluorescently labeled DNA substrates as in the EMSAs, giving curves such as shown in Figure 4.21. The decline of fluorescence in section

A shown in Figure 4.21 is indicative of photobleaching of the Cy5 fluorescent dye. The rapid decrease in fluorescence that occurs at the start of section B represents the fluorescent molecules diffusing away from the localized area of heat induced by IR. This eventually reaches a steady state where the curve flattens at the end of section B. At this point, the heat is switched off and the increase in fluorescence represents back diffusion of the fluorescent molecules. Fluorescence measurements are measured before the heat is applied (F_{COLD}), shown as a blue shaded panel on the MST trace, and again once the heat is applied (F_{HOT}), shown as the red shaded panel on the MST trace. The difference between the two is calculated to give ΔF [222] by:

$$\Delta F = F_{COLD} - F_{HOT}$$

This is used to plot the fraction of labeled molecules bound, and good data should generate a sigmoidal binding curve from which kinetic data can be derived in which there is a baseline, that represents minimal binding, an increase occurring due to protein binding and finally a plateau in which binding saturation occurs. The fraction of molecules bound is equivalent to ΔF, as described earlier. For effective kinetic analysis the concentrations used of unlabeled molecules should be within the expected k_d range. However, this was unknown for His-SUMO-HsaHelQ, and therefore His-SUMO-HsaHelQ concentration ranges were utilized until sigmoidal curves were obtained (Figure 4.24). This established that titration of His-SUMO-HsaHelQ in the range 7.6 pM to 250 nM was optimal to produce a suitable curve from which kinetic data could be

derived. Each His-SUMO-HsaHelQ DNA binding assay, to various DNA substrates, was subsequently carried out in triplicate.

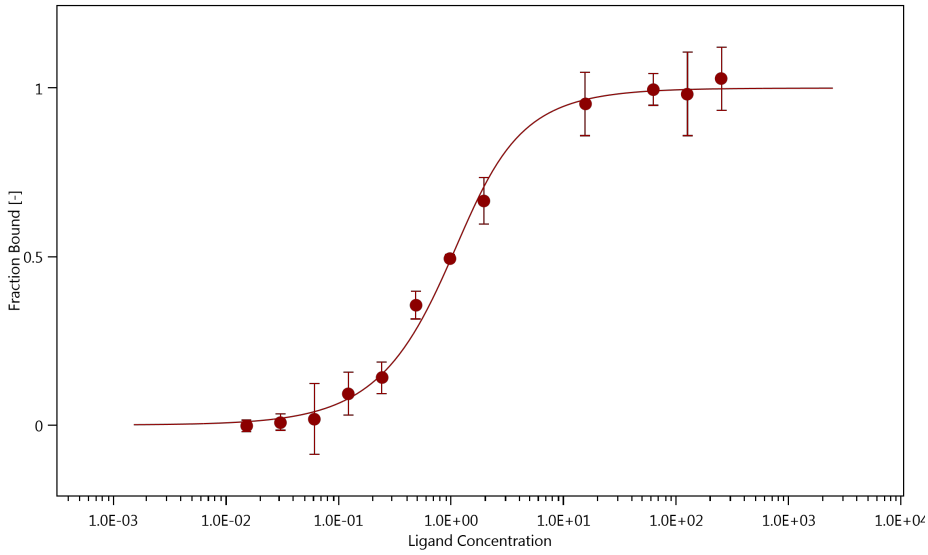


Figure 4.23: A sigmoidal curve for HsaHelQ – fork 2 (flayed duplex) binding:

Fraction of Cy5 labeled fork 2 molecules bound by His-SUMO-HsaHelQ generates a sigmoidal binding curve.

Table 4.4 lists DNA binding data for HsaHelQ that was obtained from MST analysis. Statistics including standard error of regression and reduced chi squared values were calculated for each fork; standard error of regression is a measure of the accuracy of the derived values. Reduced chi squared is a calculation of regression analysis and indicated whether a model is a good fit to the data or not. A reduced chi squared value of one indicates a good fit, a value greater than one represents that the data is not a good fit to the model and a value less than one indicated that the model is 'over fitting' the data.

Table 4.4: K_d constants and associated statistics derived from HsaHelQ MST analysis: Data highlighted in green is most statistically relevant, and is detailed further in the main text. MST data shows that His-SUMO-HsaHelQ has much lower k_d values for branched DNA molecules in contrast to linear ssDNA or linear duplex DNA, indicative of higher affinity for such substrates. One set of binding assays, mixing HelQ with fork-4 at MST power 20, failed to register any values for reasons unknown, but which are most likely due to failure of the capillary tubing in this instance.

SUBSTRATE	EXCITATION %	MST POWER	Kd	+/-	Std. error of regression	Reduced chi squared
ssDNA	15.00	20.00	713.84	835.94	4.57	9.01
		40.00	124.99	38.47	7.80	NA
Linear duplex	20.00	20.00	180.15	207.37	11.90	7.07
		40.00	50.38	13.89	7.37	41.54
Fork 2	30.00	20.00	1.41	0.20	1.48	0.92
		40.00	0.52	0.08	3.26	0.55
Fork 3	30.00	20.00	0.52	0.10	2.17	3.60
		40.00	2.47	0.54	3.56	101.13
Fork 4	15.00	20.00	FAILED	FAILED	FAILED	FAILED
		40.00	3.87	4.13	5.13	1.78
Fork 5	30.00	20.00	1.57	0.37	4.73	30.17
		40.00	2.04	0.68	11.41	2.81
Fork 9	30.00	20.00	2.71	0.63	4.88	1.76
		40.00	5.90	3.01	14.02	223.88

K_d dissociation constants displayed in table 4.4 that are significant are highlighted in green. K_d values are a type of equilibrium constant that interrogates the ability of a protein and its ligand to dissociate. From this data, k_d values for linear duplex DNA and linear ssDNA, at 180 nM, are far greater for k_d values for branched DNA substrates that are typically <5 nM. K_d values for all branched DNA substrates are similar, and from this data His-SUMO-HsaHelQ appears to have little preference between substrates. However, this data must be treated with caution because the statistics associated with each measurement set suggest that at least some of the measurements are sub-optimal.

Hill constants can also be derived from the MST binding data, to assess whether a protein binds its ligand, in this case DNA, in a cooperative or non-cooperative manner. This is a common phenomenon in molecular biology, for example, recombinase proteins RecA and Rad51 nucleate on ssDNA which then stimulates binding of subsequent recombinase monomers to assemble into a nucleoprotein filament in a positively cooperative manner to initiate HR. Hill constants of less than one indicate negative cooperativity, where binding of one molecule negatively influences binding of subsequent monomers. A value of one indicates binding of only one monomer, and a value greater than one is indicative of positive cooperativity. Coooperativity data, Table 4.5 where statistically significant data is highlighted in green, shows Hill constants greater than 1 for linear duplex DNA, fork 5 (fully base paired fork) and fork 9 (fully base paired fork wth recessed DNA at the end of the branched arms), suggesting positive cooperativity. This is noticeable in comparison to all other substrates that have values of 1 or less, as these molecules contain substantial areas of duplex DNA. Fork 2 has a Hill coefficient of approximately 1, indicating binding of one monomer and all other substrates have values of less than 1. Although interesting, the poor statistical quality of these data means that conclusions from these results should be treated with caution.

Table 4.5: Hill constants and associated statistics derived from HsaHelQ MST analysis:

Data highlighted in green is most statistically relevant. MST data shows that HsaHelQ has Hill constants greater than 1 for linear duplex, fork 5 and fork 9 all of which contain substantial regions of duplex DNA, indicative of cooperative binding. Fork 2 has a Hill constant of approximately 1 indicative of binding of 1 monomer. All other substrates have Hill constants of one or less indicative of negative co-operativity.

SUBSTRATE	EXCITATION %	MST POWER	Hill coefficient	Std. error of regression	Reduced chi squared
ssDNA	15.00	20.00	1.56	4.48	9.42
		40.00	2.36	3.89	NA
Linear duplex	20.00	20.00	3.53	10.29	6.12
		40.00	1.55	5.23	25.22
Fork 2	30.00	20.00	0.98	1.48	0.83
		40.00	1.03	2.26	0.25
Fork 3	30.00	20.00	0.99	1.84	2.95
		40.00	0.94	3.62	36.20
Fork 4	15.00	20.00	FAILED	FAILED	FAILED
		40.00	2.03	3.26	0.85
Fork 5	30.00	20.00	1.98	2.69	65.10
		40.00	2.55	8.04	1.18
Fork 9	30.00	20.00	1.61	4.21	1.08
		40.00	3.51	7.82	13.24

It is evident that the majority of the kinetic data obtained using NanoTemper Microscale Electrophoresis technology is not statistically significant given the standard error of regression values and reduced chi squared values. Therefore it would not be advisable to use this data in its current form to draw conclusions about HsaHelQ binding to DNA substrates. It seems, based on this first analysis, that this technique is not suitable for studying HsaHelQ DNA interactions. Literature that reports using MST technology is primarily regarding protein-protein interactions, and there is limited data available for protein-DNA interaction studies using this technique. This may suggest an inherent problem in studying protein-DNA interactions using MST.

4.10 TIRF microscopy for analysis of HsaHelQ: DNA reeling?

His-SUMO-HsaHelQ-DNA binding studies using biochemical or biophysical approaches did not prove successful, however biochemical studies of His-SUMO-HsaHelQ helicase activity has revealed, that similarly to archaeal Hel308, His-SUMO-HsaHelQ helicase activity is most efficient at unwinding branched DNA substrates, particularly fork 4, a model fork with a lagging strand only. To build on this, His-SUMO-HsaHelQ helicase activity was investigated further using smFRET and TIRF microscopy, described in sections 2.28. This technique can provide His-SUMO-HsaHelQ kinetic data in real time, such as translocation rate along ssDNA and HsaHelQ step size[237].

smFRET has revealed some DNA helicases such as Srs2 and PcrA demonstrate a reeling activity whilst anchored at a ds/ssDNA junction[107,237]. This is a common feature among anti-recombinases used for stripping DNA and/ or proteins from ssDNA to allow diversion towards other repair pathways. Such an activity is consistent with a hypothesis that HsaHelQ and Hel308 have roles to control recombination. In a typical smFRET experiment for protein catalysed-DNA processing a fluorescently labeled partial duplex DNA molecule with a 3'-OH overhang is immobilised onto a polyethylene glycol (PEG) coated quartz slide *via* a neutravidin-biotin linkage, Figure 4.24.

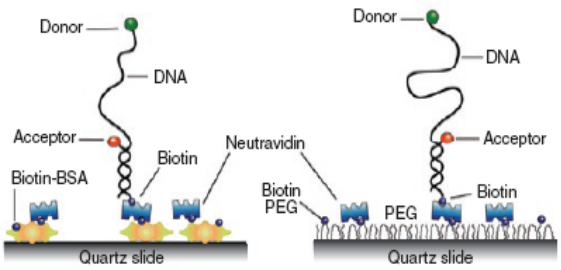


Figure 4.24: Immobilization of DNA molecules

Single partial duplex DNA molecules with 3'-OH overhangs were immobilized on PEG or BSA coated quartz slides *via* biotin-neutravidin linkages. Adapted from Roy *et. al.* 2008[223].

This method of microscopy enables focusing on single molecules and reducing background fluorescence, to enable single molecule FRET (smFRET) to occur, thereby detecting changes in distance between their attached counterparts. An idea of the distance to be measured is required in order to use the correct dyes, which must have overlapping wavelengths. When the two fluorophores come within close proximity of one another, FRET occurs. Here, Cy3, the donor fluorophore was located 3'-OH at the end of the ssDNA overhang, and the Cy5 acceptor was located 5' at the ds/ssDNA junction. Two partial duplex substrates were used, which had either 30 and 40 nucleotides of ssDNA.

Figure 4.25 shows traces displaying fluorescence intensity of Cy5 and Cy3 fluorophores in the absence of HsaHelQ protein. As described in section 2.28, the camera records data in two channels that represent the fluorescence intensity of each fluorophore, here indicated by green and red. This trace represents tha baseline fluorescence intensity and FRET

values. When FRET occurs, this is represented by a reciprocal exchange in fluorescence intensity for each fluorophore. This can be observed as a change in FRET, indicated by an increase from the baseline. When this occurs there are two or more populations of DNA molecules demonstrating different FRET values and this can be observed by a shift in the frequency of FRET values on a histogram. In this case, in the presence of no protein, baseline FRET efficiency is 0.3.

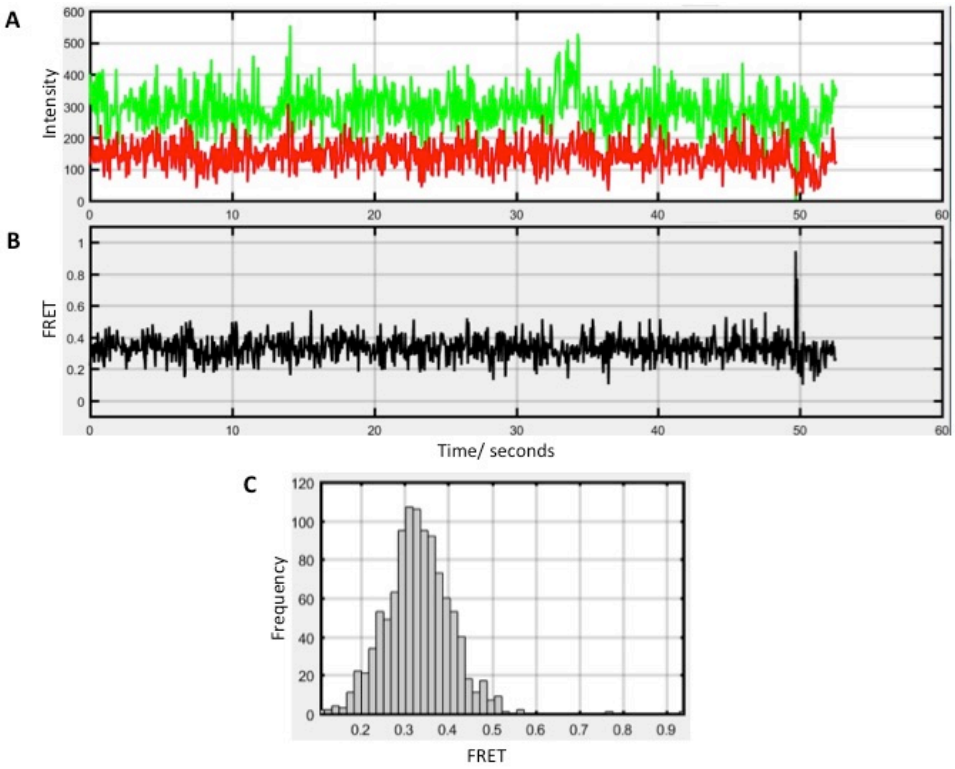


Figure 4.25: smFRET baseline trace

(A) Fluorescence intensity of Cy5 and Cy3 fluorophores.

(B) FRET baseline in the absence of protein is 0.3.

(C) The majority of DNA molecules are displaying FRET of approximately 0.3.

Poor results were initially obtained, and this was thought to be due to protein sticking to the slides hence explaining why no FRET was observed. Therefore bovine serum albumin (BSA) coated quartz slides were used instead of PEG coated slides. This is one of the reasons why fluorescently labeling the protein is advantageous as we can observe where it is and how it is behaving. However, this was not possible for HsaHelQ during this project, hence why labeled DNA was used. Unfortunately no improvements were observed and despite testing protein concentrations within picomolar to micromolar ranges, and altering buffer conditions, no FRET was observed mediated by His-SUMO-HsaHelQ. This is indicative of sub-optimal reaction conditions that must be optimized in future.

Reeling would be indicated on a FRET trace by a steady increase in FRET from the baseline value, when no protein is present, and an increase in FRET as the ssDNA of the partial duplex is reeled in and the two fluorescent dyes approach one another. This would likely be followed by a rapid decrease in FRET as the protein lets go of the ssDNA region and there is a greater distance between the two florescent dyes. See Figure 4.26. Here, when high FRET states occur, this is indicative of the ssDNA end moving near the ds/ssDNA junction, a motion mediated by the helicase, anchored at the junction.

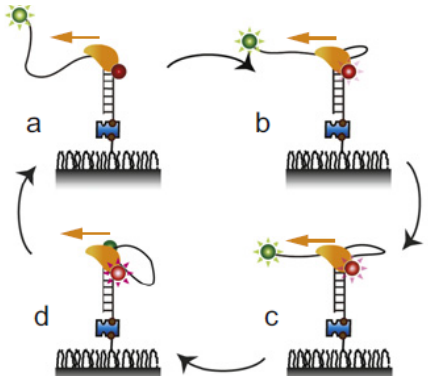


Figure 4.26: Repetitive reeling by DNA helicases can be detected by smFRET

The helicase anchors itself at the ds/ssDNA junction where it pulls through ssDNA therefore bringing the two fluorophores closer together leading to an increase in FRET. This is rapidly released leading to a sharp decrease in FRET before occurring again. Adapted from Park *et. al.* 2010[237].

4.11. HsaHelQ – RPA interaction

Direct physical interaction has been reported between archaeal Hel308 and archaeal RPA[186], and between HsaHelQ and RPA70[185]. As described in Chapter 1, RPA is important for all aspects of DNA metabolism and is utilized heavily during DNA replication and acts during DNA damage signaling to initiate the DNA damage response in the event of physical DNA damage and stalled replication forks. It is possible that physical interaction between HsaHelQ and RPA is used as a means of HsaHelQ recruitment to sites of damage or sites of stalled replication, or that HsaHelQ is required for clearance of RPA from ssDNA to allow for loading of other proteins such as machineries required for replication fork restart.

The report of RPA70-HsaHelQ interaction by pull down of FLAG-tagged HsaHelQ from human cell extracts, followed by mass spectrometry, also showed various other interactions of HsaHelQ with recombination proteins[185]. We probed for direct physical interaction between HsaHelQ and commercially available human RPA complex using EMSAs as described in section 2.22, Figure 4.27. Cy5 labeled fork 3, a model fork with a ssDNA gap, was used as a means of detection. HsaRPA was mixed with Fork 3 at 0.4 pM and 4 pM gave two stable protein-DNA complexes (Fig. 4.28 lanes 6 and 7), which became a single complex when 40 pM HsaRPA was added (lane 8). His-SUMO-HsaHelQ also gave a complex with this fork (lanes 2-4) that migrated as a doublet, probably because of some proteolytic clipping of His-SUMO-HsaHelQ that was typical of the purification of this protein. Titration of HsaRPA into 128 nM His-SUMO-HsaHelQ mixed with Fork 3 resulted in the observation of a new complex that was distinct from complexes formed either with His-SUMO-HsaHelQ alone or HsaRPA alone (lanes 3 and 4). Significantly, this new complex migrated more slowly than either single protein complex, and is therefore consistent with a "supershifted" complex resulting from both HsaRPA and His-SUMO-HsaHelQ binding together.

To verify that this supershift occurred as a result of both His-SUMO-HsaHelQ and HsaRPA being present with Fork 3, a western blot (section 2.15.2) of the EMSA gel was carried out to try to detect each protein in positions consistent with the position of the new complex, a method successful for analysis of archaeal Hel308-RPA complexes[186]. This

was unsuccessful however, because no protein could be detected following probing for RPA using anti-RPA antibodies (not shown). A simple explanation for this could be that the quantities of protein used in the EMSA and present in the super-shift complex were too low to be detected *via* the antibodies used. This was found to be the case, summarised in Figure 4.28. In this, His-SUMO-HsaHelQ and HsaRPA were titrated from 0 – 2,000 μg and analysed by SDS-PAGE and western blot to test sensitivity of the anti-His and anti-HsaRPA primary antibodies that were used to probe the HsaHelQ-HsaRPA-Fork 3 interaction EMSA. The anti-His antibody could only detect 10 ng protein, shown in Figure 4.28B, lane 6, and the anti-RPA antibody could only detect 100 ng, shown in Figure 4.28B, lane 7. The highest mass of HelQ and RPA used for EMSA analysis was 350 and 0.05 ng respectively and therefore RPA could not be detected *via* this method. Although the EMSA indicates that both HelQ and RPA are co-incident in binding to the fork, we cannot rule out the possibility that the proteins do not physically interact with one another but rather are binding to DNA in different parts of the substrate. This would also appear as a "super-shifted" RPA-HelQ-fork complex. However, although both RPA and HelQ bind avidly to ssDNA they do not bind to duplex DNA. There is only one region of 25-nucleotide ssDNA present on this fork substrate, and it is questionable whether this could accommodate both RPA and HelQ without interaction. Furthermore, it is interesting to note that physical association between HelQ and RPA has been detected *via* mass spectrometry[185]. In summary, the supershift

observed from His-SUMO-HsaHelQ and HsaRPA is worth pursuing using other methods for verification and to establish if it is functionally important.

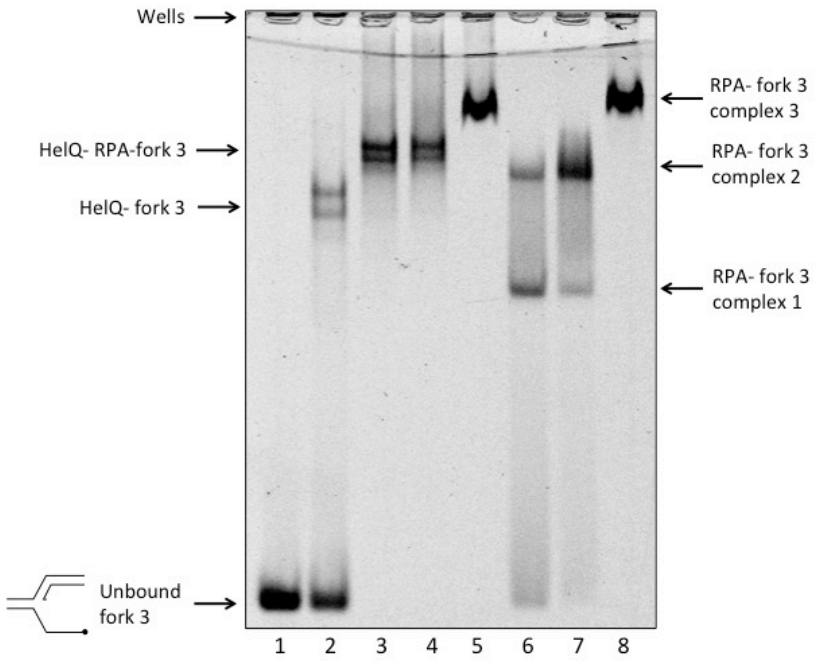


Figure 4.27: HsaHelQ and HsaRPA supershift on EMSAs. Reactions were at RT for 10 minutes and contained 1x HB, 128 nM His-SUMO-HsaHelQ, and 25 nM Cy5 labeled DNA. Unbound fork 3 DNA is shown in lane 1, 128 nM His-SUMO-HsaHelQ causes a band shift with fork 3 shown in lane 2. Human RPA complex was titrated at 0, 0.4, 4 and 40 pM against fork 3 (lanes 6-8). Addition of RPA to HsaHelQ resulted in supershifts observed in lanes 3 and 4.

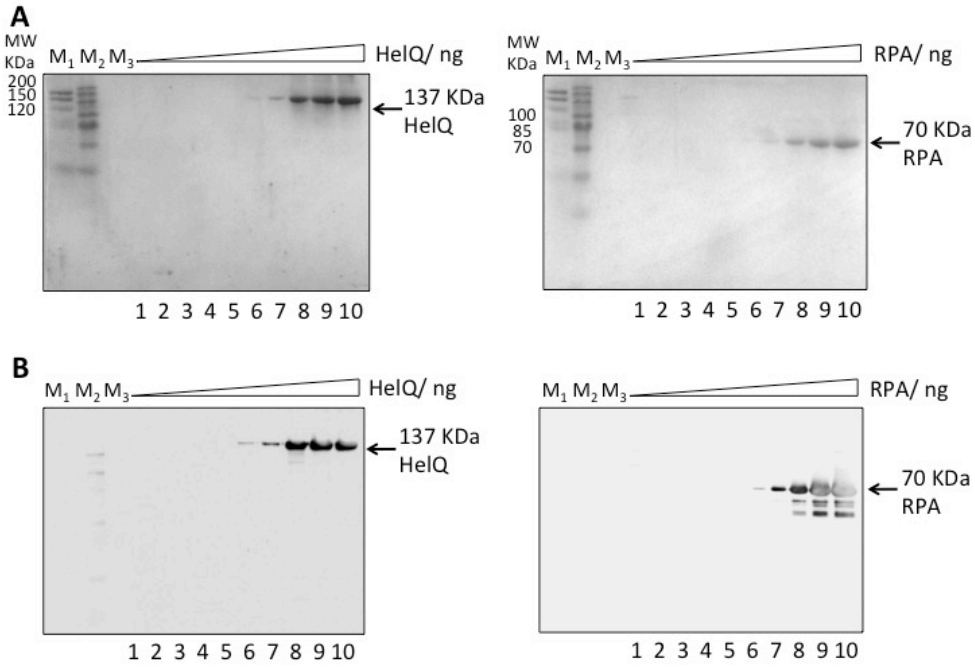


Figure 4.28: Anti-His and anti-RPA antibody sensitivity.

(A) SDS-PAGE analysis of HsaHelQ and RPA proteins titrated from 0-2,000 ng.

(B) Western blot analysis of HsaHelQ and RPA proteins titrated from 0-2,000 ng. Membranes were probed with 1:1,000 anti-His and anti-RPA antibodies respectively. Anti-His only detects protein in lane 5, corresponding to 10 ng protein and anti-RPA only detects protein

4.12 Summary

This chapter has generated new data suggesting that human HelQ exists as an oligomer, and that it is active as a helicase in this form. The oligomeric state of HsaHelQ may be consistent with being a tetramer, as has been observed for the inactive helicase domain of its homolog HsaPolQ. We also offer a rationale for the molecular basis of this oligomerisation, based on existing crystal structures, which can now be tested experimentally to understand the mechanism of helicase action by

HsaHelQ in more detail. Analyses of DNA binding preferences of HelQ were inconclusive and preliminary biophysical analysis of HelQ was useful for determining the correct parameters for subsequent experiments, although did not yield useful data at this stage. We also demonstrate biochemically a physical interaction between HsaHelQ and HsaRPA, which may be physiologically relevant for how HsaHelQ gains access to helicase substrates *in vivo*.

Chapter 5: A conserved winged helix domain in Hel308 and HelQ binds to duplex DNA to promote helicase activity

5.1 Introduction

Hel308 enzymes are conserved across all clades of archaea, but are absent from bacteria. Hel308 homologues in metazoans, including humans are known as HelQ (HsaHelQ), Figure 5.1 phylogenetic tree. Hel308 and HsaHelQ are most closely related in sequence and structure to the Ski2 family of RNA processing helicases, and to a putative helicase HFM1, although Hel308 is a ssDNA stimulated enzyme.

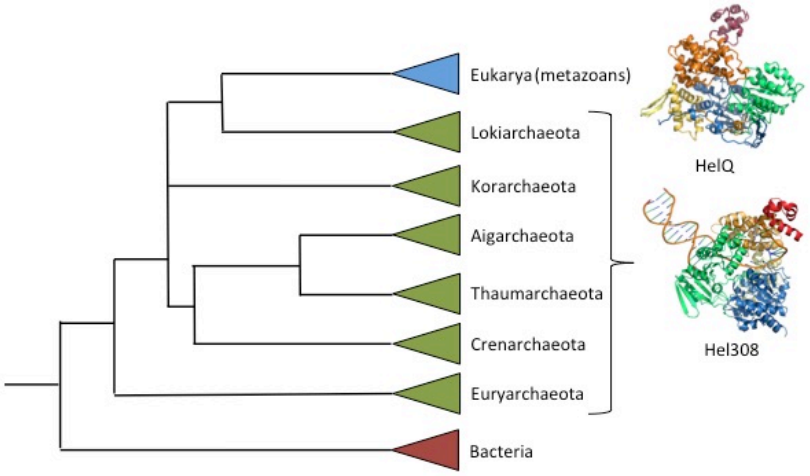


Figure 5.1: Representation of a phylogenetic tree displaying distribution of Hel308 and Hel308 homologues across the bacteria, archaea and eukaryotes. Hel308 is conserved across all archaea, and has homologues in metazoans; in humans Hel308 is known as HelQ. Hel308 is noticeably absent from the bacteria. Adapted from Spang *et. al.*[19].

Hel308 is suggested to participate in DNA repair at stalled replication forks and is known to preferentially process branched DNA structures. In order to understand how this is achieved, it is important to know how Hel308 engages with DNA branches, which may contain single stranded and duplex DNA regions.

The aim of this chapter was biochemical analysis of the Winged Helix domains (WHDs) of HsaHelQ and archaeal Hel308 proteins. It was to test the hypothesis that the WHDs of these helicases directly participate in DNA binding and in so doing activate or promote helicase activities at branched substrates. The rationale for this was as follows; Hel308 possesses three core helicase domains that confer translocation and helicase activities by encircling and making direct physical contacts with ssDNA (Figure 5.2, in grey)[191]. The core comprises RecA-like domains one and two that assemble to form motor domains that hydrolyse ATP, positioned at the domain 1 and 2 interface where the Walker A and B motifs are located. This is the driving force for Hel308 translocation along single stranded DNA. A β-hairpin located on domain 2 is responsible for separation of duplex DNA. The core helicase unit is shown in in Figure 5.2A. The third component of the helicase core is domain 4, the helicase ratchet, that 'grips' DNA to prevent slippage and facilitates unidirectional 3' to 5' translocation by the RecA-like motor domains. This is suggested to operate in an 'inchworm' like mechanism. Hel308 comprises two additional accessory helicase domains, shown in red in Figure 5.2B, domain five is known to act as a molecular brake and makes extensive

contacts with single stranded DNA, and a winged helix domain (WHD) (Hel308 domain three) of unknown function. Accessory domains confer unique activities to helicases that enable them to carry our specific functions/ unwinding at specific sites or situations. In the crystal structure of AfuHel308 the WHD is packed closely against RecA-like domain 1 and is located far from the DNA, making no contacts with this particular 3'-OH tailed partial duplex substrate, Figure 5.2C. It is important to note that this is the minimal helicase substrate for Hel308, and that Hel308 enzymes are more active when unwinding branched DNA molecules, such as model replication fork structures and D-loops[5]. The position of the WHD and its distance from DNA in the minimal DNA substrate suggests a role for the WHD that is independent of the translocation mechanism.

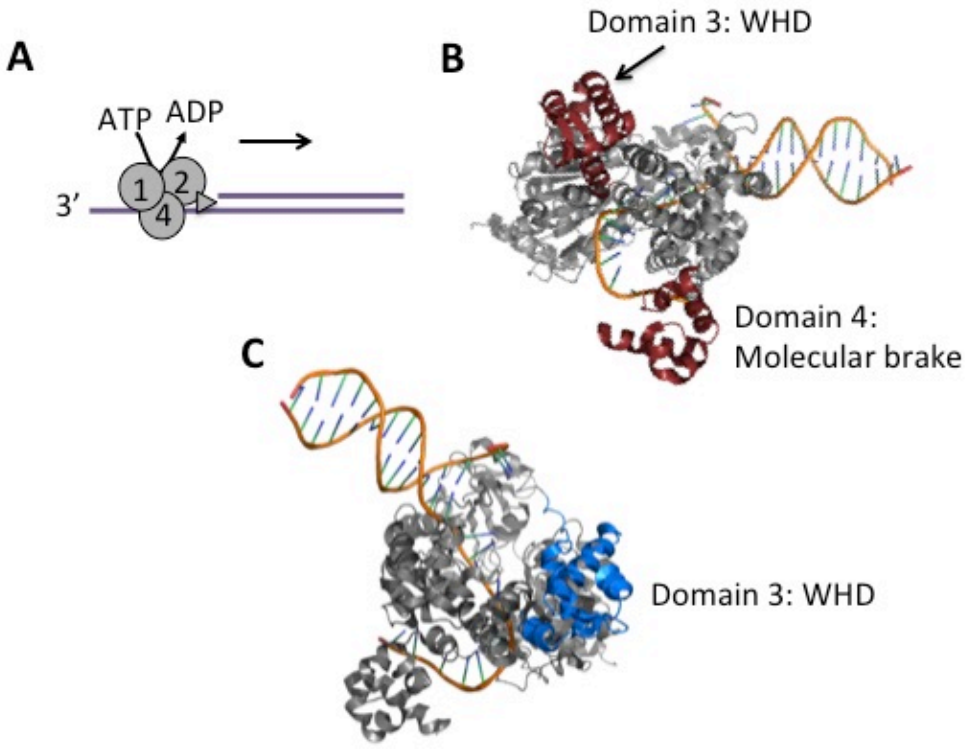


Figure 5.2: Hel308 core helicase and accessory domains:

(A) Schematic displaying Hel308 domains 1, 2 and 4 that comprise the helicase core for ATP-coupled to 3' to 5' translocation along ssDNA aided by domain 4, the helicase ratchet.

(B) AfuHel308 atomic resolution structure in complex with partial 3'-OH ssDNA-tailed duplex DNA. Helicase core domains are presented in gray and two accessory domains are highlighted in red: These are domain 3, the WHD located at the top of the image, and domain 5, the molecular brake, that engages with ssDNA. Atomic resolution coordinates were downloaded from PDB 2P6R and were rendered using MacPymol.

(C) AfuHel308 domain 3 WHD highlighted in blue. The WHD is solvent exposed and makes no contacts with the DNA substrate in this structure.

WHDs are commonly found in many DNA binding proteins, including DNA repair helicases that may have analogous functions to Hel308/HelQ, such as bacterial PriA and human RecQ1. In PriA and RecQ1 the WHDs bind to duplex DNA through the major groove *via* a WHD solvent exposed α recognition helix, and without sequence specificity[140]. WHDs have also been reported to act as protein-protein interaction units. For example, RPA contains a WHD that interacts with essential DNA repair proteins[238]. This is a potential role for the WHD of Hel308, for example because direct physical interactions between archaeal Hel308 and RPA have been detected[186]. In addition to this, multiple protein-protein interactions have been reported for HsaHelQ and DNA repair proteins, one of these being between RPA70[185].

The WHD of Hel308 and HelQ had not, until now, been investigated biochemically. Crystal structures from several archaeal Hel308 enzymes give few clues about the role of the WHD, including if or how it might interact with DNA. We investigated if the Hel308 WHD may have a role binding duplex DNA, leading from an hypothesis that the WHD may be able to bind duplex regions in branched DNA structures, preferred substrates for helicase activity, as part of the process for the enzyme engaging its physiological targets for remodeling during DNA repair.

5.2 A proposed DNA recognition helix (α -20) in the MthHel308 WHD

Sequence alignments using Clustal Omega revealed 34% overall sequence similarity between structurally determined Hel308 from

AfuHel308 and that of MthHel308, the archaeal Hel308 enzyme studied in this work. Alignments of Hel308 and HelQ sequences are presented in the appendices. There is 31% identity between the residues of the WHD. Some conserved residues in the WHD (Y460, Q459 and N464 according to amino acid numbering for MthHel308), are present within a solvent exposed proposed recognition α-helix, referred to as α-20 from hereon. In addition, conserved phenylalanine residues were present within a buried region of the WHD that forms an interface with RecA-like domain 1 (Figure 5.3), to be discussed later.

The AfuHel308 WHD structure highlights the position of solvent exposed α-20 recognition helix (Figure 5C). To investigate the properties of MthHel308 WHD we focused on the conserved α-20 tyrosine residue (Y460), because of its chemical properties giving potential to interact with DNA bases hydrophobically and by hydrogen bonding. In addition, conserved polar residues located close to Y460, Q459 and N464, are typical DNA binding amino acids and can hydrogen bond to DNA. Genetic, biochemical and biophysical analysis was used to investigate the roles of Y460, Q459 and N464 residues on the α-20 recognition helix. This includes analysis of a triple mutant of the MthHel308 WHD, creating Hel308 with the following amino acid substitutions: Q459G, Y460A and N464G, herein referred to as MthHel308^{GAG}.

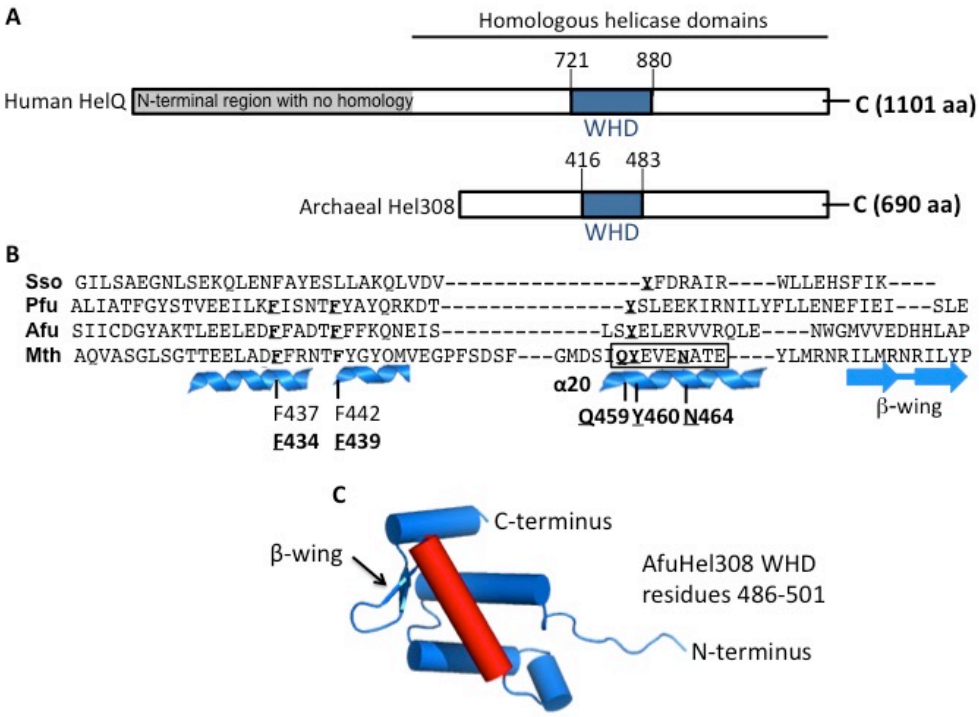


Figure 5.3: WHDs in archaeal Hel308 and human HelQ proteins.

(A) Cartoon representation of WHD position in human HelQ and archaeal Hel308 proteins where they are located between amino acids 721-880 and approximately 416-483 respectively, highlighted in blue.

(B) Amino acid sequence alignments of WHD regions of *Sulfolobus solfataricus*, *Pyrococcus furiosus*, *Archaeoglobus fulgidus* and *Methanothermobacter thermautotrophicus* generated using Clustal Omega, and corresponding secondary structure elements. Conserved residues are highlighted in bold, located on the α-20 helix.

(C) AfuHel308 WHD atomic resolution structure. The solvent exposed α-20 helix is highlighted in red. Coordinates were downloaded from PDB 2P6R and were rendered using MacPymol.

5.3. Genetic analysis of MthHel308 WHD α-20 mutants

A genetic assay, described in section 2.30, was used to assess if MthHel308^{GAG} was functional when compared to WT MthHel308, Figure 5.4A. The assay uses *E. coli* strain *dnaE486* Δ*recQ*, which is described fully in Hishida *et. al.* 2004[239], and was used previously to characterize mutations in Hel308 or RecQ helicases[5]. Briefly, a point mutation in DNA polymerase III (encoded by *dnaE*) destabilizes the active replisome at a semi-permissive temperature 37^{o}C, giving a propensity for replication fork stalling that is manifest as 100-1,000 fold reduced cell viability. No such effect is apparent at 30^{o}C, a permissive temperature. The negative effect of the *dnaE* mutation on replication at 37^{o}C can be alleviated by deletion of the *recQ* gene (Δ*recQ*), encoding the DNA repair helicase RecQ. These phenotypes are summarised in Figure 5.4A: when *dnaE* Δ*recQ* cells were transformed with an "empty" plasmid vector control (pEmpty), cell viability is robust, compared to the same cells in which RecQ helicase is re-introduced by being expressed from a plasmid, in agreement with previous studies[5]. Inactivation of MthHel308 ATPase and helicase activity by mutation of its Walker A lysine residue (pHel308^{K51L}) restored cells to normal growth like pEmpty (Figure 5.4A and Table 5.1). The MthHel308^{GAG} protein was associated with intermediate viability, improved at least 10 fold from WT MthHel308, but much reduced compared to fully inactive Hel308^{K51L}. This phenotype indicated sub-optimal activity of Hel308^{GAG} that was worthy of further analysis biochemically. Expression levels of Hel308^{GAG} and WT MthHel308 were very similar during the assay (Figure

5.4B) confirming a functional effect of the MthHel308^{GAG} mutation, rather than an effect caused by reduced or unstable protein. This indicated that pHel308^{GAG} is defective, demonstrating that amino acid residues Q459, Y460 and N464 are required for efficient MthHel308 function and therefore are worth pursuing experimentally. This interesting MthHel308 mutant was subsequently purified for biochemical analysis.

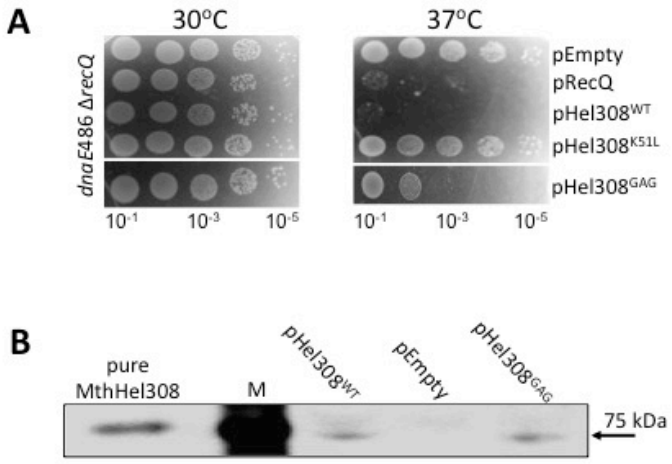


Figure 5.4: Genetic analysis of pHel308^{GAG} in *dnaE*486*ΔrecQ E. coli.*

(A) Genetic assay for assessing activity of the Hel308^{GAG} mutant, which has amino acid substitutions in WHD α-20. The *E. coli* strain (*dnaE486* Δ*recQ*) has a conditional DNA replication mutation, detailed in section 5.2. Plasmid expression of WT MthHel308 (pHel308^{WT}) is toxic, causing low cell viability but cells expressing MthHel308^{GAG} at 37°C showed much improved viability more similar to cells containing empty plasmid lacking MthHel308 (pEmpty).

(B) Western blot detection using purified antibodies for MthHel308, showing similar expression of Hel308^{WT} and Hel308^{GAG} proteins in *E. coli dnaE486* Δ*recQ* cells, as indicated alongside cells lacking Hel308 expression (pEmpty), during the course of the genetic assay described above. The marker was a Biotinylated protein ladder from Cell Signalling, which includes an 80-kDa marker band closely corresponding to the size of Hel308.

Table 5.1: Viability of *dnaE*486 Δ*recQ* *E. coli* expressing MthHel308 proteins.

Relative viability values for *E. coli* strain *dnaE*486 Δ*recQ* transformed by plasmids expressing either no helicase (pEmpty), or expressing RecQ or Hel308 helicases. See Figure 5.4A for a summary picture. Viability was calculated from "spot-test" assays in triplicate grown at the same time from a single starter culture in each case. The mean value for pEmpty (12.3 colonies, at cell dilution of 10^{-5}) was designated as 1.0, from which to compare viabilities of other plasmids.

Plasmid	Viability (standard deviation)
pEmpty	1.0 (+/- 0.16)
pRecQ	2.7×10^{-4} (+/- 6.2×10^{-5})
pHel308	6.5×10^{-5} (+/- 1.6×10^{-5})
pHel308^{K51L}	1.0 (+/- 0.14)
pHel308^{GAG}	1.4×10^{-2} (1.4×10^{-3})

5.4. Biochemical analysis of purified Mth Hel308 WHD α-20 mutant proteins.

The mutations of the MthHel308 WHD described by genetic analysis were introduced into the full length Hel308 protein for purification, described in section 2.16.5. MthHel308^{GAG} was purified as described for MthHel308^{WT}, and did not exhibit any differences to purification of WT MthHel308 protein. SDS-PAGE analysis of purified MthHel308^{WT} and MthHel308^{GAG} proteins is shown in Figure 5.5, in addition to other MthHel308 WHD mutant proteins that are discussed later. I also attempted purification of MthHel308 WHD mutants in which the α-20 recognition helix was replaced by a flexible linker

(PSVAVEVAPGVPAVEEGAVPAV), and introduced amino acid substitutions R477G and R479G in the β-wing, but we were unable to express these proteins.

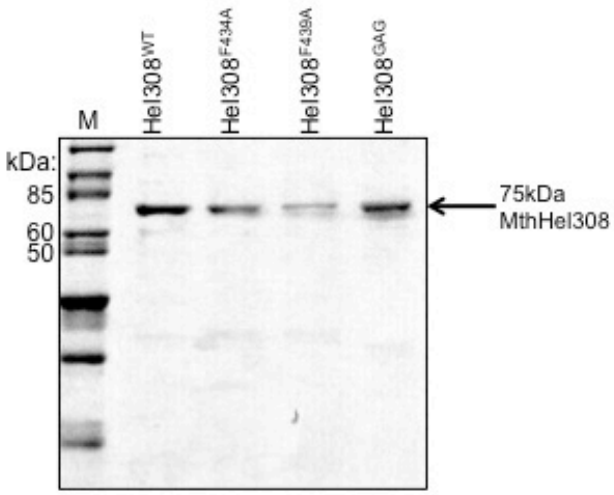


Figure 5.5: Purified WT and mutant MthHel308 proteins.
8% acrylamide SDS-PAGE gel showing purified WT and mutant Hel308 proteins at 75 kDa.

Genetic analysis had demonstrated that MthHel308^{GAG} protein was to some degree inactive, compared to wild-type protein. Catalytic activity of MthHel308^{GAG} was measured in DNA unwinding assays, as described in section 2.23, as a function of time using a 3'-OH overhang partial duplex, fork 2 and fork 5, each containing regions of duplex DNA, Figure 5.6. In comparison to MthHel308^{WT}, MthHel308^{GAG} showed diminished DNA unwinding of all DNA substrates by at least 50%, especially the fully base paired fork that showed approximately 70% reduction in unwinding efficiency, Figure 5.6A. Although a reduction in DNA unwinding by MthHel308^{GAG} was observed, this protein still unwound all DNA substrates to release the same products as for

MthHel308^{WT}. This demonstrates that these residues located on α-20 are required for efficient unwinding of the DNA substrates listed above.

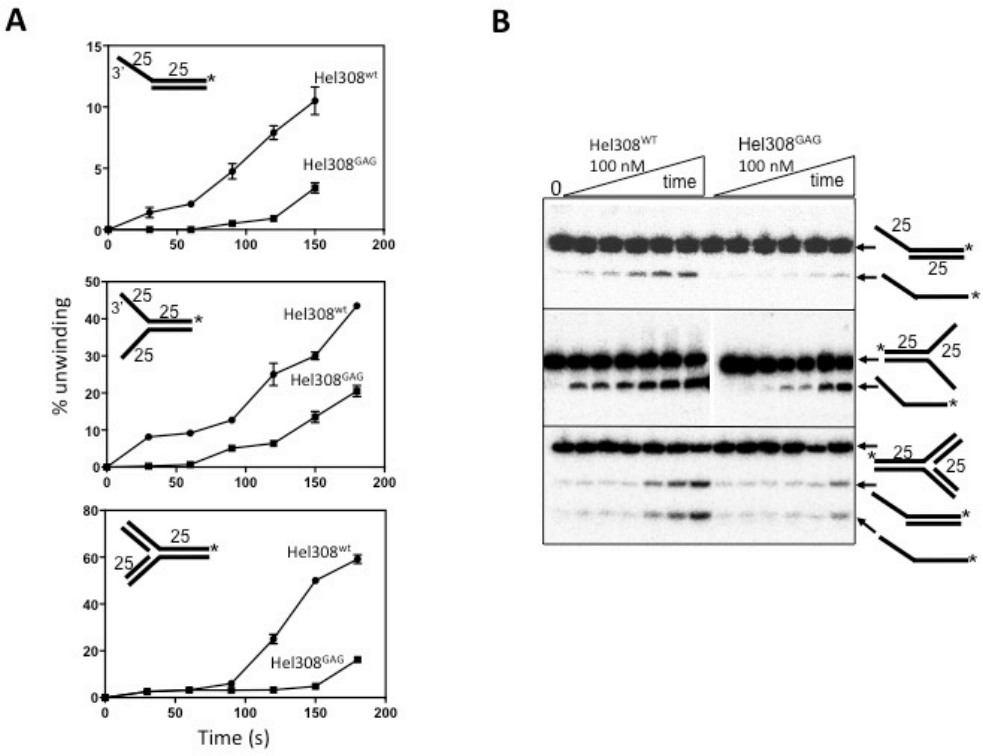


Figure 5.6: Helicase activity of Hel308^{GAG} compared to wild type Hel308.

(A) Graphs show unwinding activity as a function of time for each protein mixed with 3'-OH ssDNA tailed duplex, frayed duplex and fork DNA substrates as indicated, with DNA strand lengths (nucleotides) indicated next to each substrate. In each case protein (100 nM) was mixed with 5' end-radiolabelled (*) DNA substrate (10 nM). Samples of the reaction mixture were taken at the following time points (in seconds), as indicated in the graphs; 0, 30, 60, 90, 120, 150 and 180, except for 3'-OH tailed duplex which lacked a 180 s sample. For each substrate the assay was done three times and error bars represent the standard deviation from the mean of each data point, calculated using GraphPad Prism software.

(B) Gel panels give one example gel for each Hel308/Hel308^{GAG} unwinding assay that was quantified, indicating substrate and product DNA molecules.

EMSAs were used to assess binding of MthHel308^{GAG} to the same DNA substrates as used in helicase assays, and a ssDNA molecule, as described in section 2.22, Figure 5.7. Titration of MthHel308^{GAG} against each substrate gave generally similar binding activity compared to MthHel308^{WT}, but with reproducibly subtle reduction in binding of MthHel308^{GAG} to all substrates except ssDNA Figure 5.7A. MthHel308^{GAG} showed formation of two stable DNA complexes with both forked substrates, indicative of numerous MthHel308 monomers binding to the DNA substrates, a phenomenon observed consistently for MthHel308^{WT}, shown in Figure 5.7B in lanes 1-4.

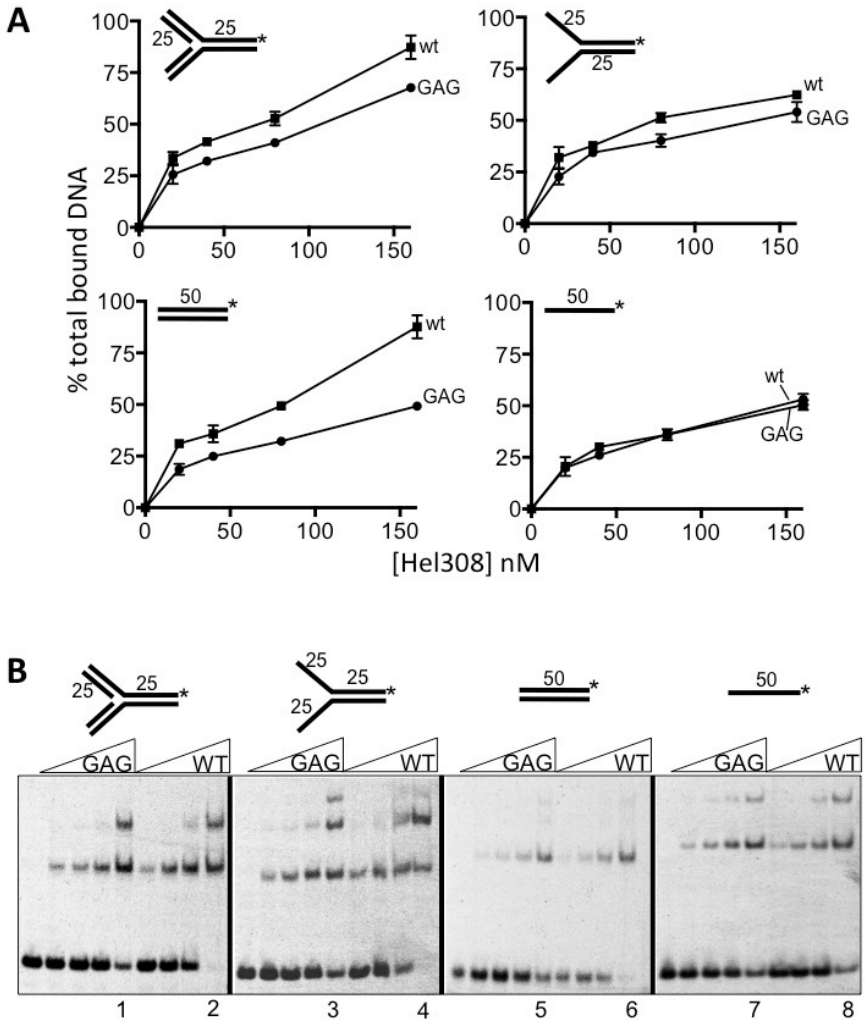


Figure 5.7: EMSAs summarizing DNA binding by MthHel308^{GAG} compared to MthHel308 wild type protein.

(A) DNA binding of WT MthHel308 and WTHel308^{GAG} (GAG) in EMSAs. Graphs show quantification of MthHel308-DNA binding complexes formed with the substrates indicated, from reactions containing MthHel308 protein at 0, 20, 40, 80 or 160 nM and DNA at 10 nM. Assays were duplicated and error bars represent the standard deviation from the mean of each data point, calculated using GraphPad Prism software.

(B) Gel panels summarizing DNA binding EMSAs, giving one example of each assay; lanes 1-6 highlight modest but reproducible differences in binding of WT MthHel308 and MthHel308^{GAG} ('GAG') to forked and duplex substrates that was not observed in binding to ssDNA (lanes 7 and 8).

We had initially proposed that α-20 might be involved in mediating DNA binding. Superficially the only very modest reduction in DNA binding by MthHel308^{GAG} compared to MthHel308^{WT} might be interpreted to argue against a role for the WHD in duplex DNA binding. However, Hel308 domains 1, 2, 4 and 5 make a large number of contacts to DNA, which raised the possibility that significant defects in DNA binding by the WHD mutations may be masked by the many other protein-DNA contacts. In light of this idea, the wild type MthHel308 WHD was next studied in isolation, in addition to two mutant isolated WHD mutant proteins; one that carries the same triple mutation on α-20 as described for full length MthHel308^{GAG}, and a mutant where α-20 was replaced with a linker containing helix disrupting prolines. We reasoned that this analysis of isolated WHD proteins would enable better understanding of any MthHel308 WHD DNA binding, in EMSAs in which reduced DNA binding resulting from WHD mutations would not be overshadowed by DNA binding by Hel308-DNA contacts from domains 1, 2, 4 and 5.

5.5. Purification of analysis of isolated WHD proteins from MthHel308

A putative WHD of MthHel308 was identified from amino acid sequence alignment with AfuHel308, for which there is an atomic resolution structure. The MthHel308 WHD was cloned and the resulting 11 kDa WHD protein was purified. Isolated WHDs were insoluble when over-expressed and were therefore dialyzed to attempt refolding. SDS-PAGE analysis of purified WHD proteins is shown in Figure 5.8.

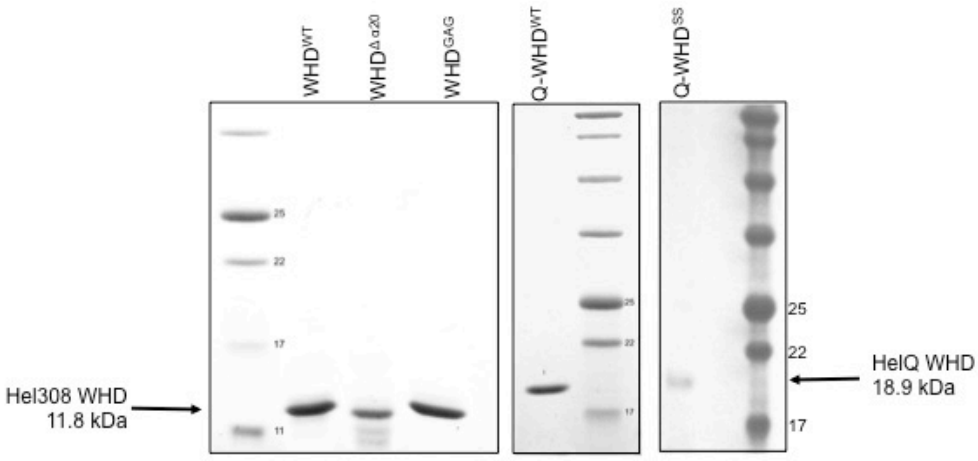


Figure 5.8: Purified WT and mutant MthHel308 WHD proteins in isolation.

12% acrylamide SDS-PAGE gel showing purified WT and mutant MthHel308 WHD proteins at 11.8 kDa, and HsaHelQ WHD proteins at 18.9 kDa.

Analytical gel filtration (AGF), described in section 2.25, was used to assess re-folding and confirmed a majority of single folded species, shown by sharp elution peaks within the expected size range of 11 kDa, indicating that the WT and mutant Mth Hel308 WHD proteins had largely refolded correctly, Figure 5.9. AGF is sensitive not only to protein size, but also protein shape, and therefore AGF traces can indicate mis-folded protein, often displayed by traces containing multiple, often broad, sloping peaks. This is due to the mis-folded proteins that although are of the correct size, adopt different shapes, therefore they elute within a broad range from the column rather than a sharp peak indicative of one species of the same conformation. In addition to individual amino acid substitutions made in α-20 of the isolated WHD, the same helix was also

replaced with a flexible linker containing prolines to act as helix breakers, $WHD^{\Delta\alpha20}$. Mutant WHD proteins were also analyzed by gel filtration to confirm correct protein folding. AGF analysis of refolded WHD^{WT} and WHD^{GAG} proteins is show in Figure 5.9.

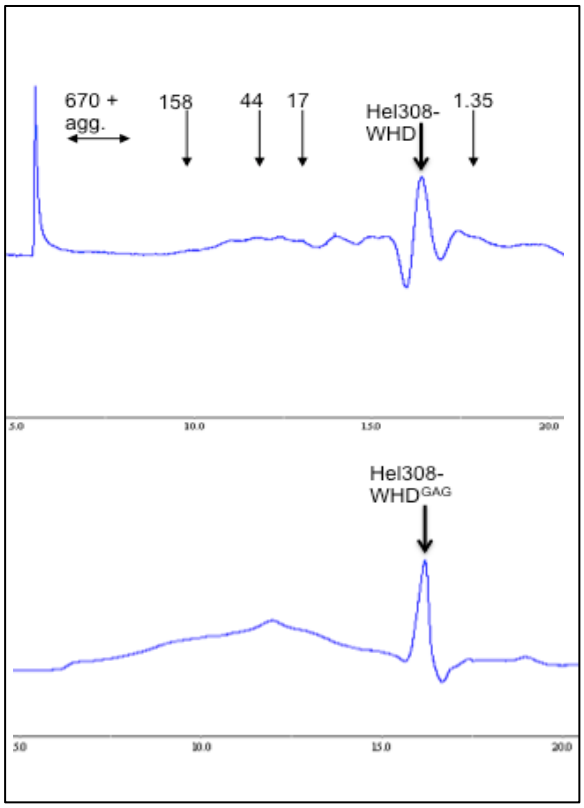


Figure 5.9: AGF indicates the majority of refolded MthHel308 WHD proteins are homogeneous.

Superose 12 Analytical gel filtration traces showing analysis of refolded WHD proteins. Elution volumes from peaks from BioRad Gel Filtration Standards and their corresponding size in kDa are indicated. Both WT MthHel308 and $MthHel308^{GAG}$ elute as sharp peaks indicating protein of one species at the correct molecular weight.

EMSAs were used to assess DNA binding by isolated WHD proteins, Figure 5.10. WHD proteins were titrated against DNA substrates containing either ssDNA or duplex DNA, or both. This showed that WHD^{WT}

forms stable complexes with 70 bp linear duplex DNA, as shown in 5.10B where 21% DNA is shifted, however, this is abolished in WHD^{GAG} and $WHD^{\Delta\alpha20}$ mutants. Stable complexes were also formed between WHD^{WT} and other linear duplex DNA molecules of different sequence to ensure binding wasn't sequence specific. This suggests that the conserved residues Q459 Y460 N464 located on α-20 are required for WHD binding to duplex DNA. WHD^{WT} binding to simple DNA substrates was next assessed; where WHD^{WT} only formed stable complexes with DNA substrates containing regions of duplex DNA, Figure 5.10C. WHD^{WT} binding to linear duplex is similar to full length $MthHel308^{WT}$. However, interestingly, WHD^{WT} binding to partial duplex is very weak in comparison to strong binding by full length $MthHel308^{WT}$, where the only difference in the DNA substrate is a short region of ssDNA. This is highlighted further where WHD^{WT} will not form stable complexes with ssDNA, in stark contrast to full length *Mth*$Hel308^{WT}$ that shifts 100% ssDNA, likely mediated by the core helicase domains 1, 2 and 4 in addition to accessory domain 5. Together, this implies a role for the MthHel308 WHD in binding to duplex DNA, rather than ssDNA. Although clearly only able to form stable complexes with linear duplex DNA, WHD^{WT} can only achieve weak DNA binding in comparison to ssDNA binding by full length MthHel308. This significant observation is discussed later as part of a model for how the Hel308 WHD may assist in processing of appropriate physiological substrates.

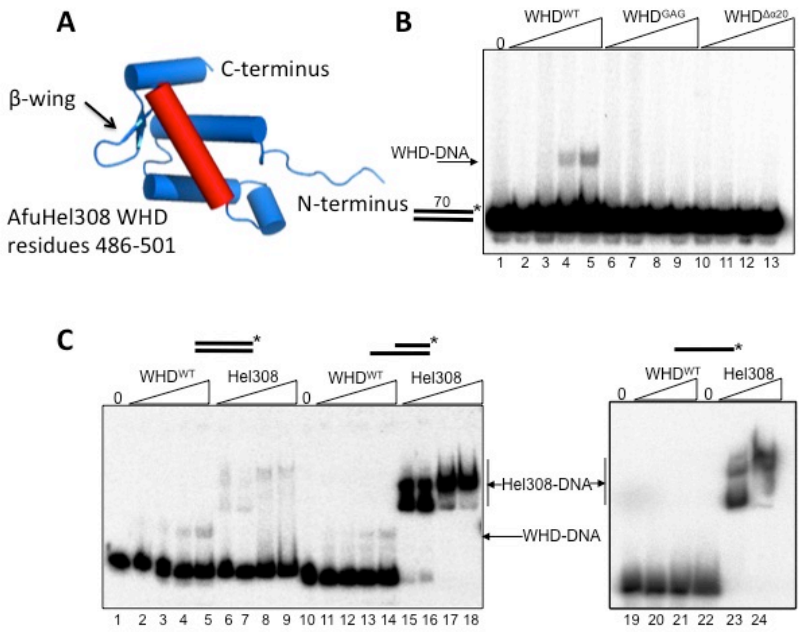


Figure 5.10: Isolated MthHel308 WHD preferentially binds duplex DNA.

(A) AfuHel308 WHD atomic resolution structure. The solvent exposed α-20 helix is highlighted in red, where WHD^{GAG} and $WHD^{\Delta\alpha20}$ mutations lie. Coordinates were downloaded from PDB 2P6R and were rendered using MacPymol.

(B) EMSAs of DNA binding by isolated WHD protein from MthHel308. Purified wild type WHD protein (WHD^{WT}) bound duplex DNA (lanes 1-5), but binding was lost from WHD proteins carrying substitution of amino acid residues within alpha helix 20 (WHD^{GAG}, lanes 6-9), or by removal of alpha helix 20 ($WHD^{\Delta\alpha20,}$ lanes 10-13). Radiolabelled duplex DNA (10 nM, *) was incubated with WHD proteins at 0, 25, 50, 100 and 200 nM, as indicated.

(C) WHD^{WT} binds to duplex DNA (lanes 1-5) and partial duplex (lanes 10-14), but not to ssDNA (lanes 19-22) in EMSAs. Binding of full length Hel308 is shown for comparison, to duplex DNA (lanes 6-9), partial duplex (lanes 15-18) and to ssDNA (lanes 23 and 24). Reactions contained 6 nM of radiolabelled (*) DNA substrate of the type indicated, and proteins at 0, 25, 50, 100 and 200 nM (lanes 1-18) or at 0, 100 and 200 nM (lanes 19-24).

5.6. Purification of analysis of isolated WHD proteins from human HelQ

Using Phyre 2 modeling of HsaHelQ and amino acid sequence conservation, detailed in Chapter 4, the HsaHelQ WHD was identified, cloned and purified in isolation, called Q-WHD. The AfuHel308 WHD atomic resolution structure and the Q-WHD and MthHel308 WHD Phyre2 models adopt a similar overall shape, Figure 5.11. According to the Phyre 2 model, Q-WHD contains an extra alpha helix and in addition, there is a long β wing that is not found in the AfuHel308 WHD. These might be considered major differences, however, it is important to note that this is a model of Q-WHD. Other differences such as the presence of additional α-helices may be attributed to differences in protein size, given that HsaHelQ is 124 kDa versus archaeal Hel308 proteins of approximately 75 kDa. The HsaPolQ WHD adopts similar topology to AfuHel308 WHD. It is possible that the HsaHelQ WHD looks similar in reality, and that any major differences may be artifacts of the computational molecular modeling process.

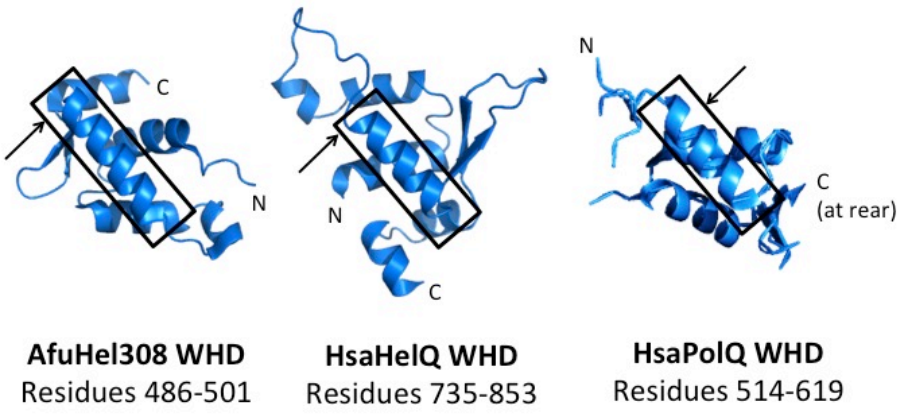


Figure 5.11: AfuHel308 WHD and HsaPolQ WHD atomic resolution structures and Phyre 2 HsaHelQ WHD models share similar topology.

Atomic resolution structures for AfuHel308 and HsaPolQ WHDs in isolation and Phyre 2 HsaHelQ WHD model. All structures adopt similar topology despite variations in size and number of helices. Structures are orientated with proposed recognition helix α-20 at the front, boxed and indicated by arrows.

HsaHelQ and MthHel308 share 25% overall sequence identity when compared in alignment using Clustal Omega (Figure 5.12). Alignment and Phyre 2 models of Q-WHD were used to generate human Q-WHD protein and mutants, in which Y818 and K819 were substituted with S and D residues respectively to create Q-WHD^{SD} and Q825 and Y826 were both substituted with S to create Q-WHD^{SS}, that correspond to those used for MthHel308 WHD described earlier. Again, these residues are typical of those involved in protein-DNA contacts. Q-WHD proteins were purified and refolded by dialysis, and AGF confirmed likely accurate refolding of the protein. SDS-PAGE analysis of Q-WHD can be seen in Figure 5.8.

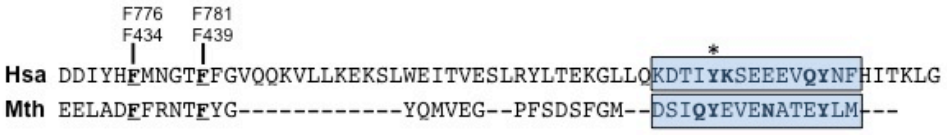


Figure 5.12: Conserved residues between HsaHelQ and MthHel308 WHDs. Clustal Omega alignment of HsaHelQ and MthHel308 WHDs amino acid sequences Hsa: 771-834 and Mth: 429-470. Conserved residues are highlighted in bold, some of which were targeted for mutagenesis.

In EMSAs (Figure 5.13) Q-WHD^{WT} formed stable protein-DNA with a 70 bp linear duplex DNA, even at concentrations as low as 5 nM. However, linear duplex DNA binding for Q-WHD^{SD} and Q-WHD^{SS} was abolished, even at high protein concentrations at 100 nM, indicating a role for these residues in duplex DNA binding. Ability of Q-WHD^{WT} to bind linear duplex, and inability of Q-WHD mutants to bind DNA is similar to activity shown by MthHel308 WHD, suggesting conservation of function between WHDs of MthHel308 and HsaHelQ WHDs.

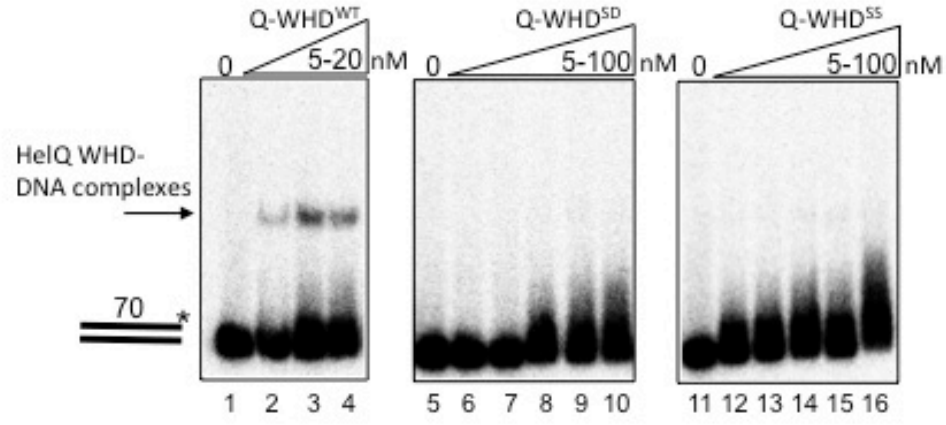


Figure 5.13: EMSAs showing that mutagenesis of conserved residues in HsaHelQ WHD abolished duplex DNA binding.

The isolated HsaHelQ WHD protein binds to duplex DNA (6 nM) at 5, 10 and 20 nM protein (lanes 1-4), but binding was abolished by substitution of amino acid residue pairs Tyr-Lys with Ser-Gly (Q-WHD^{SD}, lanes 7-12) or Gln-Tyr with Ser-Ser (Q-WHD^{SS}, lanes 13-18), at protein concentrations up to 100 nM (0, 5, 10, 20, 50 and 100 nM). Duplex DNA was radiolabelled (*) and incubated with Q-WHD proteins at as indicated.

5.7. Analysis of isolated WHD protein structure by circular dichroism.

Circular dichroism (CD) of the isolated *Mth*WHD^{WT} and mutant WHDs, WHD^{GAG} and WHD$^{\Delta\alpha20}$, was carried out to assess if loss of α-20 helix integrity may be a direct cause of lost duplex DNA binding, described in section 2.29. Circular dichroism (CD) is a biophysical technique used to detect elements of secondary structure in biological molecules by differential absorption of circularly polarized light. Each type of secondary structure, α-helix, β-sheet or random coil, possesses a characteristic CD spectrum. Analysis of WHD^{WT} spectra from 180-240 nm, Figure 5.14, was consistent with α-helices, and showed 16% alpha helix content that represented a significant proportion of secondary structure content in the protein, Table 5.2. This is expected considering it is a structure comprised largely of α-helices in addition to elements of β-sheet secondary structure. In comparison, WHD^{GAG} and WHD$^{\Delta\alpha20}$ proteins showed lower total α-helix content, which is expected for WHD$^{\Delta\alpha20}$ as α-20 had been replaced with a flexible linker. This supports that DNA binding occurs *via* residues located on the α-helix, and that the WT α-20 residues may also contribute to α-helix formation. Although the total sum of secondary structure percentages is greater than 100, values of approximately 120% are just within the limits of reliability.

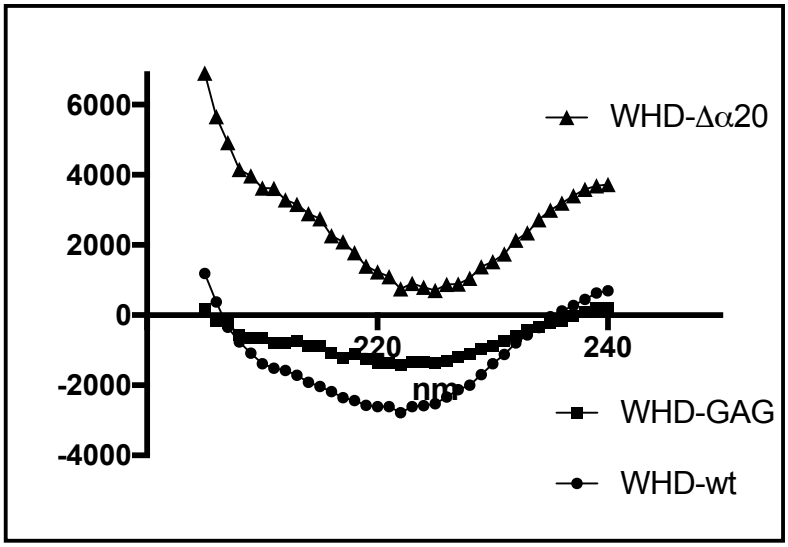


Figure 5.14: Structural analysis of MthHel308 WHD mutants reveals reduced alpha helical content

(A) Typical CD spectra of protein secondary structure elements found in biological molecules.

(B) Circular dichroism spectra for WHD^{WT}, WHD^{GAG} and $WHD^{\Delta\alpha20}$ between 180-240 nm, where WHD mutants show reduced alpha helical content than WHD^{WT}. This was quantified using CDNN software and values can be seen in table 5.3.

Table 5.2: Secondary structure content of WT MthHel308, mutant MthHel308 and HsaHel308 isolated WHD proteins

%	Winged Helix Domain Protein			
	Wild type WHD	WHD^{GAG}	$WHD^{\Delta\alpha20}$	Wild type HsaHelQ WHD
α-helix	**16.0**	**14.1**	**12.2**	14.8
Antiparallel	16.7	18.6	20.1	18.1
Parallel	16.9	18.2	21.3	17.5
β-turn	20.8	21.6	22.0	21.4
Random coil	51.0	52.5	59.0	51.2
Total sum	121.4	125.0	134.7	123.1

5.8. Functional significance for packing of Hel308 domain 1 with WHD

In the crystal structure of AfuHel308 the WHD packs tightly against RecA-like domain 1, a topology common amongst Ski-2 like helicases. Alignment of AfuHel308 with MthHel308 (Figure 5.3B), revealed two conserved phenylalanine residues within the WHDs that in the AfuHel308 atomic resolution structures are located at the tight interface of the WHD with domain 1 (Figure 5.15). These residues, F434 and F439 in MthHel308, are also conserved in HsaHelQ, Figure 5.12.

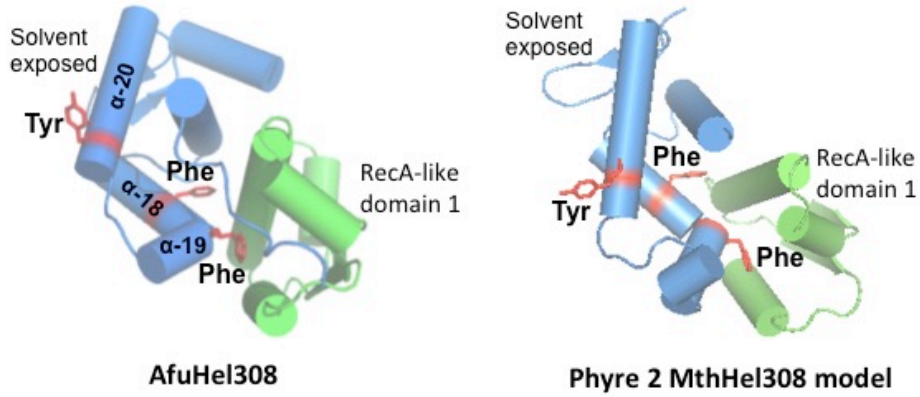


Figure 5.15: AfuHel308 WHD packs against domain 1.
AfuHel308 WHD and domain 1 atomic resolution structure and MthHel308 WHD and domain 1 Phyre 2 model demonstrate packing of WHD, blue, against domain 1, green, that is part of the RecA like motor. There are two conserved phenylalanine residues located at this interface.

Using MthHel308, the two conserved phenylalanine residues were individually substituted for alanine and valine residues generating clones F439A and F434V. Genetic analysis of these two mutants was carried out in *dnaE*486*ΔrecQ E. coli* according to the same assay as described in

section 5.3. As expected, pEmpty and pHel308^{K51L} showed cell viabity consistent with a lack of Hel308 catalytic activity. pHel308^{F434V} showed a phenotype that was intermediate between pEmpty and pHel308^{K51L} and pHel308^{WT}, indicating some abnormality of Hel308 catalytic activity. pHel308^{F439A} gave a phenotype consistent with fully inactive Hel308. These interesting phenotypes gave grounds for further investigation using *in vitro* biochemistry.

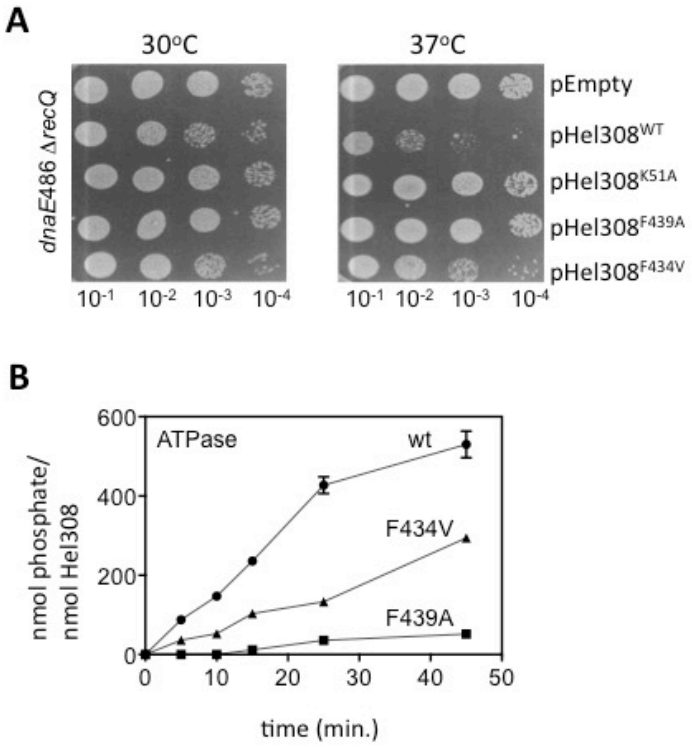


Figure 5.16: MthHel308 F434A and F439A mutants are defective in catalysis.

(A). Genetic analysis of *hel308* expression, as described in section 5.2. Cells expressed either wild type Hel308 (pHel308^{WT}), ATPase defective MthHel308 K51A (5), or Hel308 with substitution of either one of the invariant WHD phenylalanine residues (F434A or F439A), as indicated beside the panel. Empty plasmid was included as control (pEmpty).

(B). Reduced ATPase activity of F434A and F439A Hel308 proteins measured in malachite green reporter assays, detecting liberation of

phosphate from ATP. Assays were in duplicate, error bars recording the standard deviation from the mean as calculated using GraphPad Prism software. Reactions contained 100 nM of protein in the presence of 5 mM ATP and 5 mM magnesium chloride, supplemented with 1 µg of ssDNA oligonucleotide to stimulate Hel308 ATPase activity.

Malachite green ATPase assays (section 2.21) measure the colour change of malachite green dye induced by binding to liberated phosphate from ATP hydrolysis. These were used to measure ssDNA stimulated ATPase activities of Hel308 F434V and F439A proteins in comparison to wild type enzyme, Figure 5.15B. MthHel308^{F439A} showed 10-50 fold reduced ATPase activity, and activity of MthHel308^{F434V} was also substantially reduced. This is interesting because neither F434 nor F439 are directly involved in helicase motifs associated with ATP hydrolysis, and do not seem to contribute to ssDNA binding, as judged by their location buried within the WHD-domain 1 interface. Genetic analysis and the ATPase assays showed most severe catalytic defects of MthHel308^{F439A}, which was therefore further analyzed for helicase activity, Figure 5.16. DNA unwinding assays on fully base paired fork substrates showed impaired helicase activity for MthHel308^{F439A}, observing almost no DNA unwinding in comparison to efficient DNA unwinding by MthHel308^{WT}.

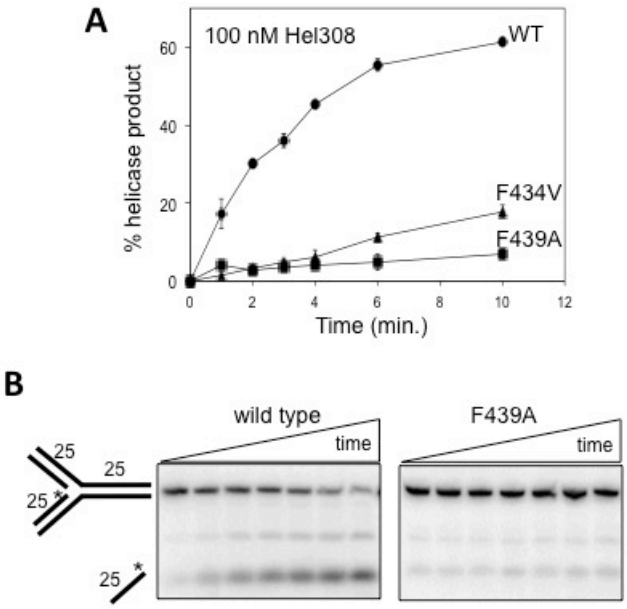


Figure 5.17: Reduced ATPase and helicase activity of MthHel308 F434A and F439A mutants.

(A) Graphical representation of mean unwinding activity of F434V and F439A mutant MthHel308 proteins to WT MthHel308 when 100 nM of protein was mixed with 6 nM of forked DNA in the presence of ATP and magnesium, measured as a function of time. The graph plots mean values for unwinding, from reactions repeated in triplicate, with error bars representing the standard deviation from the mean as calculated using GraphPad Prism software.

(B) Representative gels used for the graph are also shown.

EMSAs were used to assess MthHel308^{F439A} binding to the same fork as that used in helicase assays, Figure 5.17. MthHel308^{F439A} binding to the fork was reduced in comparison to wild type protein, but binding by MthHel308^{F439A} remained robust, in stark contrast to a total lack of helicase activity displayed by this mutant protein on the same substrate. This clarifies that lack of MthHel308^{F439A} helicase activity is unlikely to be due to defects in protein-DNA binding. This is as expected, because of the

location of F439 at the WHD- domain 1 interface making it unlikely to be available for DNA binding. The reduced catalytic effects for the two mutant proteins for ATPase activity suggest that the positioning and structural choreography between the WHD and RecA-like domain interface residues is crucial for correct functioning of Hel308 helicase.

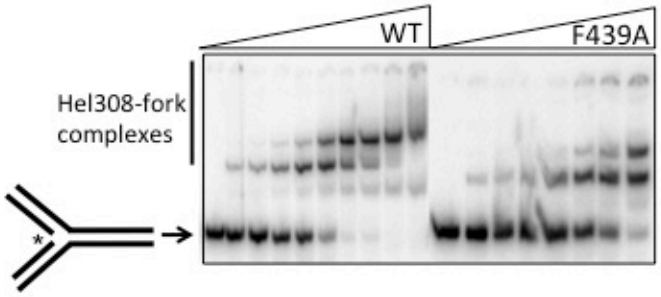


Figure 5.18: EMSA summary of forked DNA binding by wild type and F439A MthHel308. EMSAs showing MthHel308^{WT} and MthHel308^{F439A} proteins titrated from 0-100 nM against 25 nM fully base paired model replication fork. Both proteins form two stable complexes.

5.9. Summary

In this chapter we have considered the role of the MthHel308 and HsaHelQ winged helix domain (WHD) as an accessory to correct protein function. The crystal structure of archaeal Hel308 bound to a partial duplex DNA is quite clear in showing how core helicase domains 1, 2 and 4 generate a motorized translocation ratchet fulfills directional translocation along ssDNA. But the WHD is located distally to DNA, providing no evidence for it having a role in DNA processing, a potential role that was explored experimentally as presented in this chapter.

In the first part, we observed DNA binding and helicase defects associated with introducing mutations into a solvent exposed α-helix ('α-20') of full length (75 kDa) Hel308 (creating Hel308^{GAG}). The partial reduction in activity from Hel308^{GAG}, compared to wild type enzyme *in vitro*, was consistent with results from a genetic assay that also showed reduced activity from Hel308^{GAG}. Isolated 11.8 kDa WHD protein from Hel308 bound to duplex DNA, and DNA binding was abolished by the same mutations as in Hel308^{GAG}. The isolated WHD was unable to bind to ssDNA in EMSAs, suggesting that targeting of duplex DNA is a specific accessory role for the Hel308 WHD. The purified WHD fold of a human Hel308 homologue, HelQ, also bound to duplex DNA. HelQ helicase, like Hel308, translocates ssDNA with 3' to 5' polarity, to unwind a variety of DNA structures including various forks[5,187], and conservation of DNA binding by WHDs of these proteins suggest helicase unwinding of target substrates by a similar mechanism. We propose that in addition to ssDNA binding by core domains for DNA translocation, Hel308 has a second mode of DNA binding, to duplex DNA *via* the WHD. It is suggested that this function of the WHD contributes to Hel308 mechanism by promoting substrate selection, targeting duplex regions in branched substrates such as forks, which are remodeled by Hel308 helicase activity as part of the recovery process from blocked or damaged DNA replication.

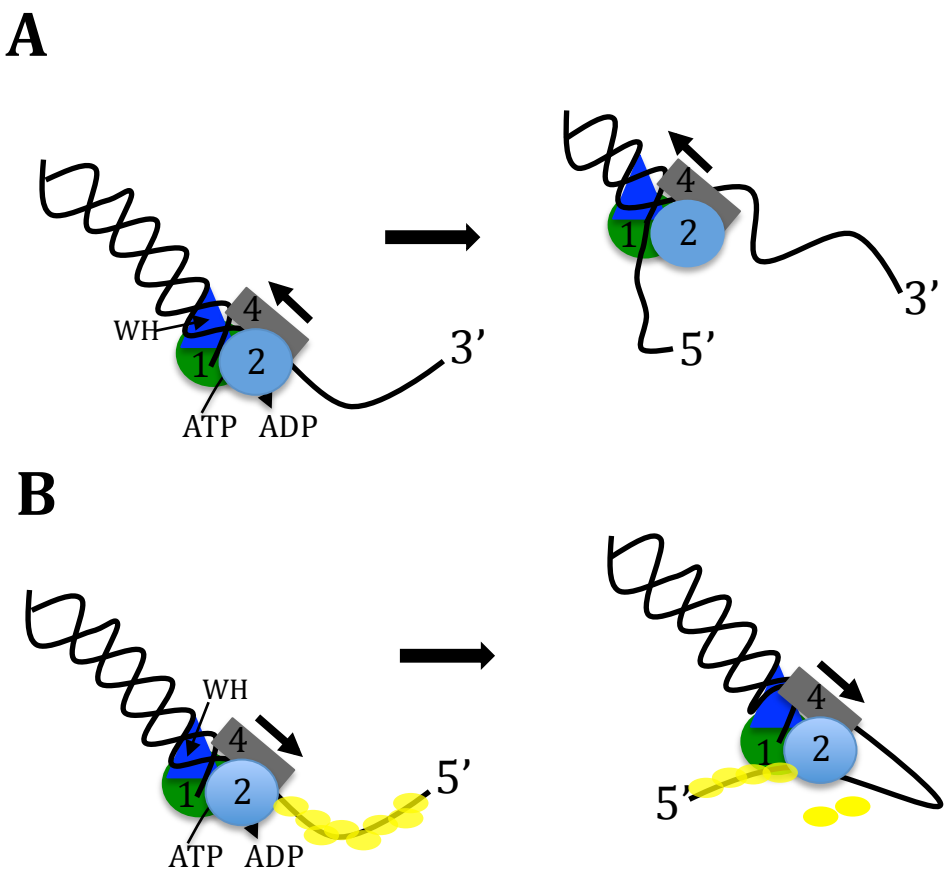


Figure 5.19. Proposal for two DNA binding and unwinding modes for Hel308 proteins. Core Hel308 domains provide ssDNA binding giving unidirectional (3' to 5') translocation. In (A) this unwinds DNA duplex *via* a β-hairpin located in domain 2. In (B) the same polarity of ssDNA translocation can "reel" ssDNA and may act to remove other bound proteins from DNA, also requiring the β-hairpin of domain 2. In both modes, the Hel308 accessory WHD (domain 3) provides duplex DNA binding close to the ssDNA-dsDNA junction, promoting helicase or "reeling" activity. Positioning of the WHD may also aid duplex DNA binding to arms of forked DNA structures.

In the second part we analyzed the interface of Hel308 WHD with RecA-like domain 1. This is a common structural theme in Ski2 helicases more generally, most obviously in Brr3 and Prp3[110,192,193,240]. This

arrangement is likely to be important for the overall integrity of protein structure, possibly explaining why our full-length *hel308* DNA constructs replacing WHD with a 22-amino acid linker, did not give protein expression. Hel308 proteins with individual substitutions in phenylalanine residues located at the interface with RecA-like domain 1, F434 and F439. showed drastically reduced ATPase and helicase activities. It is unclear how the WHD-RecA domain interface could control ATPase and helicase activity. Alteration to protein structure in the mutants is a possible explanation, but they bound to DNA, albeit with affinity reduced from wild type. It is possible that defective DNA binding to fork DNA by F434 and F439 mutants, and their loss of helicase activity, is because the WHD is no longer orientated optimally for associating with DNA substrates, or that the stability of the protein-DNA complex formed after association is sub-optimal compared to wild type. Inter-connectivity of RecA domains 1 and 2 with the WHD in the Hel308 structures may also be disturbed by mutations in the WHD – domain 1 interface. There is conformational movement between Hel308 RecA-like domain 2 and the WHD[241] facilitated by a linker located between RecA-like domain 2 and the WHD. The WHD may position RecA-like domains 1 and 2 correctly for cycles of ATP binding and hydrolysis required for ssDNA translocation. Although roles for the invariant phenylalanine residues are unclear, their mutation highlights an important role for the WHD integrated with core helicase domains, with wider relevance for Ski2 family helicases, which show this same structural topology.

Chapter 6: Summary discussion and future research

The original aims of the PhD project were to investigate the activities of human DNA helicase HelQ by (a) assessing its DNA binding and helicase activities in biochemical and biophysical assays, and (b) starting to characterize the molecular basis of known interactions between HelQ, Rad51 and Rad51 paralogues, also using the human forms of these enzymes. Significant progress was made for part (a), but part (b) remained largely intractable because of technical issues in generating and assaying human Rad51 paralogues. A significant technical breakthrough during the PhD project will sustain future research into molecular mechanisms of HelQ: reliable overexpression of human HelQ protein using insect-cell protein expression, and its rapid purification to yield milligram quantities of stable enzyme that is highly active as a helicase. This has, at the time of writing, already facilitated analysis of HelQ by FRET/TIRF and CryoEM, which were previously not possible for this enzyme, and has now led to some new biochemical characterization of HelQ helicase activity that can be compared to work on Hel308, the archaeal homologue of HelQ. Experimental results presented in Chapter 4 offer interesting first clues into the oligomeric state of HelQ, supported by molecular modeling, which might be relevant to its helicase mechanism. Data in Chapter 5 suggests that the winged helix domain of both HelQ and Hel308 may be an important accessory domain that helps to direct these helicases to forked DNA, and to control ATP-dependent helicase activity at

these and other substrates. Below is a general discussion of further research questions that may now be investigated using HelQ protein.

1. HelQ helicase mechanisms: oligomeric state and DNA pumping or translocation

Native gels in the presence of DTT as a reducing agent, and the requirement of DTT for HelQ helicase activity (Chapter 4), suggested that the enzyme might be active as a tetramer, unusually for super-family 2 helicases that are predominantly monomeric. Initial data from CryoEM may support a tetrameric state for HelQ, although refinement of those studies is currently underway. Human HelQ could be modeled well as a tetramer (Chapter 4), in agreement with the tetrameric crystallization state revealed in the atomic resolution crystal structure of the HelQ-like helicase portion of human DNA polymerase θ (PolQ)[189]. Intriguingly, tetrameric models of HelQ orientate each protein monomer with the domain 4 helicase ratchet alpha-helix facing inward toward a central pore, which is now hypothesized to accommodate at least one strand of DNA for translocation *via* an "inchworm" mechanism reminiscent of monomeric SF2 helicases. The availability of sufficient quantities of HelQ protein now make possible defined studies into helicase mechanism, and may also allow HelQ to be trialed for crystallography:

- Does incorporation of defined chemical modifications into DNA strands modulate HelQ helicase activity? Studies on translocation

properties of helicases using DNA substrates modified with backbone methyl-phosphonates, vinyl-phosphonates, phosphorothioates and abasic sites have helped to define how tracking along ssDNA may be achieved *via* electrostatic and base-stacking interactions between the helicase and DNA[242]. In the absence of a crystal structure for HsaHelQ, conservation of arginine and histidine amino acid residues in its predicted helicase ratchet with the same or similar residues in the crystal structure of archaeal Hel308 bound to DNA allow for predictions about how translocation may occur. These residues can be targeted by mutagenesis, informed by data from using chemically modified DNA substrates.

- Is a tetrameric state for HelQ required for helicase activity? By disrupting HelQ oligomers it should be now possible to test directly if this negatively affects helicase activity. Several amino acid residues at monomer interfaces of the PolQ tetramer are conserved in HelQ. Significantly, monomer-monomer interactions in PolQ are different, creating apparent dimer-pairs; mutagenesis of these residues may make it possible to isolate stable HelQ protein that is defective in tetramer formation an can be assayed for helicase activity. Native gels and denatured tetrameric HelQ can also be exploited for a similar purpose, to assay helicase activity from different oligomeric forms.

- Can biophysical analysis of helicase active HelQ give insights into the physiological role of HelQ? TIRF and smFRET methods reported in Chapter 5 were so far unsuccessful, but are being re-visited in the knowledge of the importance of DTT in HelQ helicase reactions, and that a

forked or splayed duplex is preferable for HelQ unwinding instead of a 3'-OH tailed partial duplex. For these experiments DTT will be substituted for β-mercaptoethanol, which I also showed is effective at stimulating HelQ activity, but which does not cause photo bleaching of the fluorescent dye conjugates to visualize DNA molecules. This has potential to investigate a "reeling" mechanism for HelQ in real time ('reel' time) by generating data that includes kinetic measurements. The potential power of this approach to understanding HelQ activity is two-fold; first, it may give credence to an hypothesis that the function of HelQ in human cells is to dissociate synaptic or pre-synaptic filaments that could otherwise mature into full-blown homologous recombination, which may be undesirable in the context of fixing blocked or broken DNA replication forks. The activity of bacterial PcrA in similar experiments gives a precedent for this combination of biophysical analysis with the physiological role of an individual helicase[237]. However, few studies of this type are completed for human helicases. Second, it should be possible to combine any successful outcomes of these biophysical experiments with use of chemically modified DNA substrates to unify the helicase mechanism with a potential physiological role for HelQ.

2. HelQ helicase mechanisms: roles for the winged helix domain

Chapter 5 included a comparison of winged helix domains from *M. thermautotrophicus* Hel308, which is closely related to the structurally determined Hel308 from *A. fulgidus*[191], with that predicted for human

HelQ. This concluded that the WHD forms interactions with double stranded DNA to give Hel308, and by analogy HelQ, a degree of substrate selectivity, for binding to and unwinding branched DNA molecules. This is consistent with a preference for Hel308 and HelQ unwinding DNA strands within forked substrates when assayed in isolation *in vitro*. In this model Hel308 and HelQ may behave similarly, in a binary mode of DNA binding comprising core helicase domains that encircle mainly ssDNA for a translocation ratchet, and duplex DNA binding between the WHD and elements of a branched substrate. This action of the Hel308/HelQ WHD is similar to that predicted for the WHD in bacterial helicase PriA, an enzyme that remodels blocked replication forks to re-load the replisome at sites remote from replication origins [243]. In this way PriA is thought to promote replication restart rather than fork collapse that leads to homologous recombination. In PriA, the WHD is proposed to bind transiently to duplex DNA ahead of the fork branch-point to orientate the core helicase motor for fork unwinding and replication restart[140]. This function of PriA involves interaction with PriB and PriC, loaders of the bacterial replicative helicase DnaB. A future line of enquiry for HelQ and Hel308 could be to assess whether they interact with Cdc6 and ORC loaders of MCM protein, acting therefore analogously to bacterial PriA in promoting DNA replication restart instead of recourse to homologous recombination.

It should also be possible to combine the current analysis of the importance of the HelQ/Hel308 WHD with biophysical analysis of these enzymes as part of the elucidation of a translocase “reeling” mechanism.

In this model the helicase is positioned at the ssDNA-duplex DNA branch-point with the WHD located to interact with duplex DNA. Therefore amino acid WHD mutations would be predicted to affect helicase activity measured using TIRF analysis, to provide a direct link between helicase mechanisms during translocation with the role of the WHD as an accessory domain.

3. HelQ in homologous recombination: protein interactions.

Defects in metazoan HelQ and archaeal Hel308 both give phenotypes that are associated with defective DNA repair or replication, but not much is known about exactly how their helicase functions contribute to genome stability. In one study HelQ from *C. elegans* interacts physically with Rad51 recombinase to remove it from duplex DNA[179], leading to a model that HelQ would therefore limit D-loops to prevent branch migration and homologous recombination. This activity of HelQ was ATP independent, raising the question, what is the function of the conserved helicase activity of HelQ and Hel308? Further analysis of how HelQ/Hel308 may utilize helicase dependent activities to displace other proteins from ssDNA may give some insight into its biological role. Studies using archaeal Hel308 and HelQ have noted their physical interaction with the ssDNA binding protein RPA[185,186]. One model may be that this interaction is able to guide the helicase to sites of compromised DNA replication, characterized by ectopic ssDNA which would presumably be bound by RPA, and that the helicase activity is necessary to provide

motive power to displace RPA and gain access to branched structures such as D-loops to dissipate them. This would require analysis of helicase activity by HelQ on DNA substrates that are not naked but have been pre-bound by human RPA. In conjunction, the use of Microscale Thermophoresis (NanoTemper), though not especially informative for HelQ-DNA interactions, may be a route to measuring precisely the interaction of HelQ with RPA, and determining if this is modulated by the presence of ssDNA.

A major outstanding question regarding the function of HelQ is the nature and functional relevance of its interaction with the Rad51 BCDX2 paralogue complex that has been detected by mass spectrometry after affinity pull-down studies[185]. This could not be investigated and verified using defined isolated proteins during this PhD project because of aforementioned technical difficulties in obtaining the Rad51 paralogue proteins. It is doubtful whether using human proteins is tractable in this respect. An emerging model organism for studying DNA repair by recombination, *Adeneta vaga*, is unusual in having a complete set of Rad51 paralogues (Rad51B, C, D and XRCC2/XRCC3) and has two HelQ homologues, identified by E. Bolt using the *A. vaga* genome sequence[244]. Steps to utilize these proteins are underway, with the aim to verify the protein interaction using isolated proteins, and determining how the interaction may modulate HelQ helicase activity, and influence the activity of Rad51. We speculate that potential regions of intrinsic disorder in human HelQ N- and C-termini, as identified by IUPred analysis in Chapter

4, may mediate protein-protein interactions. It may be possible to confirm the disordering of these regions using small angle X-ray scattering (SAXS)[236] or small angle neutron scattering (SANS)[245], which may be informative to plan how to study these regions in the future, or homologous regions within *A. vaga* HelQ enzymes.

Acknowledgements

I could not have undertaken this project without the help and support of numerous people both within the laboratory and outside it.

First and foremost I owe my utmost thanks and gratitude to my supervisors Ed and Panos for giving me the opportunity to carry out this project that was of particular interest to me. Especially their guidance, helpful discussion and support for both technical aspects and other has been indispensable over the last few years, and their encouragement to pursue a career in science (at least for a bit longer!). Thanks to the University of Nottingham School of Chemistry and School of Life Sciences for financial support.

Many thanks to those who helped me with technical aspects of my project, namely all at the Oxford Protein Production Facility UK (OPPF-UK) for high throughput cloning and protein expression, Christin Pelzel (Vinkimeyer lab) for teaching me insect cell culture and Lucio Manzi (Oldham lab) for protein mass spectrometry analysis. In addition thanks to Dr Dennis Ptchelkine for preliminary Cryo-EM analysis of HelQ (good luck in the future!) and to Dr Carlos Penedo and Cibran Perez Gonzalez at the Single Molecule Biophysics laboratory at St Andrews for giving me the opportunity to try a new interesting technique that I hope to use again in the future.

Also, thanks to lab members past and present in CBS A20, particularly Vasileios Paschalis, Matt Green, Manthos Pithoulias, Faadil Fawzy and Eleyna Martin of the Soultanas group and Dave Harvey, Ali

Waleed, Alex Slater, Juliet Morgan for making A20 a friendly place to work in. Also, thanks to lab members past and present in QMC D53/55, in particular Emily Walker, Tom Killelea, Ryan Buckley, George Blundell-Hunter and Shivham Bhatt. Thanks for putting up with me for the last three years, and good luck!

Thanks to those outside the lab, especially Katie Belfield, Lucy Simmonds and Dan Kent. Finally, thanks to my Mum, Denise, and sister, Becky and family for their support and understanding.

References

1. Forterre, P. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 3669–3674 (2006).
2. Krawinkel, U., Zoebelein, G., Brüggemann, M., Radbruch & Rajewsky, K. Recombination between antibody heavy chain variable-region genes: evidence for gene conversion. *Proc. Natl. Acad. Sci. U. S. A.* **80,** 4997–5001 (1983).
3. Hunter, N. Meiotic Recombination : The Essence of Heredity. *Cold Spring Harb. Perspect. Biol.* 1–36 (2015).
4. Marini, F. & Wood, R. D. A human DNA helicase homologous to the DNA cross-link sensitivity protein Mus308. *J. Biol. Chem.* **277,** 8716–23 (2002).
5. Guy, C. P. & Bolt, E. L. Archaeal Hel308 helicase targets replication forks in vivo and in vitro and unwinds lagging strands. *Nucleic Acids Res.* **33,** 3678–3690 (2005).
6. Watson, J. C. & Crick, F. H. Molecular structure of nucleic acids. *Nature* **171,** 737–738 (1953).
7. Drew, H. R. *et al.* Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci.* **78,** 2179–2183 (1981).
8. Travers, A; Muskhelishvili, G. DNA structure and function. *FEBS J.* **282,** 2279–95 (2015).
9. Woese, C. R., Sogin, M. L. & Sutton, L. A. Procaryote phylogeny. *J. Mol. Evol.* **3,** 293–299 (1974).
10. Huet, J., Schnabell, R. & Sentenac, A. Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. *EMBO J.* **2,** 1291–1294 (1983).
11. Langer, D., Hain, J., Thuriauxt, P. & Zillig, W. Transcription in Archaea : Similarity to that in Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 5768–5772 (1995).
12. Woese, C. R., Kandlert, O. & Wheelis, M. L. Towards a natural system of organisms : Proposal for the domains. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 4576–4579 (1990).
13. Coker, J. A. Extremophiles and biotechnology: current uses and prospects. *F1000 Research* **5,** 1–7 (2016).
14. Bult, C. J. *et al.* Complete Genome Sequence of the Methanogenic Archaeon , *Methanococcus jannaschii. Science* **273,** 1058–1073 (1996).
15. Graham, D. E., Overbeek, R., Olsen, G. J. & Woese, C. R. An archaeal genomic signature. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 3304–3308 (2000).
16. Grabowski, B. & Kelman, Z. Archaeal DNA Replication : Eukaryal Proteins in a Bacterial Context. *Annu. Rev. Microbiol.* **57,** 487–516 (2003).
17. Yutin, N. & Koonin, E. V. Archaeal origin of tubulin. *Biol. Direct* **7,** 10

(2012).
18. 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).
19. Spang, A. *et. al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521,** 173–179 (2015).
20. López-García, P. & Moreira, D. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem. Sci.* **24,** 88–93 (1999).
21. Azam, T. & Ishihama, A. Twelve species of the nucleoid-associated protein 481 from Escherichia coli. *J. Biol. Chem.* **274,** 33105–33113 (1999).
22. Macvanin, M. & Adhya, S. Architectural organization in E. coli nucleoid. *Biochim. Biophys. Acta* **1819,** 830–835 (2012).
23. Ho Lee, J. *et. al.* TonEBP suppresses adipogenesis and insulin sensitivity by blocking epigenetic transition of PPARγ2. *Sci. Rep.* **5,** 10937 (2015).
24. Jelinska, C. *et. al.* Obligate heterodimerization of the archaeal Alba2 protein with Alba1 provides a mechanism for control of DNA Packaging. *Structure* **13,** 963–971 (2005).
25. Sandman, K. & Reeve, J. N. Archaeal histones and the origin of the histone fold. *Curr. Opin. Microbiol.* **9,** 520–525 (2006).
26. Lawrence, M., Daujat, S. & Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends Genet.* **32,** 42–56 (2016).
27. Burma, S., Chen, B. P., Murphy, M., Kurimasa, A. & Chen, D. J. ATM Phosphorylates Histone H2AX in Response to DNA Double-strand Breaks. *J. Biol. Chem.* **276,** 42462–42467 (2001).
28. Arents, G., Burlingame, R. W., Wangt, B.-C., Love, W. E. & Moudrianakis, E. N. The nucleosomal core histone octamer at 3.1 A resolution: A tripartite protein assembly and a left-handed superhelix. *Biochemistry* **88,** 10148–10152 (1991).
29. Thomas, J. O. & Kornberg, R. D. An octamer of histones in chromatin and free in solution. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 2626–30 (1975).
30. Li, H. J. A model for chromatin structure. *Nucleic Acids Res.* **2,** 1275–1290 (1975).
31. Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. U. S. A.* **73,** 1897–901 (1976).
32. House, N. C. M., Koch, M. R. & Freudenreich, C. H. Chromatin modifications and DNA repair: Beyond double-strand breaks. *Front. Genet.* **5,** 1–18 (2014).
33. Bartholomew, B. Regulating the chromatin landscape: structural and mechanistic perspectives. *Annu. Rev. Biochem.* **83,** 671–696 (2014).
34. Guy, C. P. *et al.* Rep Provides a Second Motor at the Replisome to Promote Duplication of Protein-Bound DNA. *Mol. Cell* **36,** 654–666 (2009).

35. Maine, G. T., Surosky, R. T. & Tye, B. K. Isolation and characterization of the centromere from chromosome V (CEN5) of *Saccharomyces cerevisiae*. *Mol Cell Biol* **4,** 86–91 (1984).
36. Yeeles, J. T. P., Deegan, T. D., Janska, A., Early, A. & Diffley, J. F. X. Regulated eukaryotic DNA replication origin firing with purified proteins. *Nature* **519,** 431–435 (2015).
37. Yeeles, J. T. P., Janska, A., Early, A. & Diffley, J. F. X. How the Eukaryotic Replisome Achieves Rapid and Efficient DNA Replication. *Mol. Cell* **65,** 105–116 (2016).
38. Macneill, S. A. The haloarchaeal chromosome replication machinery. *Biochem. Soc. Trans.* **37,** 108–113 (2009).
39. O'Donnell, M. *et. al.* Principles and Concepts of DNA Replication in Bacteria, Archaea and Eukarya. *Cold Spring Harb. Perspect. Biol.* 1–14 (2013).
40. Meselson, M. & Stahl, F. W. The Replication of DNA in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **44,** 671–682 (1958).
41. Norais, C. *et. al.* Genetic and Physical Mapping of DNA Replication Origins in *Haloferax volcanii*. *PLOS Genet.* **3,** e77 (2007).
42. Robinson, N. P. *et. al.* Identification of Two Origins of Replication in the Single Chromosome of the Archaeon *Sulfolobus solfataricus*. *Cell* **116,** 25–38 (2004).
43. Duggin, I. G., McCallum, S. a & Bell, S. D. Chromosome replication dynamics in the archaeon *Sulfolobus acidocaldarius*. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 16737–42 (2008).
44. Bochman, M. L. & Schwacha, A. The Mcm Complex: Unwinding the Mechanism of a Replicative Helicase. *Microbiol. Mol. Biol. Rev.* **73,** 652–683 (2009).
45. Chong, J. P. J., 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).
46. Evrin, C. *et. al.* A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 20240–5 (2009).
47. 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).
48. Krastanova, I. *et. al.* Structural and functional insights into the DNA replication factor Cdc45 reveal an evolutionary relationship to the DHH family of phosphoesterases. *J. Biol. Chem.* **287,** 4121–4128 (2012).
49. Soultanas, P. Loading mechanisms of ring helicases at replication origins. *Mol. Microbiol.* **84,** 6–16 (2012).
50. Chandler, M., Bird, R. E. & Caro, L. The replication time of the *Escherichia coli* K12 chromosome as a function of cell doubling time. *J. Mol. Biol.* **94,** 127–132 (1975).
51. Frick, D. N. & Richardson, C. C. DNA Primases. *Annu. Rev. Biochem.*

70, 39–80 (2001).
52. Liu, B., Eliason, W. K. & Steitz, T. A. Structure of a helicase-helicase loader complex reveals insights into the mechanism of bacterial primosome assembly. *Nat. Commun.* **4,** 2495 (2013).
53. Bauer, R. J., Graham, B. W. & Trakselis, M. A. Novel interaction of the bacterial-like DnaG primase with the MCM helicase in archaea. *J. Mol. Biol.* **425,** 1259–1273 (2013).
54. Zuo, Z., Rodgers, C. J., Mikheikin, A. L. & Trakselis, M. A. Characterization of a Functional DnaG-Type Primase in Archaea: Implications for a Dual-Primase System. *J. Mol. Biol.* **397,** 664–676 (2010).
55. Liu, B. *et. al.* A primase subunit essential for efficient primer synthesis by an archaeal eukaryotic-type primase. *Nat. Commun.* **6,** 1–11 (2015).
56. Steitz, T. a & Steitz, J. a. A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 6498–6502 (1993).
57. McHenry, C. S. Bacterial replicases and related polymerases. *Curr. Opin. Chem. Biol.* **15,** 587–594 (2011).
58. Johansson, E. & Dixon, N. Replicative DNA Polymerases. *Cold Spring Harb. Perspect. Biol.* **5,** 1–14 (2013).
59. Čuboňová, L. *et. al.* Archaeal DNA polymerase D but not DNA polymerase B is required for genome replication in *Thermococcus kodakarensis*. *J. Bacteriol.* **195,** 2322–2328 (2013).
60. Henneke, G., Flament, D. & Hu, U. The Hyperthermophilic Euryarchaeota *Pyrococcus abyssi* Likely Requires the Two DNA Polymerases D and B for DNA Replication. *J. Mol. Biol.* **350,** 53–64 (2005).
61. Bravo, R., Frank, R., Blundell, P. A. & Macdonald-Bravo, H. Cyclin/PCNA is the auxiliary protein of DNA polymerase-δ. *Nature* **326,** 515–517 (1987).
62. Wagner, J., Fujii, S., Gruz, P., Nohmi, T. & Fuchs, R. P. The beta clamp targets DNA polymerase IV to DNA and strongly increases its processivity. *EMBO Rep.* **1,** 484–488 (2000).
63. Yao, N. Y. & O'Donnell, M. SnapShot: The Replisome. *Cell* **141,** 1088–1088 (2010).
64. Wobbe, C. R. *et. al.* Replication of simian virus 40 origin-containing DNA in vitro with purified proteins. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 1834–1838 (1987).
65. Wold, M. S. & Kelly, T. Purification and characterization of replication protein A, a cellular protein required for in vitro replication of simian virus 40 DNA. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 2523–7 (1988).
66. Meyer, R. R., Glassberg, J. & Kornberg, A. An *Escherichia coli* mutant defective in single-strand binding protein is defective in DNA replication. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 1702–5 (1979).
67. Bates, A. D. & Maxwell, A. DNA topology: topoisomerases keep it simple. *Curr. Biol.* **7,** R778–R781 (1997).

68. Murzin, A. G. OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO J.* **12,** 861–867 (1993).
69. Kelly, T. J., Simancek, P. & Brush, G. S. Identification and characterization of a single-stranded DNA-binding protein from the archaeon *Methanococcus jannaschii*. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14634–9 (1998).
70. Gupta, M. K. *et. al.* Protein-DNA complexes are the primary sources of replication fork pausing in *Escherichia coli*. *Proc Natl Acad Sci U S A* **110,** 7252–7257 (2013).
71. Brüning, J., Howard, J. L. & Mcglynn, P. Accessory Replicative Helicases and the Replication of Protein-Bound DNA. *J. Mol. Biol.* **426,** 3917–3928 (2014).
72. Florés, M. J., Sanchez, N. & Michel, B. A fork-clearing role for UvrD. *Mol. Microbiol.* **57,** 1664–1675 (2005).
73. Boubakri, H., Langlois De Septenville, A., Viguera, E. & Né Dicte Michel, B. The helicases DinG, Rep and UvrD cooperate to promote replication across transcription units in vivo. *EMBO J.* **29,** 145–157 (2009).
74. Lloyd, R. G. & Rudolph, C. J. 25 years on and no end in sight: a perspective on the role of RecG protein. *Curr. Genet.* **62,** 827–840 (2016).
75. Bessler, J. B., Torres, J. Z., Zakian, V. A., Torres, J. Z. & Zakian, V. A. The Pif1p subfamily of helicases : region- specific DNA helicases? *Trends Cell Biol.* **11,** 60–65 (2001).
76. Ivessa, A. S. *et al.* The *Saccharomyces cerevisiae* Helicase Rrm3p Facilitates Replication Past Nonhistone Protein-DNA Complexes. *Mol. Cell* **12,** 1525–1536 (2003).
77. Torres, J. Z., Schnakenberg, S. L. & Zakian, V. A. *Saccharomyces cerevisiae* Rrm3p DNA Helicase Promotes Genome Integrity by Preventing Replication Fork Stalling : Viability of rrm3 Cells Requires the Intra-S-Phase Checkpoint and Fork Restart Activities. *Mol. Cell. Biol.* **24,** 3198–3212 (2004).
78. Ivessa, A. S., Zhou, J., Schulz, V. P., Monson, E. K. & Zakian, V. A. Saccharomyces Rrm3p , a 5' to 3' DNA helicase that promotes replication fork progression through telomeric and subtelomeric DNA. *Genes Dev.* **16,** 1383–1396 (2002).
79. Boulé, J. B. & Zakian, V. a. Roles of Pif1-like helicases in the maintenance of genomic stability. *Nucleic Acids Res.* **34,** 4147–4153 (2006).
80. Deans, A. J. & West, S. C. DNA interstrand crosslink repair and cancer. *Nat. Rev. Cancer* **11,** 467–80 (2011).
81. Tubbs, A. N. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168,** 644–656 (2017).
82. Hoeijmakers, J. H. J. Genome Maintenance Mechanisms for Preventing Cancer. *Nature* **411,** 366–374 (2001).
83. Brady, C. A. *et. al.* Distinct p53 transcriptional programs dictate

acute DNA-damage responses and tumor suppression. *Cell* **145,** 571–583 (2011).

84. Lindahl, T., Karran, P. & Wood, R. D. DNA excision repair pathways. *Curr. Opin. Genet. Dev.* **7,** 158–169 (1997).
85. A Aguilera, T. G.-M. R Loops : From Transcription Byproducts to Threats to Genome Stability. *Mol. Cell* **46,** 115–124 (2012).
86. Ivancic-Bace, I., Howard, J. A. L. & Bolt, E. L. Tuning in to Interference : R-Loops and Cascade Complexes in CRISPR Immunity. *J. Mol. Biol.* **422,** 607–616 (2012).
87. Santos-Pereira, J. M. & Aguilera, A. R loops: new modulators of genome dynamics and function. *Nat. Rev. Genet.* **16,** 583–597 (2015).
88. Andrea, A. D. D., Grompe, M., Jackson, S. & Road, P. The Fanconi Anaemia / BRCA Pathway. *Nat. Rev. Cancer* **3,** 23–34 (2003).
89. Aggarwal, M. *et al.* Werner Syndrome Helicase Has a Critical Role in DNA Damage Responses in the Absence of a Functional Fanconi Anemia Pathway. *Cancer Res.* 5497–5507 (2013).
90. Alter, B. P. Fanconi's anemia and malignancies. *Am. J. Hematol.* **53,** 99–110 (1996).
91. Smerdon, M. J. DNA repair and the role of chromatin structure. *Curr. Opin. Cell Biol.* **3,** 422–428 (1991).
92. Cortez, D. Unwind and slow down: Checkpoint activation by helicase and polymerase uncoupling. *Genes Dev.* **19,** 1007–1012 (2005).
93. Fanning, E., Klimovich, V. & Nager, A. R. A dynamic model for replication protein A (RPA) function in DNA processing pathways. *Nucleic Acids Res.* **34,** 4126–4137 (2006).
94. Byun, T. S., Pacek, M., Yee, M. C., Walter, J. C. & Cimprich, K. A. Functional uncoupling of MCM helicase and DNA polymerase activities activates the ATR-dependent checkpoint. *Genes Dev.* **19,** 1040–1052 (2005).
95. Ball, H., Meyers, J. & Cortez, D. ATRIP Binding to Replication Protein A-Single Stranded DNA Promotes ATR-ATRIP Localization but Is Dispensable for Chk1 Phosphorylation. *Mol. Biol. Cell* **16,** 2372–2381 (2005).
96. Trenz, K., Smith, E., Smith, S. & Costanzo, V. ATM and ATR promote Mre11 dependent restart of collapsed replication forks and prevent accumulation of DNA breaks. *EMBO J.* **25,** 1764–74 (2006).
97. Bass, T. E. *et al.* ETAA1 acts at stalled replication forks to maintain genome integrity. *Nat. Cell Biol.* **18,** (2016).
98. Price, B. D. & D'Andrea, A. D. Chromatin remodeling at DNA double-strand breaks. *Cell* **152,** 1344–1354 (2013).
99. Cimprich, K. A. & Cortez, D. ATR: an essential regulator of genome integrity. *Nat. Rev. Mol. Cell Biol.* **9,** 616–27 (2008).
100. Shiloh, Y. & Ziv, Y. The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. *Nat. Rev. Mol. Cell Biol.* **14,** 197–210 (2013).

101. Singleton, M. R., Dillingham, M. S. & Wigley, D. B. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.* **76,** 23–50 (2007).
102. Brosh, R. M. DNA helicases involved in DNA repair and their roles in cancer. *Nat. Rev. Cancer* **13,** 542–558 (2013).
103. Suhasini, A. N. & Brosh, R. M. Mechanistic and biological aspects of helicase action on damaged DNA. *Cell Cycle* **4101,** 2317–2329 (2017).
104. Bernstein, K., Gangloff, S. & Rothstein, R. The RecQ DNA Helicases in DNA Repair. *Annu. Rev. Genet.* **44,** 393–417 (2014).
105. Wu, L. & Hickson, I. D. DNA helicases required for homologous recombination and repair of damaged replication forks. *Annu. Rev. Genet.* **40,** 279–306 (2006).
106. Petrova, V. *et al.* Active displacement of RecA filaments by UvrD translocase activity. *Nucleic Acids Res.* **43,** 4133–4149 (2015).
107. Qiu, Y. *et al.* Srs2 prevents Rad51 filament formation by repetitive motion on DNA. *Nat. Commun.* **4,** 2281 (2013).
108. Prakash, R. *et al.* Yeast Mphl helicase dissociates Rad51-made D-loops: Implications for crossover control in mitotic recombination. *Genes Dev.* **23,** 67–79 (2009).
109. Ivessa, A. S., Zhou, J., Zakian, V. A., Teng, H. S. & Orf, T. R. R. M. The Saccharomyces Pif1p DNA Helicase and the Highly Related Rrm3p Have Opposite Effects on Replication Fork Progression in Ribosomal DNA. *Cell* **100,** 479–489 (2000).
110. Johnson, S. J. & Jackson, R. N. Ski2-like RNA helicase structures Common themes and complex assemblies. *RNA Biol.* **10,** 33–43 (2013).
111. Kim, J. L. *et al.* Hepatitis C virus NS3 RNA helicase domain with a bound oligonucleotide: the crystal structure provides insights into the mode of unwinding. *Structure* **6,** 89–100 (1998).
112. Kisker, C., Kuper, J. & Houten, B. Van. Prokaryotic Nucleotide Excision Repair. *Cold Spring Harb. Perspect. Biol.* **5,** 1–18 (2013).
113. Jaciuk, M., Nowak, E., Skowronek, K., Tańska, A. & Nowotny, M. Structure of UvrA nucleotide excision repair protein in complex with modified DNA. *Nat. Struct. Mol. Biol.* **18,** 191–7 (2011).
114. Truglio, J. J., Croteau, D. L., Houten, B. Van & Kisker, C. Prokaryotic Nucleotide Excision Repair: The UvrABC System. *Chem. Rev.* **106,** 233–252 (2006).
115. Sancar, A. Excision Repair in Mammalian Cella. *J. Biol. Chem.* **270,** 15915–15918 (1995).
116. Hoeijmakers, J. A. N. H. J. Nucleotide excision repair II : from yeast to mammals. *Trends Genet.* **9,** (1993).
117. Shuck, S. C., Short, E. a & Turchi, J. J. Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. *Cell Res.* **18,** 64–72 (2008).
118. Räschle, M. *et al.* Mechanism of replication-coupled DNA interstrand crosslink repair. *Cell* **134,** 969–80 (2008).

119. Cole, R. S. Repair of interstrand cross-links in DNA induced by psoralen plus light. *Yale J. Biol. Med.* **46,** 492 (1973).
120. Andreassen, P. & Ren, K. Fanconi Anaemia Proteins, DNA Interstrand Crosslink Repair Pathways, and Cancer Therapy. *Curr. Cancer Crug Targets* **9,** 101–117 (2009).
121. Gari, K., Décaillet, C., Delannoy, M., Wu, L. & Constantinou, A. Remodeling of DNA replication structures by the branch point translocase FANCM. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 16107–12 (2008).
122. Gari, K., Décaillet, C., Stasiak, A. Z., Stasiak, A. & Constantinou, A. The Fanconi anemia protein FANCM can promote branch migration of Holliday junctions and replication forks. *Mol. Cell* **29,** 141–8 (2008).
123. Robertson, A. B., Klungland, A., Rognes, T. & Leiros, I. Base excision repair: The long and short of it. *Cell. Mol. Life Sci.* **66,** 981–993 (2009).
124. T. Lindhal. DNA Glycosylases, Endonucleases for Apurinic/ Apyrimidinic Sites and Base Excision Repair. *Prog. Nucl. Acid. Res. Mol. Biol.* 135–192 (1970).
125. Maynard, S., Schurman, S. H., Harboe, C., Souza-pinto, N. C. De & Ã, V. A. B. Base excision repair of oxidative DNA damage and association with cancer and aging. *Carcinogenesis* **30,** 2–10 (2009).
126. Andersson, D. I., Koskiniemi, S. & Hughes, D. Micro Review Biological roles of translesion synthesis DNA polymerases in eubacteria. *Mol. Microbiol.* **77,** 540–548 (2010).
127. Jansen, J. G., Tsaalbi-shtylik, A. & Wind, N. De. Roles of mutagenic translesion synthesis in mammalian genome stability, health and disease. *DNA Repair (Amst).* **29,** 56–64 (2015).
128. Waters, L. S. *et al.* Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiol. Mol. Biol. Rev.* **73,** 134–154 (2009).
129. Ling, H., Boudsocq, F., Woodgate, R. & Yang, W. Crystal Structure of a Y-Family DNA Polymerase in Action. *Cell* **107,** 91–102 (2001).
130. Boudsocq, F. *et. al.* Investigating the role of the little finger domain of Y-family DNA polymerases in low fidelity synthesis and translesion replication. *J. Biol. Chem.* **279,** 32932–32940 (2004).
131. Matsuda, T., Bebenek, K., Masutani, C., Hanaoka, F. & Kunkel, T. A. Low fidelity DNA synthesis by human DNA polymerase-eta. *Nature* **404,** 1011–1013 (2000).
132. Bridges, B. A. Error-prone DNA repair and translesion synthesis : Focus on the replication fork. *DNA Repair (Amst).* **4,** 618–634 (2005).
133. Cox, M. M. *et al.* The importance of repairing stalled replication forks. *Nature* **404,** 37–41 (2000).
134. Rudolph, C. J., Dhillon, P., Moore, T. & Lloyd, R. G. Avoiding and resolving conflicts between DNA replication and transcription. *DNA Repair (Amst).* **6,** 981–993 (2007).

135. Rudolph, C. J., Upton, A. L., Stockum, A., Nieduszynski, C. A. & Lloyd, R. G. Avoiding chromosome pathology when replication forks collide. *Nature* **500,** 608–611 (2013).
136. Hiasa, H. & Marians, K. J. Tus prevents overreplication of oriC plasmid DNA. *J. Biol. Chem.* **269,** 26959–26968 (1994).
137. Pandey, M. *et al.* Two mechanisms coordinate replication termination by the Escherichia coli Tus-Ter complex. *Nucleic Acids Res.* **43,** 5924–5935 (2015).
138. Heller, R. C. & Marians, K. J. Unwinding of the Nascent Lagging Strand by Rep and PriA Enables the Direct Restart of Stalled Replication Forks. *J. Biol. Chem.* **280,** 34143–34151 (2005).
139. Sasaki, K. *et. al.* Structural basis of the 3' end recognition of a leading strand in stalled replication forks by PriA. *EMBO J.* **26,** 2584–2593 (2007).
140. Bhattacharyya, B. *et al.* Structural mechanisms of PriA-mediated DNA replication restart. *Proc Natl Acad Sci U S A* **111,** 1373–1378 (2014).
141. Mahdi, A. A., Buckman, C., Harris, L. & Lloyd, R. G. Rep and PriA helicase activities prevent RecA from provoking unnecessary recombination during replication fork repair. *Genes Dev.* **20,** 2135–2147 (2006).
142. Xu, L. & Marians, K. J. PriA Mediates DNA Replication Pathway Choice at Recombination Intermediates. *Mol. Cell* **11,** 817–826 (2003).
143. Northall, S. J., Ivancic-Bace, I., Soultanas, P. & Bolt, E. L. Remodeling and control of homologous recombination by DNA helicases and translocases that target recombinases and synapsis. *Genes.* **7,** 1–12 (2016).
144. Somyajit, K., Saxena, S., Babu, S., Mishra, A. & Nagaraju, G. Mammalian RAD51 paralogs protect nascent DNA at stalled forks and mediate replication restart. *Nucleic Acids Res.* **43,** (2015).
145. Badie, S. *et. al.* RAD51C facilitates checkpoint signaling by promoting CHK2 phosphorylation. *J. Cell Biol.* **185,** 587–600 (2009).
146. Fugger, K. & West, S. C. Keeping homologous recombination in check. *Cell Res.* **26,** 397–398 (2016).
147. Benson, F. E., Stasiak, A. & West, S. C. Purification and characterisation of the human Rad51 protein, an analog of E. coli RecA. *EMBO J.* **13,** 5764–5771 (1994).
148. Conway, A. B. *et al.* Crystal structure of a Rad51 filament. *Nat. Struct. Mol. Biol.* **11,** 791–6 (2004).
149. Bishop, D. K., Park, D., Xu, L. & Kleckner, N. DMC1: A meiosis-specific yeast homolog of E. coli RecA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell* **69,** 439–456 (2016).
150. Liu, J., Doty, T., Gibson, B. & Heyer, W.-D. Human BRCA2 protein promotes RAD51 filament formation on RPA-covered single-stranded DNA. *Nat. Struct. Mol. Biol.* **17,** 1260–1262 (2010).

151. Tombline, G. & Fishel, R. Biochemical characterization of the human RAD51 protein. I. ATP hydrolysis. *J. Biol. Chem.* **277,** 14417–14425 (2002).
152. Tombline, G., Shim, K. S. & Fishel, R. Biochemical characterization of the human RAD51 protein. II. Adenosine nucleotide binding and competition. *J. Biol. Chem.* **277,** 14426–14433 (2002).
153. Dunn, K., Chrysogelos, S. & Griffith, J. Electron microscopic visualization of RecA-DNA filaments: Evidence for a cyclic extension of duplex DNA. *Cell* **28,** 757–765 (1982).
154. Bell, J. C. & Kowalczykowski, S. C. RecA: Regulation and Mechanism of a Molecular Search Engine. *Trends Biochem. Sci.* **41,** 491–507 (2016).
155. VanLoock, M. S. *et. al.* Complexes of RecA with LexA and RecX differentiate between active and inactive RecA nucleoprotein filaments. *J. Mol. Biol.* **333,** 345–354 (2003).
156. Liu, J. *et. al.* Rad51 paralogues Rad55–Rad57 balance the antirecombinase Srs2 in Rad51 filament formation. *Nature* **479,** 245–248 (2011).
157. Amunugama, R., Groden, J. & Fishel, R. The HsRAD51B-HsRAD51C stabilizes the HsRAD51 nucleoprotein filament. *DNA Repair (Amst).* **12,** 723–32 (2013).
158. Taylor, M. R. G. *et. al.* Rad51 Paralogs Remodel Pre-synaptic Rad51 Filaments to Stimulate Homologous Recombination. *Cell* **162,** 271–286 (2015).
159. Qi, Z. *et. al.* DNA sequence alignment by microhomology sampling during homologous recombination. *Cell* **160,** 856–869 (2015).
160. West, S. *et. al.* Resolution of Recombination Intermediates : Mechanisms and Regulation. *Cold spring Harb. Symp. Quant. Biol.* **80,** 103–109 (2015).
161. Ciccia, A., McDonald, N. & West, S. C. Structural and functional relationships of the XPF/MUS81 family of proteins. *Annu. Rev. Biochem.* **77,** 259–87 (2008).
162. Blanco, M. G. & Matos, J. Hold your horSSEs: controlling structure-selective endonucleases MUS81 and Yen1/GEN1. *Front. Genet.* **6,** 253 (2015).
163. Bocquet, N. *et. al.* Structural and mechanistic insight into Holliday-junction dissolution by Topoisomerase IIIα and RMI1. *Nat. Struct. Mol. Biol.* **21,** 261–8 (2014).
164. Hiom, K. Homologous Recombination. *Curr. Biol.* **10,** 359–361 (2000).
165. Malkova, A. & Ira, G. Break-induced replication: Functions and molecular mechanism. *Curr. Opin. Genet. Dev.* **23,** 271–279 (2013).
166. Anand, R. P., Lovett, S. T. & Haber, J. E. Break-Induced DNA Replication. (2016).
167. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5,** (2009).

168. Veaute, X. *et. al.* The Srs2 helicase prevents recombination by disrupting Rad51 nucleoprotein filaments. *Nature* **423,** 309–312 (2003).
169. Krejci, L., Altmannova, V., Spirek, M. & Zhao, X. Homologous recombination and its regulation. *Nucleic Acids Res.* **40,** 5795–818 (2012).
170. Meddows, T. R., Savory, A. P. & Lloyd, R. G. RecG helicase promotes DNA double-strand break repair. *Mol. Microbiol.* **52,** 119–132 (2004).
171. Mosig, G. Recombination and recombination-dependent DNA replication in bacteriophage T4. *Annu. Rev. Genet.* **32,** 379–413 (1998).
172. Masai, H., Asai, T., Kubota, Y., Arai, K. & Kogoma, T. *Escherichia coli* PriA protein is essential for inducible and constitutive stable DNA replication. *EMBO J.* **13,** 5338–5345 (1994).
173. Kogoma, T. Recombination by replication. *Cell* **85,** 625–627 (1996).
174. Hong, X., Cadwell, G. W. & Kogoma, T. Activation of stable DNA replication in rapidly growing *Escherichia coli* at the time of entry to stationary phase. *Mol. Microbiol.* **21,** 953–961 (1996).
175. Kogoma, T. Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol. Mol. Biol. Rev.* **61,** 212–238 (1997).
176. Antony, E. *et al.* Srs2 Disassembles Rad51 Filaments by a Protein-Protein Interaction Triggering ATP Turnover and Dissociation of Rad51 from DNA. *Mol. Cell* **35,** 105–115 (2009).
177. Barber, L. J. *et al.* RTEL1 Maintains Genomic Stability by Suppressing Homologous Recombination. *Cell* **135,** 261–271 (2008).
178. Schwendener, S. *et. al.* Physical interaction of RECQ5 helicase with RAD51 facilitates its anti-recombinase activity. *J. Biol. Chem.* **285,** 15739–15745 (2010).
179. Ward, J. D. *et. al.* Overlapping mechanisms promote postsynaptic RAD-51 filament disassembly during meiotic double-strand break repair. *Mol. Cell* **37,** 259–72 (2010).
180. Boyd, J. B., Sakaguchi, K. & Harris, P. V. *mus308* Mutants of Drosophila Exhibit Hypersensitivity to DNA crosslinking agents and are defective in a deoxyribonuclease. *Genetics* **125,** 813–819 (1990).
181. Sharief, F. S., Vojta, P. J., Ropp, P. a & Copeland, W. C. Cloning and chromosomal mapping of the human DNA polymerase theta (POLQ), the eighth human DNA polymerase. *Genomics* **59,** 90–6 (1999).
182. Seki, M., Marini, F. & Wood, R. D. POLQ (Pol), a DNA polymerase and DNA-dependent ATPase in human cells. *Nucleic Acids Res.* **31,** 6117–6126 (2003).
183. Marini, F., Kim, N., Schuffert, A. & Wood, R. D. POLN, a nuclear PolA family DNA polymerase homologous to the DNA cross-link sensitivity protein Mus308. *J. Biol. Chem.* **278,** 32014–9 (2003).

184. Woodman, I. L. & Bolt, E. L. Molecular biology of Hel308 helicase in archaea. *Biochem. Soc. Trans.* **37,** 74–8 (2009).
185. Adelman, C. *et. al.* HELQ promotes RAD51 paralogue-dependent repair to avert germ cell loss and tumorigenesis. *Nature* **502,** 381–4 (2013).
186. Woodman, I. L., Brammer, K. & Bolt, E. L. Physical interaction between archaeal DNA repair helicase Hel308 and Replication Protein A (RPA). *DNA Repair (Amst).* **10,** 306–13 (2011).
187. Tafel, A. a, Wu, L. & McHugh, P. J. Human HEL308 localizes to damaged replication forks and unwinds lagging strand structures. *J. Biol. Chem.* **286,** 15832–40 (2011).
188. Bolt, E. L. & Guy, C. P. Homologous recombination in Archaea: new Holliday junction helicases. *Biochem. Soc. Trans.* **31,** 703–5 (2003).
189. Newman, J. A., Cooper, C. D. O., Aitkenhead, H. & Gileadi, O. Structure of the Helicase Domain of DNA Polymerase Theta Reveals a Possible Role in the Microhomology-Mediated End-Joining Pathway. *Structure* **23,** 2319–2330 (2015).
190. Richards, J. D. *et al.* Structure of the DNA repair helicase Hel308 reveals DNA binding and autoinhibitory domains. *J. Biol. Chem.* **283,** 5118–26 (2008).
191. Büttner, K., Nehring, S. & Hopfner, K.-P. Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nat. Struct. Mol. Biol.* **14,** 647–52 (2007).
192. Woodman, I. L. & Bolt, E. L. Winged helix domains with unknown function in Hel308 and related helicases. *Biochem. Soc. Trans.* **39,** 140–4 (2011).
193. Pena, V. *et. al.* Common design principles in the spliceosomal RNA helicase Brr2 and in the Hel308 DNA helicase. *Mol. Cell* **35,** 454–66 (2009).
194. Woodman, I. L., Briggs, G. S. & Bolt, E. L. Archaeal Hel308 domain V couples DNA binding to ATP hydrolysis and positions DNA for unwinding over the helicase ratchet. *J. Mol. Biol.* **374,** 1139–44 (2007).
195. Takata, K., Reh, S., Tomida, J., Person, M. D. & Wood, R. D. Human DNA helicase HELQ participates in DNA interstrand crosslink tolerance with ATR and RAD51 paralogs. *Nat. Commun.* **4,** 2338 (2013).
196. Moldovan, G.-L. *et al.* DNA polymerase POLN participates in cross-link repair and homologous recombination. *Mol. Cell. Biol.* **30,** 1088–96 (2010).
197. Ben-yehoyada, M. *et al.* Checkpoint signalling from a single DNA interstrand crosslink. *Mol. Cell* **35,** 704–715 (2010).
198. Albala, J. S. *et al.* Identification of a novel human RAD51 homolog, RAD51B. *Genomics* **46,** 476–9 (1997).
199. Dosanjh, M. K. *et al.* Isolation and characterization of RAD51C, a new human member of the RAD51 family of related genes. *Nucleic Acids Res.* **26,** 1179–84 (1998).

200. Pittman, D. L., Weinberg, L. R. & Schimenti, J. C. Identification, characterization, and genetic mapping of Rad51d, a new mouse and human RAD51/RecA-related gene. *Genomics* **49,** 103–11 (1998).
201. Cartwright, R., Tambini, C. E., Simpson, P. J. & Thacker, J. The XRCC2 DNA repair gene from human and mouse encodes a novel member of the recA/RAD51 family. *Nucleic Acids Res.* **26,** 3084–9 (1998).
202. Liu, N. *et al.* XRCC2 and XRCC3, new human Rad51-family members, promote chromosome stability and protect against DNA cross-links and other damages. *Mol. Cell* **1,** 783–93 (1998).
203. Suwaki, N., Klare, K. & Tarsounas, M. RAD51 paralogs: roles in DNA damage signalling, recombinational repair and tumorigenesis. *Semin. Cell Dev. Biol.* **22,** 898–905 (2011).
204. Miller, K. a, Sawicka, D., Barsky, D. & Albala, J. S. Domain mapping of the Rad51 paralog protein complexes. *Nucleic Acids Res.* **32,** 169–78 (2004).
205. Compton, S., Ozgür, S. & Griffith, J. D. Ring-shaped Rad51 paralog protein complexes bind Holliday junctions and replication forks as visualized by electron microscopy. *J. Biol. Chem.* **285,** 13349–56 (2010).
206. Nagaraju, G., Hartlerode, A., Kwok, A., Chandramouly, G. & Scully, R. XRCC2 and XRCC3 regulate the balance between short- and long-tract gene conversions between sister chromatids. *Mol. Cell. Biol.* **29,** 4283–94 (2009).
207. Takata, M. *et. al.* Chromosome Instability and Defective Recombinational Repair in Knockout Mutants of the Five Rad51 Paralogs. *Mol. Cell. Biol.* **21,** 2858–2866 (2001).
208. Liu, Y., Masson, J., Shah, R., Regan, P. O. & West, S. C. RAD51C Is Required for Holliday Junction Processing in Mammalian Cells. **303,** 243–246 (2004).
209. Yokoyama, H. *et. al.* Preferential binding to branched DNA strands and strand-annealing activity of the human Rad51B, Rad51C, Rad51D and Xrcc2 protein complex. *Nucleic Acids Res.* **32,** 2556–65 (2004).
210. Sigurdsson, S. *et. al.* Mediator function of the human Rad51B-Rad51C complex in Rad51/RPA-catalyzed DNA strand exchange. *Genes Dev.* **15,** 3308–18 (2001).
211. Sung, P., Trujillo, K. M. & Komen, S. Van. Recombination factors of *Saccharomyces cerevisiae*. *Mutat. Res.* **451,** 257–275 (2000).
212. Gao, Y. *et al.* Genetic variants at 4q21, 4q23 and 12q24 are associated with esophageal squamous cell carcinoma risk in a Chinese population. *Hum. Genet.* (2013). doi:10.1007/s00439-013-1276-5
213. Pelttari, L. M. *et al.* Screening of HELQ in breast and ovarian cancer families. *Fam. Cancer* **15,** 19–23 (2016).
214. Li, Y. A. P., Yang, J. U. N. J., Xu, H. U. I., Guo, E. N. Y. U. & Yu, Y. A. N. Structure-function analysis of DNA helicase HELQ : A new diagnostic marker in ovarian cancer. *Oncol. Lett.* **12,** 4439–4444 (2016).

215. Vaz, F. *et al.* Mutation of the RAD51C gene in a Fanconi anemia-like disorder. *Nat. Genet.* **42,** 406–9 (2010).
216. Kee, Y. & D'Andrea, A. D. Expanded roles of the Fanconi anemia pathway in preserving genomic stability. *Genes Dev.* **24,** 1680–94 (2010).
217. Bimboim, H. C. & Doly, J. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* **7,** 1513–1523 (1979).
218. Kelley, L. a, Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling , prediction and analysis. *Nat. Protoc.* **10,** 845–858 (2015).
219. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12,** 7–8 (2014).
220. Mukherjee, S. & Zhang, Y. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure* **19,** 955–966 (2011).
221. Lee, H., Park, H., Ko, J. & Seok, C. GalaxyGemini: A web server for protein homo-oligomer structure prediction based on similarity. *Bioinformatics* **29,** 1078–1080 (2013).
222. Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P. & Duhr, S. Molecular Interaction Studies Using Microscale Thermophoresis. *Assay Drug Dev. Technol.* **9,** 342–353 (2011).
223. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat.Methods* **5,** 507–516 (2008).
224. Perez-Gonzalez, C., Lafontaine, D. A. & Penedo, J. C. Fluorescence-Based Strategies to Investigate the Structure and Dynamics of Aptamer-Ligand Complexes. *Front. Chem.* **4,** 33 (2016).
225. Morten, M. J. *et al.* Binding dynamics of a monomeric SSB protein to DNA: A single-molecule multi-process approach. *Nucleic Acids Res.* **43,** 10907–10924 (2015).
226. Bohm, G., Muhr, R. & Jaenicke, R. Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng.* **5,** 191–195 (1991).
227. Hishida, T., Han, Y., Shibata, T., Kubota, Y. & Ishino, Y. Role of the Escherichia coli RecQ DNA helicase in SOS signaling and genome stabilization at stalled replication forks. *Genes Dev.* 1886–1897 (2004). doi:10.1101/gad.1223804
228. Masson, J. Y. *et al.* Identification and purification of two distinct complexes containing the five RAD51 paralogs. *Genes Dev.* **15,** 3296–307 (2001).
229. Bugreev, D. V & Mazin, A. V. Ca2+ activates human homologous recombination protein Rad51 by modulating its ATPase activity. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9988–93 (2004).
230. Fujikane, R., Komori, K., Shinagawa, H. & Ishino, Y. Identification of a novel helicase activity unwinding branched DNAs from the hyperthermophilic archaeon, *Pyrococcus furiosus*. *J. Biol. Chem.* **280,** 12351–12358 (2005).

231. Li, Z. *et al.* Hjm/Hel308A DNA helicase from *Sulfolobus tokodaii* promotes replication fork regression and interacts with Hjc endonuclease in vitro. *J. Bacteriol.* **190,** 3006–17 (2008).
232. Oyama, T. *et al.* Atomic structures and functional implications of the archaeal RecQ-like helicase Hjm. *BMC Struct. Biol.* **9,** 2 (2009).
233. Girard, P. M. *et al.* Oxidative Stress in Mammalian Cells Impinges on the Cysteines Redox State of Human XRCC3 Protein and on Its Cellular Localization. *PLoS One* **8,** (2013).
234. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434 (2005).
235. Kennedy, J., Daughdrill, G. W. & Schmidt, K. H. A transient α-helical molecular recognition element in the disordered N-terminus of the Sgs1 helicase is critical for chromosome stability and binding of Top3/Rmi1. *Nucleic Acids Res.* **41,** 10215–27 (2013).
236. J, S. *et al.* A Helical Bundle in the N-terminal Domain of the BLM Helicase Mediates Dimer and Potentially Hexamer Formation. (2017).
237. Park, J. *et al.* PcrA helicase dismantles RecA filaments by reeling in DNA in uniform steps. *Cell* **142,** 544–555 (2010).
238. Feldkamp, M. D., Mason, A. C., Eichman, B. F. & Chazin, W. J. Structural analysis of replication protein a recruitment of the DNA damage response protein SMARCAL1. *Biochemistry* **53,** 3052–3061 (2014).
239. Hishida, T. *et al.* Role of the Escherichia coli RecQ DNA helicase in SOS signaling and genome stabilization at stalled replication forks. *Genes.* **18,** 1886–1897 (2004).
240. Zhang, L. *et al.* Structural evidence for consecutive Hel308-like modules in the spliceosomal ATPase Brr2. *Nat. Struct. Mol. Biol.* **16,** 731–9 (2009).
241. Flechsig, H., Popp, D. & Mikhailov, A. S. In silico investigation of conformational motions in superfamily 2 helicase proteins. *PLoS One* **6,** e21809 (2011).
242. Raney, K. D. Chemical Modifications of DNA for Study of Helicase Mechanisms. *Bioorganic Med. Chem.* **22,** 4399–4406 (2015).
243. Rudolph, C. J., Upton, A. L., Stockum, A., Nieduszynski, C. A. & Lloyd, R. G. Avoiding chromosome pathology when replication forks collide. *Nature* **500,** 608–611 (2003).
244. Flot, J. *et al.* Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* **500,** 453–457 (2014).
245. Green, M., Hatter, L., Brookes, E., Soultanas, P. & Scott, D. J. Defining the Intrinsically Disordered C-Terminal Domain of SSB Reveals DNA-Mediated Compaction. *J. Mol. Biol.* **428,** 357–364 (2016).

Appendix A: GeneArt DNA sequences and project details

The following pages list synthesized DNA sequences and stratagies utilized in GeneArt (Life Technologies) for cloning and codon optimized expression of human proteins in *E. coli*. The sequences listed are: HelQ (shown as "human hel308", hHel308), Rad51B, Rad51C, Rad51D, XRCC2, XRCC3 and Rad51. The project summaries are included for details of restriction endonuclease sites that were introduced into each DNA fragment, to facilitate sub-cloning of GeneArt constructs from a standard vector provided in the synthesis, to an *E. coli* over-expression plasmid vector, as described in the main results and methods.

Sequence name: hHel308
Sequence type: DNA
Biosafety level: Level 1

Vector resistancy: standard vector contains ampicillin resistance
TSE free: No

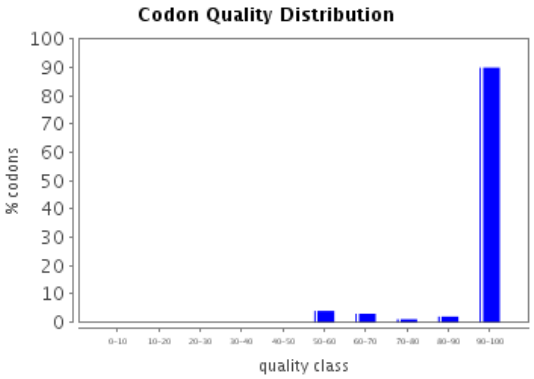
Sequence name / optimized for

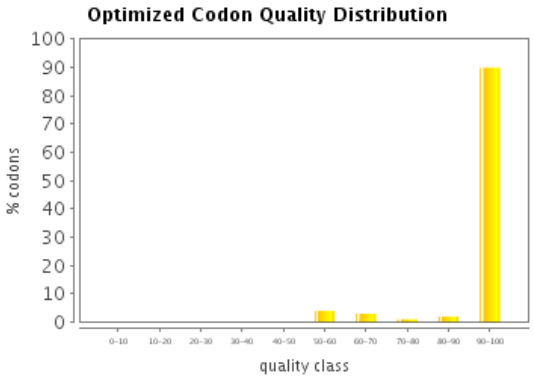
hHel308/ Escherichia coli

ORF	Protected sites	Protected areas	Motifs to avoid
3-3305 [ATG...GCA]		1-2 [CC] 3306-3313 [GCGGCCGC]	

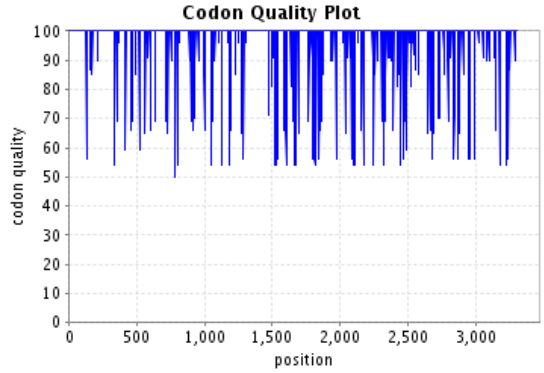
```
      M  D  E  C  G  S  R  I  R  R  R  V  S  L  P  K  R  N  R  P  S  L  G
1.    CCATGGATGAATGTGGTAGCCGTATTCGTCGTCGTGTTAGCCTGCCGAAACGTAATCGTCCGAGCCTGG
      C  I  F  G  A  P  T  A  A  E  L  E  P  G  D  E  G  K  E  E  E  E  M
70.   GTTGTATTTTTGGTGCACCGACCGCAGCAGAACTGGAACCGGGTGATGAAGGTAAAGAAGAGGAAGAAA
      V  A  E  N  R  R  R  K  T  A  G  V  L  P  V  E  V  Q  P  L  L  L  S
139.  TGGTTGCAGAAAATCGTCGTCGCAAAACCGCAGGCGTTCTGCCGGTTGAAGTTCAGCCGCTGCTGCTGA
      D  S  P  E  C  L  V  L  G  G  G  D  T  N  P  D  L  L  R  H  M  P  T
208.  GCGATAGTCCGGAATGTCTGGTTCTGGGTGGTGGTGATACCAATCCGGATCTGCTGCGTCATATGCCGA
      D  R  G  V  G  D  Q  P  N  D  S  E  V  D  M  F  G  D  Y  D  S  F  T
277.  CCGATCGTGGTGTTGGTGATCAGCCGAATGATAGCGAAGTTGATATGTTTGGTGATTACGATAGCTTTA
      E  N  S  F  I  A  Q  V  D  D  L  E  Q  K  Y  M  Q  L  P  E  H  K  K
346.  CCGAAAACAGCTTTATTGCCCAGGTTGATGATCTGGAACAGAAATATATGCAGCTGCCGGAACATAAAA
      H  A  T  D  F  A  T  E  N  L  C  S  E  S  I  K  N  K  L  S  I  T  T
415.  AACACGCAACCGATTTTGCAACCGAAAATCTGTGTAGCGAAAGCATCAAAAACAAACTGAGCATTACCA
      I  G  N  L  T  E  L  Q  T  D  K  H  T  E  N  Q  S  G  Y  E  G  V  T
484.  CCATTGGCAATCTGACCGAACTGCAGACCGATAAACACACAGAAAATCAGAGCGGTTATGAAGGTGTTA
      I  E  P  G  A  D  L  L  Y  D  V  P  S  S  Q  A  I  Y  F  E  N  L  Q
553.  CCATTGAACCTGGTGCCGATCTGCTGTATGATGTGCCGAGCAGCCAGGCAATCTATTTTGAAAATCTGC
      N  S  S  N  D  L  G  D  H  S  M  K  E  R  D  W  K  S  S  S  H  N  T
622.  AGAATAGCAGCAACGATCTGGGTGATCATAGCATGAAAGAACGTGATTGGAAAAGCAGCAGCCATAATA
      V  N  E  E  L  P  H  N  C  I  E  Q  P  Q  Q  N  D  E  S  S  S  K  V
691.  CCGTTAATGAAGAACTGCCGCATAACTGTATTGAACAGCCTCAGCAGAATGATGAAAGCAGCTCAAAAG
      R  T  S  S  D  M  N  R  R  K  S  I  K  D  H  L  K  N  A  M  T  G  N
760.  TTCGTACCAGCAGCGATATGAATCGTCGGAAAAAGCATTAAAGACCATCTGAAAAATGCCATGACCGGTA
      A  K  A  Q  T  P  I  F  S  R  S  K  Q  L  K  D  T  L  L  S  E  E  I
829.  ATGCAAAAGCACAGACCCCGATTTTTAGCCGTAGCAAACAGCTGAAAGATACCCTGCTGAGTGAAGAAA
      N  V  A  K  K  T  I  E  S  S  S  N  D  L  G  P  F  Y  S  L  P  S  K
898.  TTAACGTGGCCAAAAAAACCATCGAATCCAGCTCAAATGATCTGGGTCCGTTTTATAGTCTGCCGAGCA
      V  R  D  L  Y  A  Q  F  K  G  I  E  K  L  Y  E  W  Q  H  T  C  L  T
967.  AAGTTCGTGATCTGTATGCACAGTTTAAAGGCATCGAAAAACTGTATGAATGGCAGCATACCTGTCTGA
      L  N  S  V  Q  E  R  K  N  L  I  Y  S  L  P  T  S  G  G  K  T  L  V
1036. CCCTGAATAGCGTTCAAGAACGTAAAAACCTGATTTATTCACTGCCGACCAGCGGTGGTAAAACCCTGG
      A  E  I  L  M  L  Q  E  L  L  C  C  R  K  D  V  L  M  I  L  P  Y  V
1105. TTGCCGAAATTCTGATGCTGCAAGAACTGCTGTGTTGTCGTAAAGATGTTCTGATGATTCTGCCGTATG
      A  I  V  Q  E  K  I  S  G  L  S  S  F  G  I  E  L  G  F  F  V  E  E
1174. TTGCCATTGTGCAAGAAAAAATCAGCGGTCTGAGCAGCTTTGGTATTGAACTGGGCTTTTTTGTTGAAG
      Y  A  G  S  K  G  R  F  P  P  T  K  R  R  E  K  K  S  L  Y  I  A  T
1243. AATATGCCGGTAGCAAAGGTCGTTTTCCGCCTACCAAACGTCGTGAGAAAAAAAGCCTGTATATTGCCA
      I  E  K  G  H  S  L  V  N  S  L  I  E  T  G  R  I  D  S  L  G  L  V
1312. CCATTGAAAAAGGTCATAGCCTGGTTAATAGCCTGATTGAAACCGGTCGTATTGATAGCCTGGGTCTGG
      V  V  D  E  L  H  M  I  G  E  G  S  R  G  A  T  L  E  M  T  L  A  K
1381. TTGTTGTTGATGAACTGCATATGATTGGTGAAGGTAGCCGTGGTGCAACCCTGGAAATGACCCTGGCAA
      I  L  Y  T  S  K  T  T  Q  I  I  G  M  S  A  T  L  N  N  V  E  D  L
1450. AAATTCTGTATACCAGCAAAACCACGCAGATTATTGGTATGAGCGCAACACTGAATAATGTGGAAGATC
      Q  K  F  L  Q  A  E  Y  Y  T  S  Q  F  R  P  V  E  L  K  E  Y  L  K
1519. TGCAAAAATTTCTGCAAGCCGAGTATTATACCAGCCAGTTTCGTCCGGTTGAACTGAAAGAATATCTGA
      I  N  D  T  I  Y  E  V  D  S  K  A  E  N  G  M  T  F  S  R  L  L  N
1588. AAATCAACGATACCATCTACGAGGTTGATAGCAAAGCAGAAAATGGTATGACATTTAGCCGTCTGCTGA
      Y  K  Y  S  D  T  L  K  K  M  D  P  D  H  L  V  A  L  V  T  E  V  I
1657. ACTATAAATACAGCGACACCCTGAAAAAAATGGATCCTGATCATCTGGTTGCACTGGTTACCGAAGTTA
      P  N  Y  S  C  L  V  F  C  P  S  K  K  N  C  E  N  V  A  E  M  I  C
1726. TTCCGAATTATAGCTGTCTGGTTTTTTGTCCGAGCAAAAAAAACTGCGAAAATGTGGCCGAAATGATTT
      K  F  L  S  K  E  Y  L  K  H  K  E  K  E  K  C  E  V  I  K  N  L  K
1795. GTAAATTCCTGAGCAAAGAGTACCTGAAACACAAAGAAAAAGAAAAATGCGAGGTGATCAAAAACCTGA
      N  I  G  N  G  N  L  C  P  V  L  K  R  T  I  P  F  G  V  A  Y  H  H
1864. AAAACATTGGTAATGGCAATCTGTGTCCGGTTCTGAAACGTACCATTCCGTTTGGTGTTGCATATCATC
```

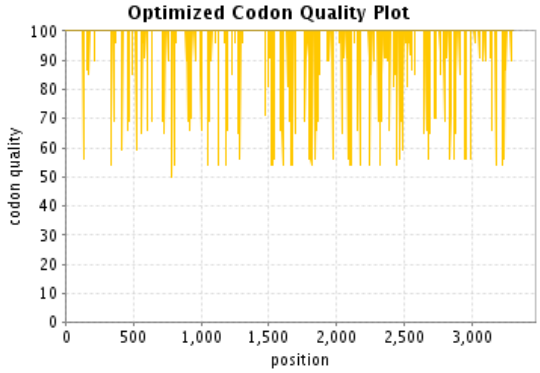
```
          S  G  L  T  S  D  E  R  K  L  L  E  E  A  Y  S  T  G  V  L  C  L  F
1933.  ATAGTGGTCTGACCTCAGATGAACGTAAACTGCTGGAAGAGGCATATAGCACCGGTGTTCTGTGTCTGT
          T  C  T  S  T  L  A  A  G  V  N  L  P  A  R  R  V  I  L  R  A  P  Y
2002.  TTACCTGTACCAGCACCCTGGCAGCCGGTGTTAATCTGCCTGCACGTCGTGTGATTCTGCGTGCACCGT
          V  A  K  E  F  L  K  R  N  Q  Y  K  Q  M  I  G  R  A  G  R  A  G  I
2071.  ATGTGGCAAAAGAATTCCTGAAACGCAACCAGTACAAACAAATGATTGGTCGTGCAGGTCGTGCCGGTA
          D  T  I  G  E  S  I  L  I  L  Q  E  K  D  K  Q  Q  V  L  E  L  I  T
2140.  TTGATACCATTGGTGAAAGCATCCTGATTCTGCAAGAGAAAGATAAACAGCAGGTTCTGGAACTGATTA
          K  P  L  E  N  C  Y  S  H  L  V  Q  E  F  T  K  G  I  Q  T  L  F  L
2209.  CCAAACCGCTGGAAAATTGTTATAGCCATCTGGTGCAAGAATTCACCAAAGGCATTCAGACCCTGTTTC
          S  L  I  G  L  K  I  A  T  N  L  D  D  I  Y  H  F  M  N  G  T  F  F
2278.  TGAGTCTGATTGGTCTGAAAATTGCAACCAACCTGGATGATATCTACCACTTTATGAACGGCACCTTTTT
          G  V  Q  Q  K  V  L  L  K  E  K  S  L  W  E  I  T  V  E  S  L  R  Y
2347.  TTGGTGTGCAGCAGAAAGTGCTGCTGAAAGAAAAAAGTCTGTGGGAAATCACCGTTGAAAGCCTGCGTT
          L  T  E  K  G  L  L  Q  K  D  T  I  Y  K  S  E  E  E  V  Q  Y  N  F
2416.  ATCTGACAGAAAAAGGCCTGCTGCAGAAAGACACCATCTATAAAAGCGAAGAAGAGGTGCAGTATAACT
          H  I  T  K  L  G  R  A  S  F  K  G  T  I  D  L  A  Y  C  D  I  L  Y
2485.  TTCACATTACAAAACTGGGTCGTGCCAGCTTTAAAGGTACAATTGATCTGGCCTATTGCGATATTCTGT
          R  D  L  K  K  G  L  E  G  L  V  L  E  S  L  L  H  L  I  Y  L  T  T
2554.  ATCGCGATCTGAAAAAAGGTCTGGAAGGCCTGGTTCTGGAAAGCCTGCTGCATCTGATTTATCTGACCA
          P  Y  D  L  V  S  Q  C  N  P  D  W  M  I  Y  F  R  Q  F  S  Q  L  S
2623.  CCCCGTATGATCTGGTTAGCCAGTGTAATCCTGATTGGATGATCTATTTTCGTCAGTTCAGTCAGCTGA
          P  A  E  Q  N  V  A  A  I  L  G  V  S  E  S  F  I  G  K  K  A  S  G
2692.  GCCCTGCAGAACAGAATGTTGCAGCAATTCTGGGTGTTTCCGAATCCTTTATTGGTAAAAAAGCAAGCG
          Q  A  I  G  K  K  V  D  K  N  V  V  N  R  L  Y  L  S  F  V  L  Y  T
2761.  GTCAGGCGATTGGCAAAAAAGTTGATAAAAATGTGGTGAACCGTCTGTATCTGAGCTTTGTGCTGTATA
          L  L  K  E  T  N  I  W  T  V  S  E  K  F  N  M  P  R  G  Y  I  Q  N
2830.  CCCTGCTGAAAGAGACTAATATTTGGACCGTTAGCGAAAAATTCAATATGCCTCGTGGCTATATTCAGA
          L  L  T  G  T  A  S  F  S  S  C  V  L  H  F  C  E  E  L  E  E  F  W
2899.  ATCTGCTGACCGGCACCGCAAGCTTTAGCAGCTGTGTTCTGCATTTTTGTGAGGAACTGGAAGAGTTTT
          V  Y  R  A  L  L  V  E  L  T  K  K  L  T  Y  C  V  K  A  E  L  I  P
2968.  GGGTTTATCGTGCACTGCTGGTTGAGCTGACCAAAAAACTGACCTATTGTGTTAAAGCCGAACTGATTC
          L  M  E  V  T  G  V  L  E  G  R  A  K  Q  L  Y  S  A  G  Y  K  S  L
3037.  CGCTGATGGAAGTTACCGGTGTGCTGGAAGGTCGTGCAAAACAGCTGTATAGTGCAGGTTATAAAAGTC
          M  H  L  A  N  A  N  P  E  V  L  V  R  T  I  D  H  L  S  R  R  Q  A
3106.  TGATGCATCTGGCAAATGCAAATCCGGAAGTTCTGGTGCGTACCATCGATCATCTGAGCCGTCGTCAGG
          K  Q  I  V  S  S  A  K  M  L  L  H  E  K  A  E  A  L  Q  E  E  V  E
3175.  CAAAACAAATTGTTAGCAGCGCAAAAATGCTGCTGCATGAAAAAGCAGAAGCACTGCAAGAAGAAGTAG
          E  L  L  R  L  P  S  D  F  P  G  A  V  A  S  S  T  D  K  A
3244.  AGGAACTGCTGCGCCTGCCGAGCGATTTTCCGGGTGCAGTTGCAAGCTCAACCGATAAAGCAGCGGCCG

3313.  C
```



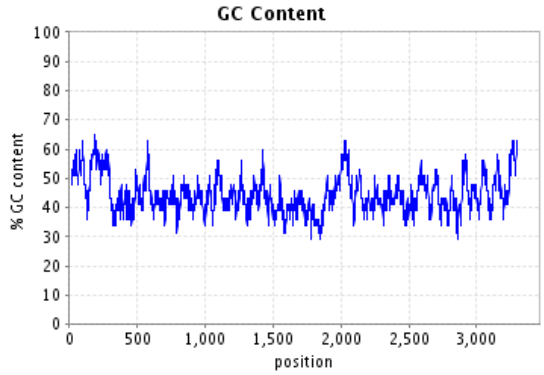


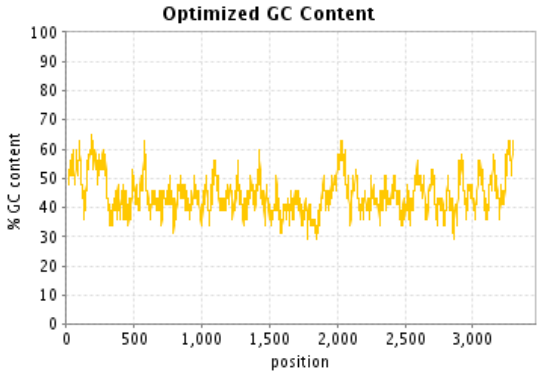
The histograms show the percentage of sequence codons which fall into a certain quality class. The quality value of the most frequently used codon for a given amino acid in the desired expression system is set to 100, the remaining codons are scaled accordingly (see also Sharp, P.M., Li, W.H., Nucleic Acids Res. 15 (3),1987).





The plots show the quality of the used codon at the indicated codon position.





The plots show the GC content in a 40 bp window centered at the indicated nucleotide position.

Project Summary Report

Project Name: Hel308
Report Created:10/03/2013 (8:17 AM)

Gene Synthesis

Construct ID	13ABQRIP
Gene Name	hHel308
Gene Length	3313bp
Host Species	Escherichia coli
Express Upgrade	-
Super speed upgrade	-
TSE free	-
Standard vector option	standard vector contains ampicillin resistance
Sequence encodes toxic protein	-
Comment:	-
ORF	3-3305 [ATG...GCA]
Protected sites	
Protected areas	1-2 Custom [CC] 3306-3313 Custom [GCGGCCGC]
Motifs to avoid	

Subcloning

Construct ID	13ABQRJP
Final construct name	pEBPS1
Vector	Custom
Vector name	pQE-HisStrep-1
Resistance marker	Ampicillin
Copy number of vector	High copy
Site 1	NcoI
Site 2	NotI
3' restriction site	NotI
5' restriction site	NcoI
Vector Part to be used	long
Direction of Insert	Sense
Comment:	-

Plasmid Preparation

Construct ID	13ABQRKP
DNA quantity	50 ml shaker culture (~100 μg pDNA)
Desired buffer	TE
Donor Species of Insert	-
Restriction 1 site for analytical digest	-
Restriction 2 site for analytical digest	-
TSE-free	-
Custom Donor Species of Insert	-
Comment:	-

Sequence name: human_Rad51B
Sequence type: PROTEIN
Biosafety level: Level 1

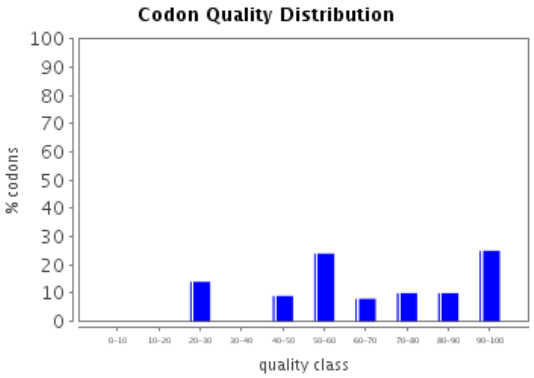
Vector resistancy: standard vector contains ampicillin resistance
TSE free: No

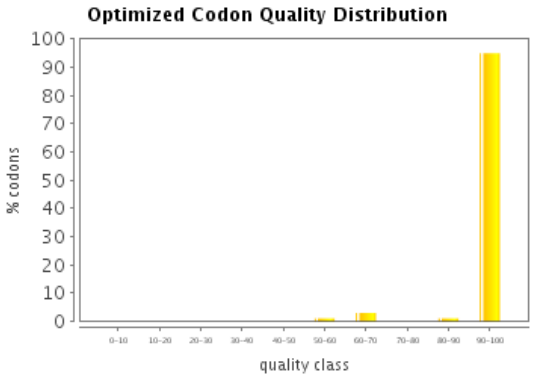
Sequence name / optimized for

human_Rad51B/ Escherichia coli

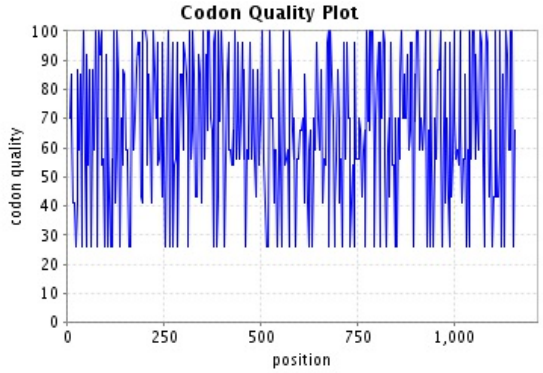
ORF	Protected sites	Protected areas	Motifs to avoid
7-1155 [GGT...TTT]			

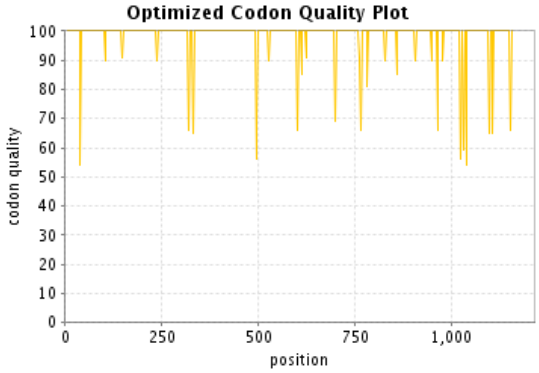
```
                G  S  K  K  L  K  R  V  G  L  S  Q  E  L  C  D  R  L  S  R  H
1.    GGATCCGGTAGCAAAAAACTGAAACGTGTTGGTCTGAGCCAAGAACTGTGTGATCGTCTGAGCCGTCAT
      Q  I  L  T  C  Q  D  F  L  C  L  S  P  L  E  L  M  K  V  T  G  L  S
70.   CAGATTCTGACCTGTCAGGATTTTCTGTGTCTGAGTCCGCTGGAACTGATGAAAGTTACCGGTCTGAGC
      Y  R  G  V  H  E  L  L  C  M  V  S  R  A  C  A  P  K  M  Q  T  A  Y
139.  TATCGTGGTGTGCATGAACTGCTGTGTATGGTTAGCCGTGCATGTGCACCGAAAATGCAGACCGCATAT
      G  I  K  A  Q  R  S  A  D  F  S  P  A  F  L  S  T  T  L  S  A  L  D
208.  GGTATTAAAGCACAGCGTAGCGCAGATTTTAGTCCGGCATTTCTGAGCACCACCCTGAGCGCACTGGAT
      E  A  L  H  G  G  V  A  C  G  S  L  T  E  I  T  G  P  P  G  C  G  K
277.  GAAGCACTGCATGGTGGTGTTGCATGTGGTAGCCTGACCGAAATCACCGGTCCGCCTGGTTGTGGTAAA
      T  Q  F  C  I  M  M  S  I  L  A  T  L  P  T  N  M  G  G  L  E  G  A
346.  ACCCAGTTTTGTATTATGATGAGCATTCTGGCAACCCTGCCGACCAATATGGGTGGTCTGGAAGGTGCA
      V  V  Y  I  D  T  E  S  A  F  S  A  E  R  L  V  E  I  A  E  S  R  F
415.  GTTGTTTATATTGATACCGAAAGCGCATTTAGCGCAGAACGTCTGGTTGAAATTGCAGAAAGCCGTTTT
      P  R  Y  F  N  T  E  E  K  L  L  L  T  S  S  K  V  H  L  Y  R  E  L
484.  CCGCGTTATTTCAACACCGAAGAAAAACTGCTGCTGACCAGCAGTAAAGTTCATCTGTATCGTGAACTG
      T  C  D  E  V  L  Q  R  I  E  S  L  E  E  E  I  I  S  K  G  I  K  L
553.  ACCTGTGATGAAGTTCTGCAGCGTATTGAAAGCCTGGAAGAAGAAATTATCAGCAAAGGCATTAAACTG
      V  I  L  D  S  V  A  S  V  V  R  K  E  F  D  A  Q  L  Q  G  N  L  K
622.  GTGATTCTGGATAGCGTTGCAAGCGTTGTTCGTAAAGAATTTGATGCACAGCTGCAGGGTAATCTGAAA
      E  R  N  K  F  L  A  R  E  A  S  S  L  K  Y  L  A  E  E  F  S  I  P
691.  GAACGTAACAAATTTCTGGCACGTGAAGCAAGCAGCCTGAAATATCTGGCAGAAGAATTTAGCATTCCG
      V  I  L  T  N  Q  I  T  T  H  L  S  G  A  L  A  S  Q  A  D  L  V  S
760.  GTGATCCTGACCAATCAGATTACAACCCATCTGAGCGGTGCACTGGCAAGCCAGGCAGATCTGGTTTCA
      P  A  D  D  L  S  L  S  E  G  T  S  G  S  S  C  V  I  A  A  L  G  N
829.  CCGGCAGATGATCTGAGCCTGAGCGAAGGCACCAGCGGTAGCAGCTGTGTTATTGCAGCACTGGGTAAT
      T  W  S  H  S  V  N  T  R  L  I  L  Q  Y  L  D  S  E  R  R  Q  I  L
898.  ACCTGGTCACATAGCGTTAATACCCGTCTGATTCTGCAGTATCTGGATAGTGAACGTCGTCAGATCCTG
      I  A  K  S  P  L  A  P  F  T  S  F  V  Y  T  I  K  E  E  G  L  V  L
967.  ATTGCAAAATCACCGCTGGCACCGTTTACCAGCTTTGTTTATACCATTAAAGAAGAGGGTCTGGTCCTG
      Q  E  T  T  F  C  S  V  T  Q  A  E  L  N  W  A  P  E  I  L  P  P  Q
1036. CAAGAAACCACCTTTTTGTAGCGTTACCCAGGCAGAACTGAATTGGGCACCGGAAATTCTGCCTCCGCAG
      P  P  E  Q  L  G  L  Q  M  C  H  H  T  Q  L  I  F
1105. CCTCCGGAACAGCTGGGTCTGCAGATGTGTCATCATACCCAGCTGATCTTTTAAGTCGAC
```



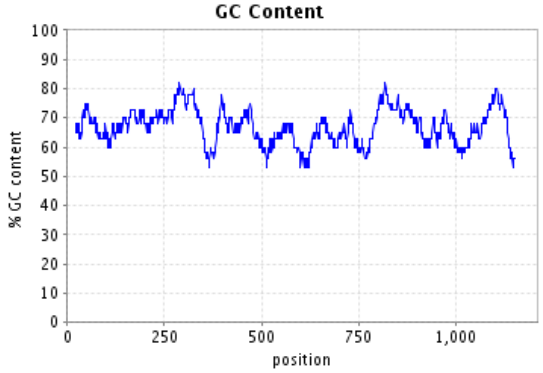


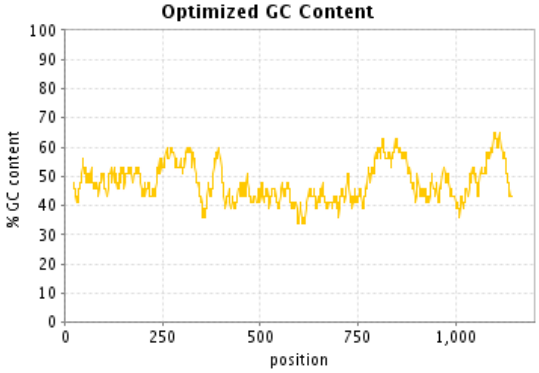
The histograms show the percentage of sequence codons which fall into a certain quality class. The quality value of the most frequently used codon for a given amino acid in the desired expression system is set to 100, the remaining codons are scaled accordingly (see also Sharp, P.M., Li, W.H., Nucleic Acids Res. 15 (3),1987).





The plots show the quality of the used codon at the indicated codon position.





The plots show the GC content in a 40 bp window centered at the indicated nucleotide position.

Sequence name: human_Rad51C
Sequence type: PROTEIN
Biosafety level: Level 1

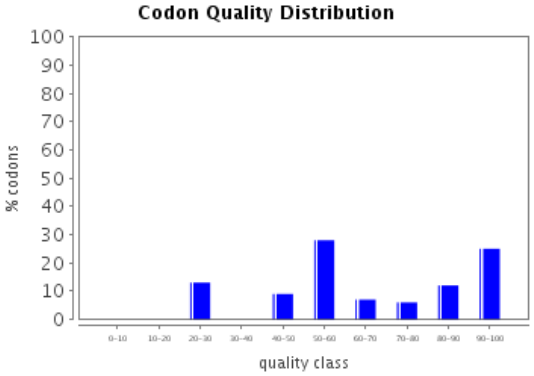
Vector resistancy: standard vector contains ampicillin resistance
TSE free: No

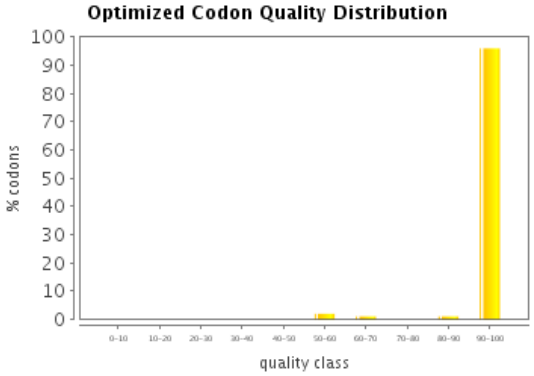
Sequence name / optimized for

human_Rad51C/ Escherichia coli

ORF	Protected sites	Protected areas	Motifs to avoid
4-1131 [ATG...CTG]			

```
        M  R  G  K  T  F  R  F  E  M  Q  R  D  L  V  S  F  P  L  S  P  A
1.      CATATGCGTGGTAAAACCTTTCGTTTTGAAATGCAGCGTGATCTGGTTAGCTTTCCGCTGAGTCCGGCA
        V  R  V  K  L  V  S  A  G  F  Q  T  A  E  E  L  L  E  V  K  P  S  E
70.     GTTCGTGTTAAACTGGTTAGCGCAGGTTTTCAGACCGCAGAAGAACTGCTGGAAGTTAAACCGAGCGAA
        L  S  K  E  V  G  I  S  K  A  E  A  L  E  T  L  Q  I  I  R  R  E  C
139.    CTGAGCAAAGAAGTTGGTATTAGCAAAGCAGAAGCACTGGAAACCCTGCAGATTATTCGTCGTGAATGT
        L  T  N  K  P  R  Y  A  G  T  S  E  S  H  K  K  C  T  A  L  E  L  L
208.    CTGACCAATAAACCGCGTTATGCAGGCACCAGCGAAAGCCACAAAAAAATGTACCGCACTGGAACTGCTG
        E  Q  E  H  T  Q  G  F  I  I  T  F  C  S  A  L  D  D  I  L  G  G  G
277.    GAACAAGAACATACCCAGGGTTTTATCATTACCTTTTGCAGCGCACTGGATGATATTCTGGGTGGTGGT
        V  P  L  M  K  T  T  E  I  C  G  A  P  G  V  G  K  T  Q  L  C  M  Q
346.    GTTCCGCTGATGAAAACCACCGAAATTTGTGGTGCACCGGGTGTGGGTAAAACCCAGCTGTGTATGCAG
        L  A  V  D  V  Q  I  P  E  C  F  G  G  V  A  G  E  A  V  F  I  D  T
415.    CTGGCAGTTGATGTTCAGATCCCGGAATGTTTTGGTGGCGTTGCCGGTGAAGCAGTTTTTATTGATACC
        E  G  S  F  M  V  D  R  V  V  D  L  A  T  A  C  I  Q  H  L  Q  L  I
484.    GAAGGTAGCTTTATGGTTGATCGTGTTGTTGATCTGGCAACCGCATGTATTCAGCATCTGCAGCTGATT
        A  E  K  H  K  G  E  E  H  R  K  A  L  E  D  F  T  L  D  N  I  L  S
553.    GCAGAAAAACACAAAGGTGAAGAACATCGTAAAGCCCTGGAAGATTTTACCCTGGATAATATCCTGAGC
        H  I  Y  Y  F  R  C  R  D  Y  T  E  L  L  A  Q  V  Y  L  L  P  D  F
622.    CACATCTATTATTTCCGCTGTCGTGATTATACCGAGCTGCTGGCACAGGTTTATCTGCTGCCGGATTTT
        L  S  E  H  S  K  V  R  L  V  I  V  D  G  I  A  F  P  F  R  H  D  L
691.    CTGAGCGAACATTCAAAAGTTCGTCTGGTTATTGTTGATGGTATCGCATTTCCGTTTCGTCATGATCTG
        D  D  L  S  L  R  T  R  L  L  N  G  L  A  Q  Q  M  I  S  L  A  N  N
760.    GATGATCTGAGCCTGCGTACCCGTCTGCTGAATGGTCTGGCACAGCAGATGATTAGCCTGGCAAATAAT
        H  R  L  A  V  I  L  T  N  Q  M  T  T  K  I  D  R  N  Q  A  L  L  V
829.    CATCGTCTGGCAGTGATTCTGACAAATCAGATGACCACCAAAATTGATCGTAATCAGGCACTGCTGGTT
        P  A  L  G  E  S  W  G  H  A  A  T  I  R  L  I  F  H  W  D  R  K  Q
898.    CCGGCACTGGGTGAAAGCTGGGGTCATGCAGCAACCATTCGTCTGATTTTTCATTGGGATCGTAAACAG
        R  L  A  T  L  Y  K  S  P  S  Q  K  E  C  T  V  L  F  Q  I  K  P  Q
967.    CGTCTGGCCACCCTGTATAAAAGCCCGAGCCAGAAAGAATGCACCGTTCTGTTTCAGATTAAACCGCAG
        G  F  R  D  T  V  V  T  S  A  C  S  L  Q  T  E  G  S  L  S  T  R  K
1036.   GGTTTTCGTGATACCGTTGTTACCAGCGCATGTAGCCTGCAGACCGAAGGTTCACTGAGCACCCGTAAA
        R  S  R  D  P  E  E  E  L
1105.   CGTAGCCGTGATCCGGAAGAGGAACTGTAAGGTACCGC
```





The histograms show the percentage of sequence codons which fall into a certain quality class. The quality value of the most frequently used codon for a given amino acid in the desired expression system is set to 100, the remaining codons are scaled accordingly (see also Sharp, P.M., Li, W.H., Nucleic Acids Res. 15 (3),1987).

Project Summary Report

Project Name: Rad51BC
Report Created:10/03/2013 (8:6 AM)

Gene Synthesis

Construct ID	13ABQXWP
Gene Name	human_Rad51C
Gene Length	1142bp
Host Species	Escherichia coli
Express Upgrade	-
Super speed upgrade	-
TSE free	-
Standard vector option	standard vector contains ampicillin resistance
Sequence encodes toxic protein	-
Comment:	-
ORF	4-1131 [ATG...CTG]
Protected sites	
Protected areas	
Motifs to avoid	

Gene Synthesis

Construct ID	13ABQXPP
Gene Name	human_Rad51B
Gene Length	1164bp
Host Species	Escherichia coli
Express Upgrade	-
Super speed upgrade	-
TSE free	-
Standard vector option	standard vector contains ampicillin resistance
Sequence encodes toxic protein	-
Comment:	-
ORF	7-1155 [GGT...TTT]
Protected sites	
Protected areas	
Motifs to avoid	

Sequence name: humanRad51D
Sequence type: PROTEIN
Biosafety level: Level 1

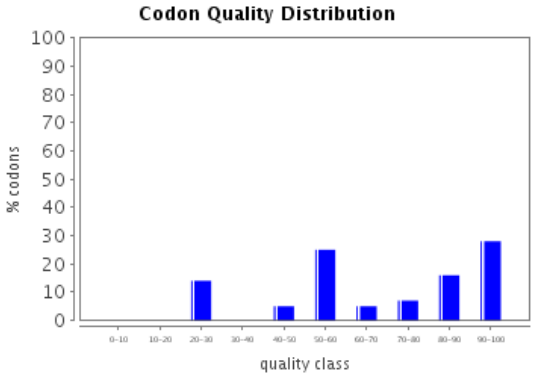
Vector resistancy: standard vector contains ampicillin resistance
TSE free: No

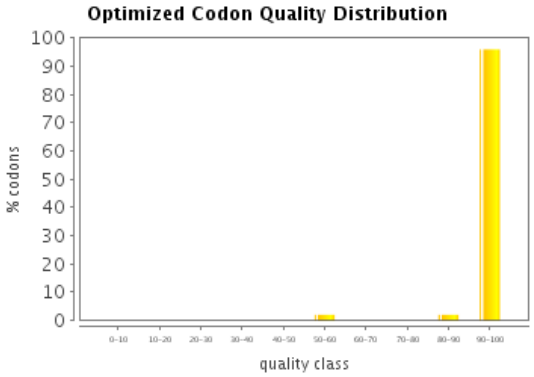
Sequence name / optimized for

humanRad51D/ Escherichia coli

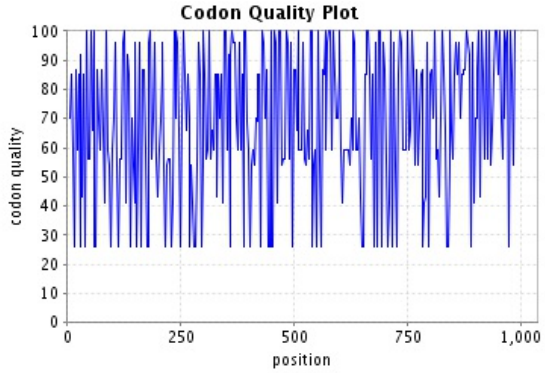
ORF	Protected sites	Protected areas	Motifs to avoid
7-987 [GGT...ACC]			

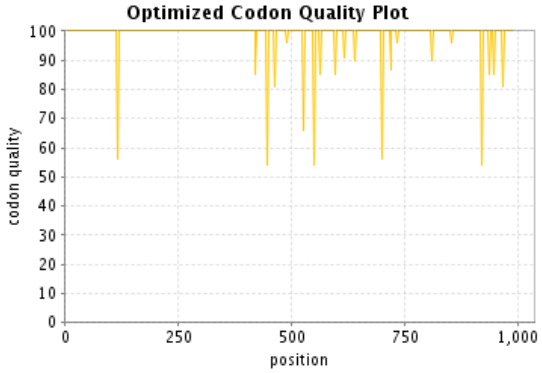
```
          G  V  L  R  V  G  L  C  P  G  L  T  E  E  M  I  Q  L  L  R  S
1.  GGATCCGGTGTTCTGCGTGTTGGTCTGTGTCCGGGTCTGACCGAAGAAATGATTCAGCTGCTGCGTAGC
    H  R  I  K  T  V  V  D  L  V  S  A  D  L  E  E  V  A  Q  K  C  G  L
70. CATCGTATTAAAACCGTTGTTGATCTGGTTAGCGCAGATCTGGAAGAGGTTGCACAGAAATGTGGTCTG
    S  Y  K  A  L  V  A  L  R  R  V  L  L  A  Q  F  S  A  F  P  V  N  G
139. AGCTATAAAGCACTGGTTGCACTGCGTCGTGTTCTGCTGGCACAGTTTAGCGCATTTCCGGTTAATGGT
    A  D  L  Y  E  E  L  K  T  S  T  A  I  L  S  T  G  I  G  S  L  D  K
208. GCAGATCTGTATGAAGAACTGAAAACCAGCACCGCAATTCTGAGCACCGGTATTGGTAGCCTGGATAAA
    L  L  D  A  G  L  Y  T  G  E  V  T  E  I  V  G  G  P  G  S  G  K  T
277. CTGCTGGATGCAGGTCTGTATACCGGTGAAGTTACCGAAATTGTTGGTGGTCCGGGTAGCGGTAAAACC
    Q  V  C  L  C  M  A  A  N  V  A  H  G  L  Q  Q  N  V  L  Y  V  D  S
346. CAGGTTTGTCTGTGTATGGCAGCAAATGTTGCACATGGTCTGCAGCAGAATGTTCTGTATGTTGATAGC
    N  G  G  L  T  A  S  R  L  L  Q  L  L  Q  A  K  T  Q  D  E  E  E  Q
415. AATGGTGGCCTGACCGCAAGCCGTCTGCTGCAACTGCTGCAGGCAAAAACACAGGATGAAGAAGAACAG
    A  E  A  L  R  R  I  Q  V  V  H  A  F  D  I  F  Q  M  L  D  V  L  Q
484. GCAGAAGCCCTGCGTCGTATTCAGGTTGTTCATGCATTTGATATCTTTCAGATGCTGGATGTTCTGCAA
    E  L  R  G  T  V  A  Q  Q  V  T  G  S  S  G  T  V  K  V  V  V  V  D
553. GAACTGCGTGGCACCGTTGCACAGCAGGTTACCGGTAGCAGCGGCACCGTTAAAGTTGTTGTTGTGGAT
    S  V  T  A  V  V  S  P  L  L  G  G  Q  Q  R  E  G  L  A  L  M  M  Q
622. AGCGTTACCGCAGTTGTTAGTCCGCTGCTGGGTGGTCAGCAGCGTGAAGGTCTGGCACTGATGATGCAG
    L  A  R  E  L  K  T  L  A  R  D  L  G  M  A  V  V  V  T  N  H  I  T
691. CTGGCACGTGAGCTGAAAACCCTGGCACGCGATCTGGGTATGGCCGTTGTTGTTACCAATCATATTACC
    R  D  R  D  S  G  R  L  K  P  A  L  G  R  S  W  S  F  V  P  S  T  R
760. CGTGATCGTGATAGCGGTCGTCTGAAACCGGCACTGGGTCGTAGCTGGTCATTTGTTCCGAGCACCCGT
    I  L  L  D  T  I  E  G  A  G  A  S  G  G  R  R  M  A  C  L  A  K  S
829. ATTCTGCTGGATACCATTGAAGGTGCCGGTGCAAGCGGTGGTCGTCGTATGGCATGTCTGGCAAAAAGC
    S  R  Q  P  T  G  F  Q  E  M  V  D  I  G  T  W  G  T  S  E  Q  S  A
898. AGCCGTCAGCCGACCGGTTTTCAAGAAATGGTTGATATTGGCACCTGGGGCACCAGCGAACAGAGCGCA
    T  L  Q  G  D  Q  T
967. ACACTGCAGGGTGATCAGACCTAAAAGCTT
```



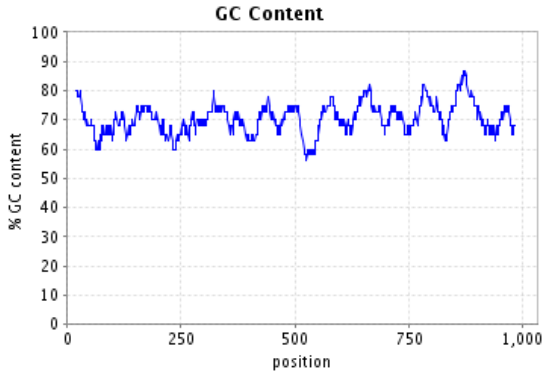


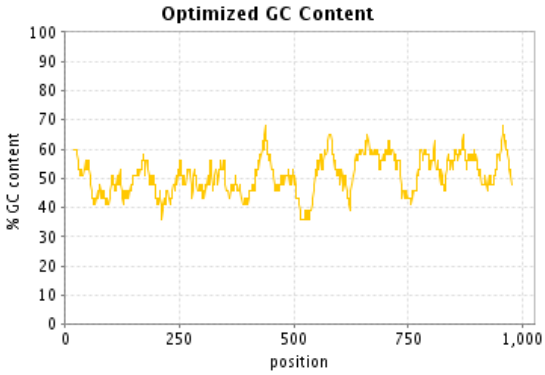
The histograms show the percentage of sequence codons which fall into a certain quality class. The quality value of the most frequently used codon for a given amino acid in the desired expression system is set to 100, the remaining codons are scaled accordingly (see also Sharp, P.M., Li, W.H., Nucleic Acids Res. 15 (3),1987).



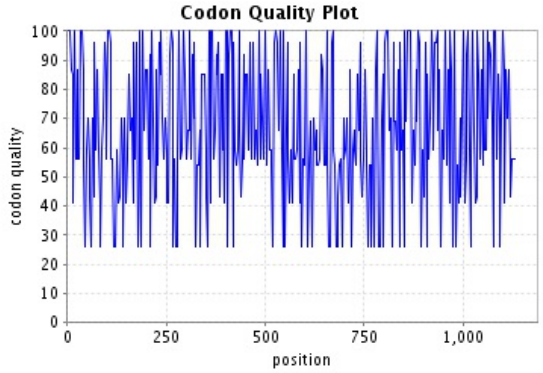


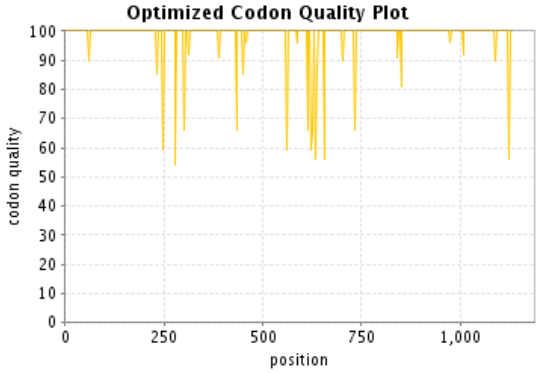
The plots show the quality of the used codon at the indicated codon position.



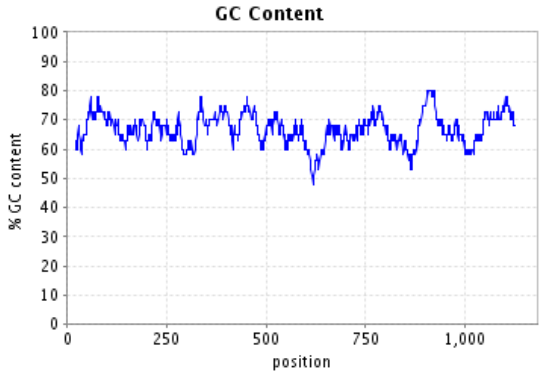


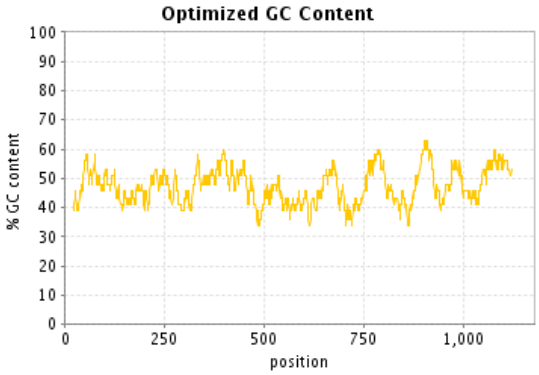
The plots show the GC content in a 40 bp window centered at the indicated nucleotide position.





The plots show the quality of the used codon at the indicated codon position.





The plots show the GC content in a 40 bp window centered at the indicated nucleotide position.

Sequence name: XRCC2
Sequence type: PROTEIN
Biosafety level: Level 1

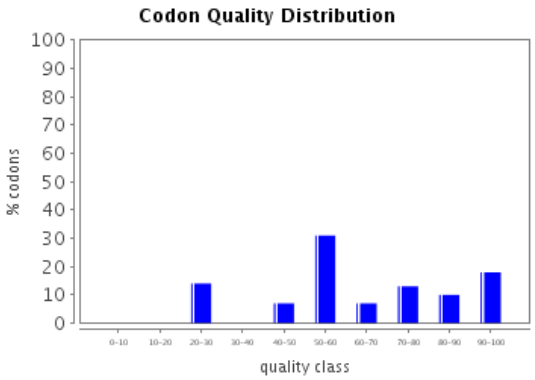
Vector resistancy: standard vector contains ampicillin resistance
TSE free: No

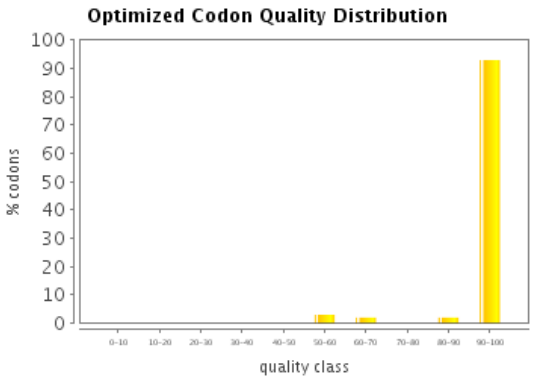
Sequence name / optimized for

XRCC2/ Escherichia coli

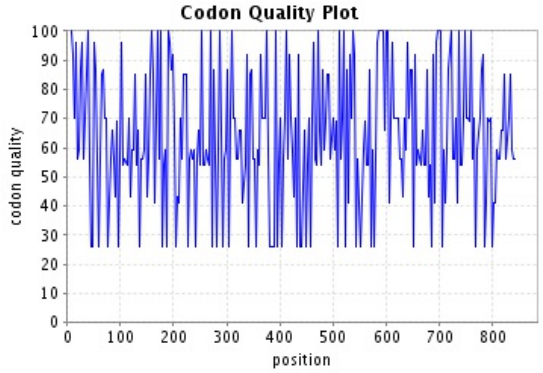
ORF	Protected sites	Protected areas	Motifs to avoid
7-843 [TGT...TGC]			

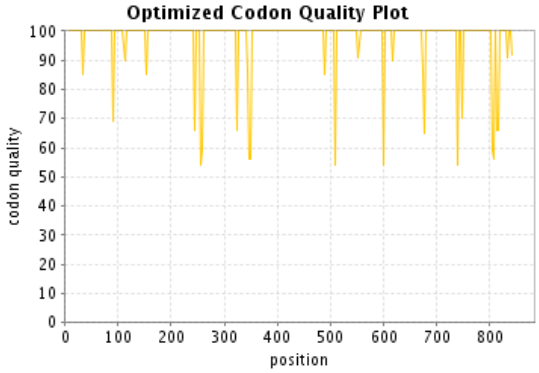
```
          C  S  A  F  H  R  A  E  S  G  T  E  L  L  A  R  L  E  G  R  S
1.  CATATGTGTAGCGCATTTTCATCGTGCAGAAAGCGGCACCGAACTGCTGGCACGTCTGGAAGGTCGTAGC
    S  L  K  E  I  E  P  N  L  F  A  D  E  D  S  P  V  H  G  D  I  L  E
70. AGCCTGAAAGAAATTGAACCGAACCTGTTTGCAGATGAAGATAGTCCGGTTCATGGTGATATTCTGGAA
    F  H  G  P  E  G  T  G  K  T  E  M  L  Y  H  L  T  A  R  C  I  L  P
139.TTTCATGGTCCGGAAGGCACCGGTAAAACCGAAATGCTGTATCATCTGACCGCACGTTGTATTCTGCCG
    K  S  E  G  G  L  E  V  E  V  L  F  I  D  T  D  Y  H  F  D  M  L  R
208.AAAAGCGAAGGTGGTCTGGAAGTTGAAGTTCTGTTTATCGATACCGATTACCACTTTGATATGCTGCGT
    L  V  T  I  L  E  H  R  L  S  Q  S  S  E  E  I  I  K  Y  C  L  G  R
277.CTGGTTACCATTCTGGAACATCGTCTGAGCCAGAGCAGCGAAGAAATTATCAAATATTGTCTGGGTCGC
    F  F  L  V  Y  C  S  S  S  T  H  L  L  L  T  L  Y  S  L  E  S  M  F
346.TTCTTCCTGGTTTATTGTAGCAGCAGCACCCATCTGCTGCTGACCCTGTATAGCCTGGAAAGCATGTTT
    C  S  H  P  S  L  C  L  L  I  L  D  S  L  S  A  F  Y  W  I  D  R  V
415.TGTAGCCATCCGAGCCTGTGTCTGCTGATTCTGGATAGCCTGAGCGCATTTTATTGGATTGATCGTGTT
    N  G  G  E  S  V  N  L  Q  E  S  T  L  R  K  C  S  Q  C  L  E  K  L
484.AATGGTGGCGAAAGCGTTAATCTGCAAGAAAGCACCCTGCGTAAATGTAGCCAGTGTCTGGAAAAACTG
    V  N  D  Y  R  L  V  L  F  A  T  T  Q  T  I  M  Q  K  A  S  S  S  S
553.GTGAATGATTATCGTCTGGTTCTGTTTGCAACCACCCAGACCATTATGCAAAAAGCAAGCAGCAGTAGC
    E  E  P  S  H  A  S  R  R  L  C  D  V  D  I  D  Y  R  P  Y  L  C  K
622.GAAGAACCGAGCCATGCAAGCCGTCGTCTGTGTGATGTTGATATTGATTATCGCCCTTATCTGTGTAAA
    A  W  Q  Q  L  V  K  H  R  M  F  F  S  K  Q  D  D  S  Q  S  S  N  Q
691.GCATGGCAGCAGCTGGTTAAACATCGTATGTTTTTTAGCAAACAGGATGACAGCCAGTCCAGCAATCAG
    F  S  L  V  S  R  C  L  K  S  N  S  L  K  K  H  F  F  I  I  G  E  S
760.TTTAGCCTGGTTAGCCGTTGTCTGAAAAGCAATAGCCTGAAAAAACACTTCTTTATCATCGGTGAAAGC
    G  V  E  F  C
829.GGTGTGGAATTTTGCTAAGGTACC
```



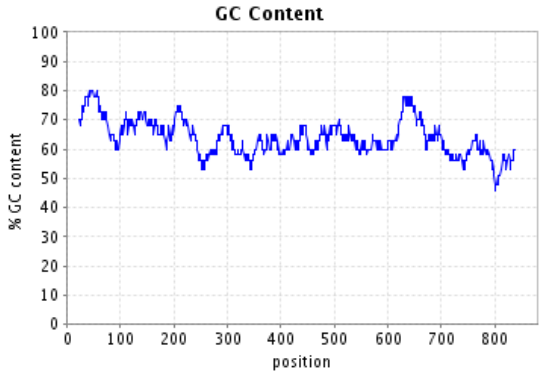


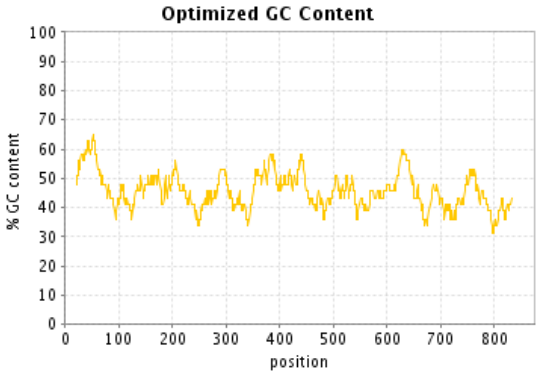
The histograms show the percentage of sequence codons which fall into a certain quality class. The quality value of the most frequently used codon for a given amino acid in the desired expression system is set to 100, the remaining codons are scaled accordingly (see also Sharp, P.M., Li, W.H., Nucleic Acids Res. 15 (3),1987).





The plots show the quality of the used codon at the indicated codon position.





The plots show the GC content in a 40 bp window centered at the indicated nucleotide position.

Sequence name: XRCC3
Sequence type: PROTEIN
Biosafety level: Level 1

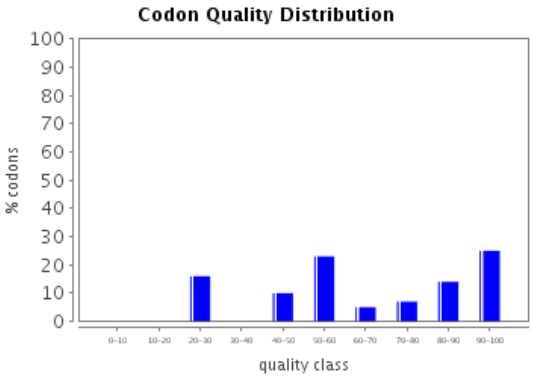
Vector resistancy: standard vector contains ampicillin resistance
TSE free: No

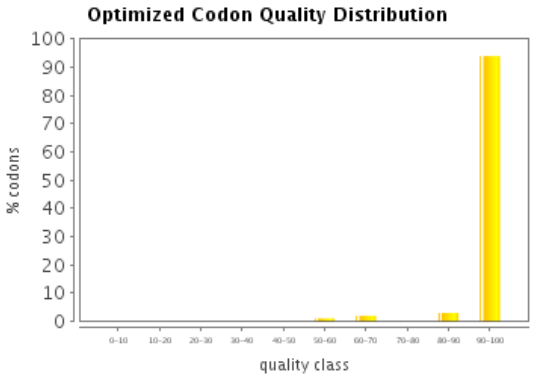
Sequence name / optimized for

XRCC3/ Escherichia coli

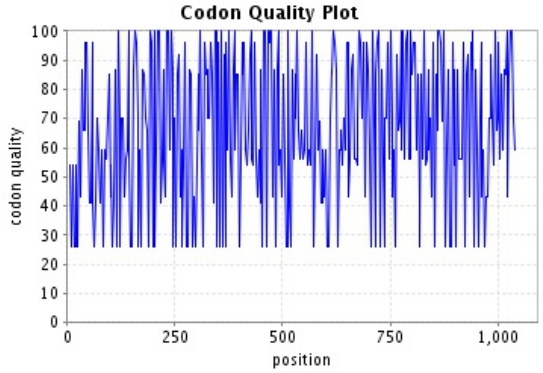
ORF	Protected sites	Protected areas	Motifs to avoid
7-1041 [GAT...CAT]		1-6 [GGATCC] 1042-1050 [TAA...CTT]	

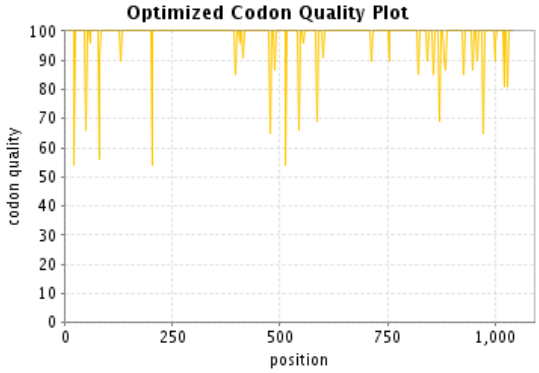
```
                D  L  D  L  L  D  L  N  P  R  I  I  A  A  I  K  K  A  K  L  K
1.    GGATCCGATCTGGATCTGCTGGACCTGAATCCGCGTATTATTGCAGCAATCAAAAAAGCCAAACTGAAA
      S  V  K  E  V  L  H  F  S  G  P  D  L  K  R  L  T  N  L  S  S  P  E
70.   AGCGTTAAAGAGGTGCTGCATTTTAGCGGTCCGGATCTGAAACGTCTGACCAATCTGAGCAGTCCGGAA
      V  W  H  L  L  R  T  A  S  L  H  L  R  G  S  S  I  L  T  A  L  Q  L
139.  GTTTGGCATCTGCTGCGTACCGCAAGCCTGCATCTGCGTGGTAGCAGCATTCTGACCGCACTGCAACTG
      H  Q  Q  K  E  R  F  P  T  Q  H  Q  R  L  S  L  G  C  P  V  L  D  A
208.  CATCAGCAGAAAGAACGTTTTCCGACCCAGCATCAGCGTCTGAGCCTGGGTTGTCCGGTTCTGGATGCA
      L  L  R  G  G  L  P  L  D  G  I  T  E  L  A  G  R  S  S  A  G  K  T
277.  CTGCTGCGTGGTGGTCTGCCGCTGGATGGTATTACCGAACTGGCAGGTCGTAGCAGCGCAGGTAAAACC
      Q  L  A  L  Q  L  C  L  A  V  Q  F  P  R  Q  H  G  G  L  E  A  G  A
346.  CAGCTGGCACTGCAGCTGTGTCTGGCAGTTCAGTTTCCGCGTCAGCATGGTGGCCTGGAAGCCGGTGCA
      V  Y  I  C  T  E  D  A  F  P  H  K  R  L  Q  Q  L  M  A  Q  Q  P  R
415.  GTGTATATTTGTACCGAAGATGCATTTCCGCATAAACGTCTGCAGCAGCTGATGGCACAGCAGCCTCGT
      L  R  T  D  V  P  G  E  L  L  Q  K  L  R  F  G  S  Q  I  F  I  E  H
484.  CTGCGCACCGATGTTCCGGGGTGAACTGCTGCAAAAACTGCGTTTTGGTAGCCAGATTTTTATCGAACAT
      V  A  D  V  D  T  L  L  E  C  V  N  K  K  V  P  V  L  L  S  R  G  M
553.  GTTGCCGATGTTGATACCCTGCTGGAATGTGTTAACAAAAAAGTTCCGGTGCTGCTGAGCCGTGGTATG
      A  R  L  V  V  I  D  S  V  A  A  P  F  R  C  E  F  D  S  Q  A  S  A
622.  GCACGTCTGGTTGTTATTGATAGCGTTGCAGCACCGTTTCGTTGTGAATTTGATAGCCAGGCAAGCGCA
      P  R  A  R  H  L  Q  S  L  G  A  T  L  R  E  L  S  S  A  F  Q  S  P
691.  CCGCGTGCACGTCATCTGCAGAGTCTGGGTGCAACCCTGCGTGAACTGAGCAGCGCATTTCAGTCACCG
      V  L  C  I  N  Q  V  T  E  A  M  E  E  Q  G  A  A  H  G  P  L  G  F
760.  GTTCTGTGTATTAATCAGGTTACCGAAGCAATGGAAGAACAGGGTGCAGCACATGGTCCGCTGGGCTTT
      W  D  E  R  V  S  P  A  L  G  I  T  W  A  N  Q  L  L  V  R  L  L  A
829.  TGGGATGAACGTGTTAGTCCGGCACTGGGCATTACCTGGGCAAACCAGCTGCTGGTGCGCCTGCTGGCA
      D  R  L  R  E  E  E  A  A  L  G  C  P  A  R  T  L  R  V  L  S  A  P
898.  GATCGTCTGCGTGAAGAAGAAGCAGCACTGGGCTGTCCGGCACGTACCCTGCGCGTTCTGAGTGCACCG
      H  L  P  P  S  S  C  S  Y  T  I  S  A  E  G  V  R  G  T  P  G  T  Q
967.  CATCTGCCTCCGAGCAGCTGTAGCTATACCATTAGTGCCGAAGGTGTTCGTGGTACACCGGGTACACAG
      S  H
1036. AGCCATTAAAAGCTT
```



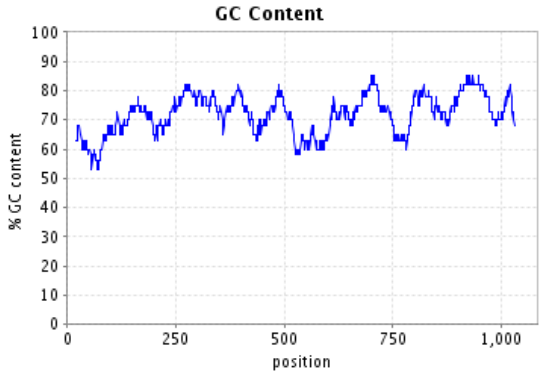


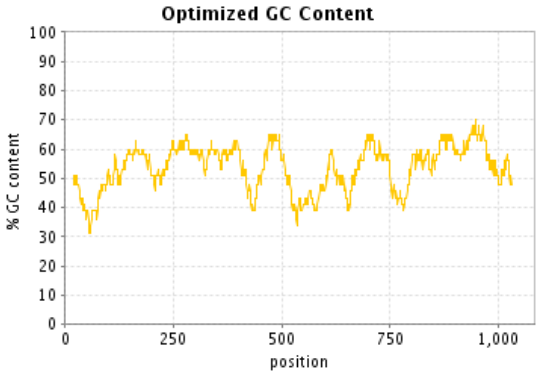
The histograms show the percentage of sequence codons which fall into a certain quality class. The quality value of the most frequently used codon for a given amino acid in the desired expression system is set to 100, the remaining codons are scaled accordingly (see also Sharp, P.M., Li, W.H., Nucleic Acids Res. 15 (3),1987).





The plots show the quality of the used codon at the indicated codon position.





The plots show the GC content in a 40 bp window centered at the indicated nucleotide position.

Sequence name: human_Rad51
Sequence type: PROTEIN
Biosafety level: Level 1

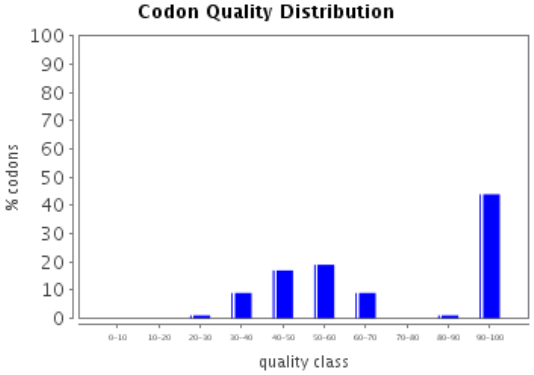
Vector resistancy: standard vector contains ampicillin resistance
TSE free: No

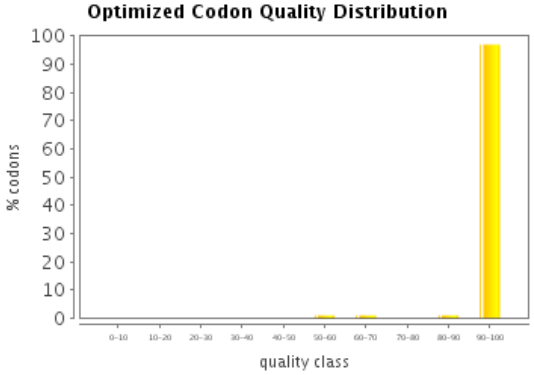
Sequence name / optimized for

human_Rad51/ Escherichia coli

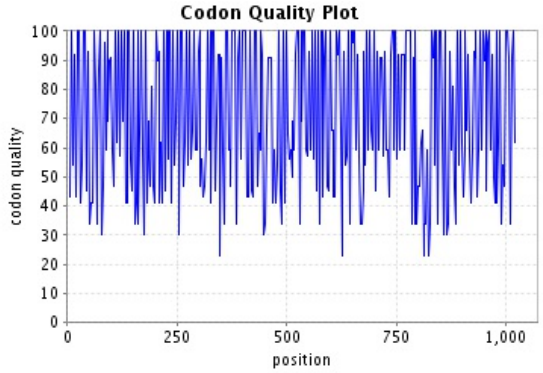
ORF	Protected sites	Protected areas	Motifs to avoid
7-1020 [GCA...GAT]			

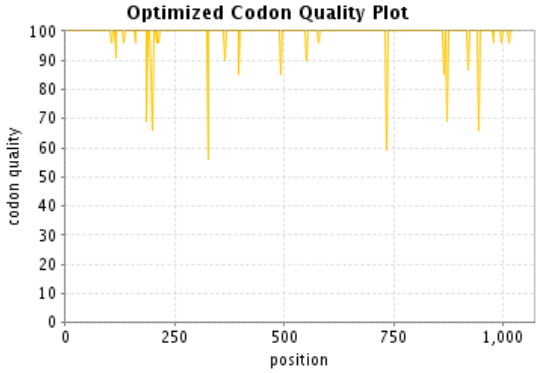
```
        A  M  Q  M  Q  L  E  A  N  A  D  T  S  V  E  E  E  S  F  G  P
1.  CTCGAGGCAATGCAGATGCAGCTGGAAGCAAATGCAGATACCAGCGTTGAAGAAGAAAGCTTTGGTCCG
    Q  P  I  S  R  L  E  Q  C  G  I  N  A  N  D  V  K  K  L  E  E  A  G
70. CAGCCGATTAGCCGTCTGGAACAGTGTGGTATTAATGCCAATGATGTGAAAAAACTGGAAGAAGCCGGT
    F  H  T  V  E  A  V  A  Y  A  P  K  K  E  L  I  N  I  K  G  I  S  E
139. TTTCATACCGTTGAAGCAGTTGCCTATGCACCGAAAAAAGAACTGATTAACATTAAAGGCATCAGCGAA
    A  K  A  D  K  I  L  A  E  A  A  K  L  V  P  M  G  F  T  T  A  T  E
208. GCCAAAGCCGATAAAATTCTGGCAGAAGCAGCAAAACTGGTTCCGATGGGTTTTACCACCGCAACCGAA
    F  H  Q  R  R  S  E  I  I  Q  I  T  T  G  S  K  E  L  D  K  L  L  Q
277. TTTCATCAGCGTCGTAGCGAAATTATTCAGATTACCACCGGTAGCAAAGAGCTGGATAAACTGCTGCAG
    G  G  I  E  T  G  S  I  T  E  M  F  G  E  F  R  T  G  K  T  Q  I  C
346. GGTGGTATTGAAACCGGTAGTATTACCGAAATGTTTGGTGAATTTCGTACCGGCAAAACCCAGATTTGT
    H  T  L  A  V  T  C  Q  L  P  I  D  R  G  G  G  E  G  K  A  M  Y  I
415. CATACCCTGGCAGTTACCTGTCAGCTGCCGATTGATCGTGGTGGTGGTGAAGGTAAAGCAATGTATATT
    D  T  E  G  T  F  R  P  E  R  L  L  A  V  A  E  R  Y  G  L  S  G  S
484. GATACCGAAGGCACCTTTCGTCCGGAACGTCTGCTGGCAGTTGCAGAACGTTATGGTCTGAGCGGTAGT
    D  V  L  D  N  V  A  Y  A  R  A  F  N  T  D  H  Q  T  Q  L  L  Y  Q
553. GATGTTCTGGATAATGTTGCATATGCCCGTGCATTTAATACCGATCATCAGACCCAGCTGCTGTATCAG
    A  S  A  M  M  V  E  S  R  Y  A  L  L  I  V  D  S  A  T  A  L  Y  R
622. GCAAGCGCAATGATGGTTGAAAGCCGTTATGCACTGCTGATTGTTGATAGCGCAACCGCACTGTATCGT
    T  D  Y  S  G  R  G  E  L  S  A  R  Q  M  H  L  A  R  F  L  R  M  L
691. ACCGATTATAGCGGTCGTGGTGAACTGAGCGCACGTCAGATGCACCTGGCACGTTTTCTGCGTATGCTG
    L  R  L  A  D  E  F  G  V  A  V  V  I  T  N  Q  V  V  A  Q  V  D  G
760. CTGCGTCTGGCAGATGAATTTGGTGTTGCAGTTGTTATTACCAATCAGGTTGTTGCACAGGTTGATGGT
    A  A  M  F  A  A  D  P  K  K  P  I  G  G  N  I  I  A  H  A  S  T  T
829. GCAGCAATGTTTGCAGCAGATCCGAAAAAAACCGATTGGCGGTAACATTATTGCACATGCAAGCACCACC
    R  L  Y  L  R  K  G  R  G  E  T  R  I  C  K  I  Y  D  S  P  C  L  P
898. CGTCTGTATCTGCGTAAAGGTCGCGGTGAAACCCGTATTTGTAAAATCTATGATAGCCCGTGTCTGCCG
    E  A  E  A  M  F  A  I  N  A  D  G  V  G  D  A  K  D
967. GAAGCAGAAGCCATGTTTGCAATTAATGCCGATGGTGTTGGTGATGCCAAAGATTGAGGTACC
```



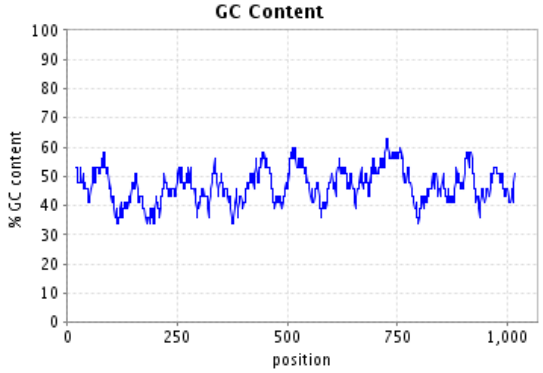


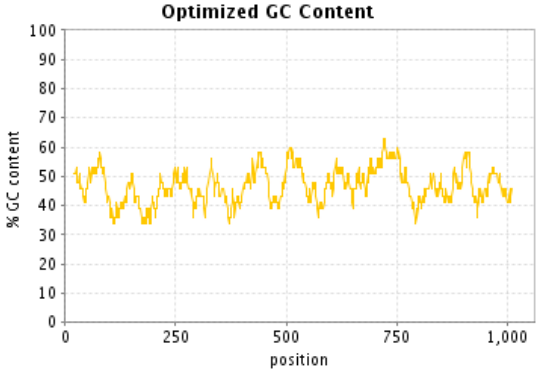
The histograms show the percentage of sequence codons which fall into a certain quality class. The quality value of the most frequently used codon for a given amino acid in the desired expression system is set to 100, the remaining codons are scaled accordingly (see also Sharp, P.M., Li, W.H., Nucleic Acids Res. 15 (3),1987).





The plots show the quality of the used codon at the indicated codon position.





The plots show the GC content in a 40 bp window centered at the indicated nucleotide position.

Project Summary Report

Project Name: Human Rad51
Report Created:10/03/2013 (7:44 AM)

Gene Synthesis

Construct ID	13ABQSPP
Gene Name	human_Rad51
Gene Length	1029bp
Host Species	Escherichia coli
Express Upgrade	-
Super speed upgrade	-
TSE free	-
Standard vector option	standard vector contains ampicillin resistance
Sequence encodes toxic protein	-
Comment:	-
ORF	7-1020 [GCA...GAT]
Protected sites	
Protected areas	
Motifs to avoid	

Subcloning

Construct ID	13ABQSQP
Final construct name	pEBPS2
Vector	Life
Vector name	pBAD/His_A
Resistance marker	Ampicillin
Copy number of vector	Low copy
Site 1	XhoI
Site 2	KpnI
3' restriction site	KpnI
5' restriction site	XhoI
Vector Part to be used	long
Direction of Insert	Sense
Comment:	-

Plasmid Preparation

Construct ID	13ABQSRP
DNA quantity	50 ml shaker culture (~100 μg pDNA)
Desired buffer	TE
Donor Species of Insert	-
Restriction 1 site for analytical digest	-
Restriction 2 site for analytical digest	-
TSE-free	-
Custom Donor Species of Insert	-
Comment:	-

Appendix B: Multiple alignments of Hel308 and HelQ sequences.

Amino acid sequences of Human HelQ (shown as "human Hel308") aligned with archaeal Hel308 sequences from *Methanothermobacter thermautotrophicus, Sulfolobus solfataricus, Archaeoglobus fulgidus and Pyrococcus furiosus.* The alignment was generated using Clustal X, and note that the first 300 amino acids of Human Hel308 (HelQ) are missing for clarity, because they have no significant homology to any archaeal Hel308, as described in Chapter 3 and 4 of the main thesis. The black box highlights the Winged Helix domains of each protein, as determined experimentally (Chapter 5) or predicted. Within this the red box highlights an alpha helix that was subjected to the analysis given in Chapter 5.

S1

```
Human-Hel308            VAKKTIESSSNDLGPFYSLPSKVRDLY-AQFKGIEKLYEWQHTCLTLNSVQERKNLIYSL   359
Methanothermobacter     ---------------MKSLPPEMRQILGDCYPHIRELNPAQRSAIEAGYLESEDNYIIAI   45
Sulfolobus              LEWMPIED------LK--LPSNVIEIIK--KRGIKKLNPPQTEAVKKGLLEG-NRLLLTS   53
Archaeoglobus           ---MKVEE------LAESISSYAVGILK--EEGIEELFPPQAEAVE-KVFSG-KNLLLAM   47
Pyrococcus              ---MRVDE------LR--VDERIKSTLK--ERGIESFYPPQAEALKSGILEG-KNALISI   46
                                     :             *..:   *  .:    ..  .. : :

Human-Hel308            PTSGGKTLVAEILMLQELLCCRKDVLMILPYVAIVQEKISGLSS---FGIELGFFVEEYA   416
Methanothermobacter     PTASGKTLLGIIAALKTV-MEGGRVIYTVPLLSIQNEKIKEFRKLEEHGIRVGKDPR---   101
Sulfolobus              PTGSGKTLIAEMGIISFLLKNGGKAIYVTPLRALTNEKYLTFKDWELIGFKVAMTSGDYD   113
Archaeoglobus           PTAAGKTLLAEMAMVREA-IKGGKSLYVVPLRALAGEKYESFKKWEKIGLRIGISTGDYE   106
Pyrococcus              PTASGKTLIAEIAMVHRILTQGGKAVYIVPLKALAEEKFQEFQDWEKIGLRVAMATGDYD   106
                        **..****:. :  :          :   *  ::  **   : .    *:.:.

Human-Hel308            GSKGRFPPTKRREKKSLYIATIEKGHSLVNSLIETGRIDSLGLVVVDELHMIGEGSRGAT   476
Methanothermobacter     -------------TSDIAVMVFESFDSLTRF--SWNILREVDLLIVDEFHMIGEYTRGPV   146
Sulfolobus              TDDAWL------KNYDIIITTYEKLDSLWRH--RPEWLNEVNYFVLDELHYLNDPERGPV   165
Archaeoglobus           SRDEHL------GDCDIIVTTSEKADSLIRN--RASWIKAVSCLVVDEIHLLDSEKRGAT   158
Pyrococcus              SKDEWL------GKYDIIIATAEKFDSLLRH--GSSWIKDVKILVADEIHLIGSRDRGAT   158
                                  .: : . *. .** .       :  :  .: **:* :..  ** .

Human-Hel308            LEMTLAKILYTSKTTQIIGMSATLNNVEDLQKFLQAEYYTSQFRPVELKEYLKINDTIYE   536
Methanothermobacter     IESAITRARTLNPSVRIVALSATLSNMDEIAGWLDARVVEHDYRPVPLHREVLDTEMFG-   205
Sulfolobus              VESVTIRA----KRRNLLALSATISNYKQIAKWLGAEPVATNWRPVPLIEGVIYPERKK-   220
Archaeoglobus           LEILVTKMRRMNKALRVIGLSATAPNVTEIAEWLDADYYVSDWRPVPLVEGVLCEGTLE-   217
Pyrococcus              LEVILAHML---GKAQIIGLSATIGNPEELAEWLNAELIVSDWRPVKLRRGVFYQGFVT-   214
                        :*    :        .::.:***  *  ::  :* *     ::*** * . :

Human-Hel308            VDSKAENGMTFSRLLNYKYSDTLKKM-DPDHLVALV-TEVIPNYSCLVFCPSKKNCENVA   594
Methanothermobacter     ----VR----E---------------KNDVVLKVLERSLEDGSQTLAFVSTRRFTESLA   241
Sulfolobus              ----KEYNVIFK-------DNTTKKVHGDDAIIAYTLDSLSKNGQVLVFRNSRKMAESTA   269
Archaeoglobus           ----LFDGAFST-------S-------RRVKFEELVEECVAENGGVLVFESTRRGAEKTA   259
Pyrococcus              ----WEDGSID---------------RFSSWEELVYDAIRKKKGALIFVNMRRKAERVA   254
                                                          :     * *   ::  *  *
```

S1 (cont.)


Human-Hel308
Methanothermobacter
Sulfolobus
Archaeoglobus
Pyrococcus
alpha helix-20

S1 (cont.)

```
Human-Hel308              NRLYLSFVLYTLLKETNIWTVSEKFNMPRGYIQNLLTGTASFSSCVLHFCEELEEFWVYR   992
Methanothermobacter       NEEATAAALIEWINERTEYEIENAFHVYAASTRRSAYEASKIVKFFGKICEIMGVYRH-S   617
Sulfolobus                NALKVALIMKDWMDEVDEDTILSKYNIGSGDLRNMVETMDWLTYSAYHLSRELKLNEH-A   638
Archaeoglobus             SEVKTALCLKDWIEEKDEDEICAKYGIAPGDLRRIVETAEWLSNAMNRIAEEVGNTSV-S   617
Pyrococcus                RAFKTALVLLAWINEVPEGEIVEKYSVEPGDIYRIVETAEWLVYSLKEIAKVLGAYEI-V   627
                               :  :   :.*     :   : :  .   .       :     .:.. :

Human-Hel308              ALLVELTKKLTYCVKAELIPLM-EVTGVLEGRAKQLYSAGYKSLMHLANANPEVLVRTID   1051
Methanothermobacter       SQLEILSARLYYGVKEDAIPLVVGVRGLGRVRARKIIKTF----GEDLRHVREDELKRID   673
Sulfolobus                DKLRILNLRVRDGIKEELLELV-QISGVGRKRARLLYNNGIKELGDVVMNPDKVKNLLGQ   697
Archaeoglobus             ----GLTERIKHGVKEELLELV-RIRHIGRVRARKLYNAGIRNAEDIVRHREKVASLIGR   672
Pyrococcus                DYLETLRVRVKYGIREELIPLM-QLPLVGRRRARALYNSGFRSIEDISQARPEELLKI-E   685
                               *  ::   :: : : *:  :  : . **: : .       .      :

Human-Hel308              HLSRRQAKQIVSSAKMLLHEKAEALQEEVEELLRLPSDFPGAVASSTDKA     1101
Methanothermobacter       GIGPKMAGAIRRYCERF--------------------------------     690
Sulfolobus                KLGEKVVQEAARLLNRFH-------------------------------     715
Archaeoglobus             GIAERVVEGIS------VK-SLNPES-----------------------     691
Pyrococcus                GIGVKTVEAIFKFLGKNVKISEKPRKSTLDYFLKS---------------     720
                           :. : .
```