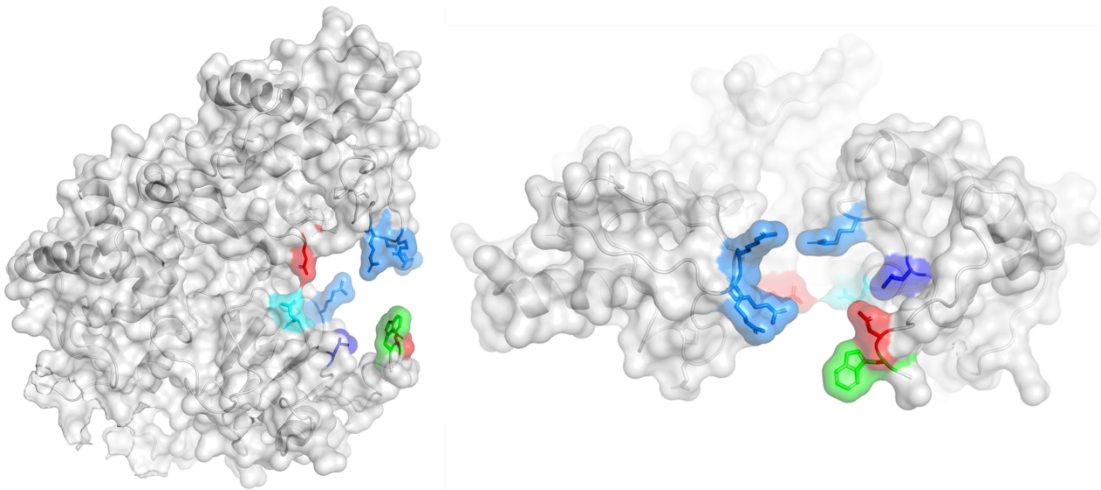


Biochemical and genetic analysis of
Lhr, a mysterious helicase/glycosylase
conserved across species



A thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

Ryan J. Buckley, MRes

November 2022

Abstract

Changes to DNA structure frequently inhibit essential processes such as DNA replication. Overcoming replication blockage or collapse requires replication-coupled DNA repair enzymes that catalyse removal of aberrant DNA structures and chemically modified bases. The Lhr family of helicases are found throughout archaea, including in the Heimdall- and Nano- archaeota, and are present in several bacterial clades. This family can be divided into 'Lhr-core' and 'Lhr-extended' protein variants with the latter containing an as yet uncharacterised extended C-terminal domain.

Through genetic analysis we identified an expression phenotype of archaeal Lhr identical to the replication-coupled DNA repair enzymes Hel308 and RecQ and implicated bacterial Lhr in a novel mutation repair pathway and in overcoming oxidative stress through interaction with a Rad51 paralogue.

In vitro analysis demonstrated archaeal Lhr preferential targeting of replication fork structures through ATP-independent binding causing melting/distortion of the branch point. This allowed loading for directional translocation and unwinding through the 'parental' DNA strands. Characterisation of bacterial Lhr-CTD revealed a newly identified d-uracil DNA glycosylase activity, building an emerging story about the contribution of Lhr and its associated proteins in prokaryotic DNA repair. Further context is afforded through phylogenetic analysis of RecA/Rad51 family proteins revealing the emergence of protein sub-families.

Here we present a substantial breakthrough in the study of Lhr proteins, implicating them for direct involvement in replication-coupled repair and a wider role in base excision repair.

Acknowledgments

I would like to start by thanking my supervisor Ed for giving me the opportunity to undertake numerous projects within his lab. I would like to thank him for his continued guidance, support and encouragement, and for giving me a space to grow as a person and as a scientist. Ed has enabled me to set a path into the scientific world and for that I will always be in his debt.

I would like to thank Chris Cooper (and his lab), Dina Grohmann and Kevin Kramm for their hard work and ideas, coming at a time when I thought I was alone battling against Lhr's many 'head-scratching' quirks. Thanks also go to the BBSRC and the University of Nottingham for their funding and support throughout my project.

I would not have been able to finish my PhD (without losing my sanity) without the support and encouragement from my friends and family. I would like to give special thanks to Tom – for his jokes and endless source of knowledge, Tabi – for the movie nights and crème eggs(!), He – for her delicious baking, Ryan Reeves – for our endless search for riffs and his support through tough times, Harry – for the tea breaks and hiking trips, Andy – for his knowledge on all things funky, Ryan Wall and Oli – for giving me much needed time away from the lab to detox my brain, and of course my family, for always being in my corner and supporting me through all the highs and lows of life.

I would also like to give thanks to the sounds of Elder, Colour Haze and ISIS (the progressive metal band) for getting me through the mind numbing spot tests and helping me focus through long writing sessions.

COVID-19 Impact Statement

This PhD project originally centred around the study of Rad51 paralogue proteins from *Trypanosoma brucei*. Our aim was to purify and characterise these proteins to be used as a model for their human counterparts. During the first COVID-19 lockdown in 2020, I performed phylogenetic analysis of these proteins to try and ascertain how related they were to the human paralogues, in order to guide characterisation once lab work resumed. This analysis highlighted the stark differences between the *T. brucei* and *H. sapiens* paralogues suggesting limitations in their use as a model. This result, as well as the great difficulties seen in trying to overexpress and purify *T. brucei* recombinant proteins from *E. coli*, lead to an abandoning of this side of the project.

In addition, the phylogenetic analysis highlighted an interesting relationship between RadA (Sms), a poorly studied RecA/Rad51 family protein from *E. coli*, and *H. sapiens* Rad51 C. From this work, along with that produced by the Lovett group¹, and building upon work performed during my MRes, we were able to change tactics and move our attention to Lhr and its potential interaction with RadA (Sms) in *E. coli*. A change in focus to Lhr family proteins allowed a continuation of study into the early stages of DNA repair, building upon knowledge/skills gained in the first year of my PhD whilst also tasking me with a project which would be more achievable within the resulting time left following the resumption of lab work.

To that end, the RecA/Rad51 family proteins will be omitted from the introduction as to not disrupt the focus from Lhr with the phylogenetic data presented in brief in the later sections due to its contribution in shaping the rest of my PhD work. Phylogenetic data also presents an interesting insight into these proteins which I feel should be documented.

Table of contents

Abstract	I
Acknowledgments	II
COVID-19 Impact Statement	III
List of Figures	XIII
List of Tables	XIX
List of Abbreviations.....	XXI
<i>Chapter 1 : Introduction</i>	<i>1</i>
1.1 Structure and function of DNA	1
1.2 DNA replication	6
1.3 Replication stress.....	10
1.4 DNA damage.....	14
1.5 DNA repair systems	16
1.5.1 A brief history of DNA glycosylases with mechanistic insights.....	16
1.5.2 Repair of DNA chemical alterations	19
1.5.3 Repair of bulky DNA adducts	21
1.5.4 Repair of mismatched bases	23
1.5.5 Homologous recombination (HR)	24
1.5.6 Rescue of stalled replication forks	27

1.6 DNA Helicases.....	29
1.6.1 Superfamily-1 and superfamily-2 helicases	29
1.7 Lhr	31
1.7.1 Initial discovery	31
1.7.2 Genetics give clues to function	32
1.7.3 Protein structure and domain organisation	34
1.7.4 Biochemical characterisation.....	37
1.7.5 Bioinformatic analysis	38
1.8 Project aims.....	41
<i>Chapter 2 Materials and Methods</i>	42
2.1 Antibiotics	42
2.2 Bacterial strains and cell lines	43
2.3 Plasmids and DNA substrates.....	46
2.4 Solution composition	59
2.4.1 Media	59
2.4.2 Electrophoresis running buffers and stains	61
2.4.3 Solutions used for substrate preparation and <i>in vitro</i> protein analysis	63
2.4.4 Buffers used during protein purification	64
2.5 General Microbiology	66
2.5.1 Protocol for making competent cells.....	66
2.5.2 Protocol for transforming competent cells	66

2.5.3 Polymerase chain reaction (PCR).....	67
2.5.4 Site directed mutagenesis.....	68
2.5.5 Production of knockout cell lines for genetic analysis	69
2.6 Gel electrophoresis	72
2.6.1 Agarose gel electrophoresis.....	72
2.6.2 Agarose gel extraction protocol.....	72
2.6.3 SDS PAGE analysis	72
2.7 Phylogenetic analysis.....	73
2.7.1 MUSCLE multiple sequence alignment.....	73
2.7.2 Gblocks alignment check	73
2.7.3 PhyML tree generation and TreeDyn visualisation.....	73
2.7.4 List of proteins used for phylogenetic analysis.....	74
2.8 In vivo experimentation	77
2.8.1 Hydrogen peroxide viability spot assay	78
2.8.2 Hydrogen peroxide growth curves	78
2.8.3 Mitomycin C viability spot assay.....	79
2.8.4 Azidothymidine viability spot assay.....	79
2.8.5 Rifampicin viability assay	79
2.8.6 Genetic analysis of <i>Methanothermobacter thermautotrophicus lhr</i>	80
2.9 In vitro experimentation	81
2.9.1 Preparation of 5'-end labelled ³² P DNA substrates	81
2.9.2 Preparation of 5'-end labelled Cy5 DNA substrates	81

2.9.3 DNA binding assays	82
2.9.4 DNA unwinding assays	83
2.9.5 DNA glycosylase assays	84
2.10 Protein overexpression and purification.....	86
2.10.1 Obtaining <i>M. thermautotrophicus</i> Lhr protein	86
2.10.2 Obtaining <i>E. coli</i> full length Lhr protein and D1536A mutant	88
2.10.3 Obtaining purified <i>E. coli</i> C-terminal Lhr protein and D1536A mutant	90
2.10.4 Obtaining purified <i>E. coli</i> RNaseT protein.....	92
2.10.5 Obtaining purified <i>E. coli</i> RadA (Sms)	92
 <i>Bioinformatic study of RecA/Rad51 family proteins and bacterial</i>	
<i>Lhr helicase</i>	94
 3.1 Phylogenetic analysis of RecA/Rad51 family of proteins	94
3.1.1 Introduction to RecA/Rad51 family of proteins	94
3.1.2 A brief introduction to <i>E. coli</i> RadA (Sms).....	95
3.1.3 Construction of phylogenetic trees	96
3.1.4 Conservation of Walker A/B active sites.....	97
3.1.5 Analysis of <i>T. brucei</i> as a model for Rad51 paralogue proteins.....	103
3.1.6 Identification of 'Rad51 C' and 'XRCC3' subfamilies.....	106
 3.2 Distribution of Lhr in bacteria and archaea	109
3.2.1 Introduction to Large Helicase Related (Lhr) proteins	109
3.2.2 Study of the genomic context of Lhr across multiple species	110
3.2.3 Analysis of <i>E. coli</i> Lhr gene regulation and codon usage	113

3.2.4 Identification of Lhr abundance in bacteria.....	117
3.3 Summary of Key findings	120
3.3.1 Phylogenetic analysis of RecA/Rad51 family proteins.....	120
3.3.2 Investigation into Lhr family protein genomic context and bacterial distribution.....	121
<i>Chapter 4 : ‘Mechanistic insights into Lhr helicase function in DNA repair’</i>	123
4.1 Introduction to Lhr-core.....	123
4.2 Genetic analysis of Lhr-core	125
4.2.1 Archaeal Lhr localises at stalled replication forks.....	125
4.3 Biochemical analysis of Lhr-core	128
4.3.1 Identification of Lhr-core polarity and optimal ATP:Mg ²⁺ ratio	128
4.3.2 Lhr-core unwinds branched DNA substrates more readily	130
4.3.3 Lhr-core preferentially unwinds fully base-paired forked DNA substrates	132
4.3.4 Lhr-core preferentially unwinds through the parental fork DNA strands	134
4.3.5 Lhr-core remodels fork DNA substrates in the absence of ATP and MgCl ₂ , prior to translocation and unwinding	136
4.3.6 Lhr novel C-terminal region matches to a glycosylase repair protein.....	139
4.4 Summary of Key findings	142
4.4.1 Initial identification of replication-coupled DNA repair protein	142

4.4.2 Assessment of <i>MthLhr</i> unwinding characteristics and substrate preference	142
4.4.3 <i>MthLhr</i> remodels replication fork DNA and unwinds through parental DNA strands	143
4.4.4 Computational identification of an α -helical bundle and identification of AlkZ-like CTD	144

Chapter 5 ‘The *E. coli* DNA helicase *Lhr* is also a DNA-uracil

***glycosylase*’145**

5.1 Introduction 145

5.2 Identification of DNA repair phenotypes..... 147

5.2.1 Production of knockout strains for genetic analysis.....148

5.2.2 *Lhr* may be involved in replication-associated DNA break repair152

5.2.3 *Lhr* is involved in a repair response following exposure to oxidative agents
.....155

5.2.4 *Lhr* is not directly related to the repair of ICL or double strand DNA breaks
.....159

5.2.5 *Lhr* and RadA (*Sms*) may be part of a mutation inducing repair pathway161

5.3 Purification of *E. coli* *Lhr*-extended..... 165

5.3.1 Optimised expression of non-tagged *E. coli* *Lhr*166

5.3.2 Purification of non-tagged *E. coli* *Lhr*-extended169

5.3.3 Purification of his-tagged *E. coli* *Lhr*-extended.....173

5.4 Purification of *E. coli* C-*Lhr* 177

5.4.1 Cloning of <i>E. coli</i> C-Lhr	177
5.4.2 Purification of <i>E. coli</i> C-Lhr	177
5.5 Investigation into <i>E. coli</i> Lhr DNA binding capacity	179
5.5.1 <i>E. coli</i> Lhr requires single stranded DNA for binding	180
5.5.2 <i>E. coli</i> Lhr stably binds 'damaged' DNA substrates.....	182
5.5.3 Lhr-CTD is unable to stably bind DNA in an EMSA.....	184
5.6 Investigation into <i>E. coli</i> Lhr DNA glycosylase activity.....	186
5.6.1 <i>E. coli</i> Lhr-CTD acts as a d-uracil stimulated DNA glycosylase.....	187
5.6.2 <i>E. coli</i> full length Lhr shows focused glycosylase activity in the presence of Mn ²⁺	189
5.6.3 Investigation of <i>E. coli</i> full length and Lhr-CTD glycosylase activities	191
5.6.4 <i>E. coli</i> full length Lhr glycosylase activity was investigated on multiple substrates.....	192
5.6.5 Determination of <i>E. coli</i> Lhr glycosylase activity in the presence of metal ions and ATP.....	194
5.6.6 Comparison of <i>E. coli</i> Lhr glycosylase activity to UDG and Fpg commercial DNA glycosylases.....	196
5.7 Investigation of Lhr-CTD glycosylase active site residues	198
5.7.1 Bioinformatic analysis of <i>E. coli</i> Lhr-CTD active site	199
5.7.2 Cloning and purification of <i>E. coli</i> Lhr CTD mutant D1536A	201
5.7.3 Cloning and purification of <i>E. coli</i> full length Lhr mutant D1536A	203
5.8 Investigation into glycosylase active site residues	205

5.8.1 D1536A mutation causes loss of glycosylase function in both full length and CTD Lhr	205
5.8.2 FL-Lhr D1536A binds and unwinds DNA	206
5.9 Summary of key findings.....	208
5.9.1 Lhr is involved in suppressing replicative stress and in oxidative damage repair.....	208
5.9.2 Lhr and RadA (Sms) may be part of a mutagenic repair pathway	208
5.9.3 Lhr-core is required for Lhr binding and loading onto exposed ssDNA....	209
5.9.4 Lhr displays d-uracil DNA glycosylase activity	209
5.9.5 Identification of a key aspartic acid residue for d-uracil glycosylase activity	210
<i>Chapter 6 :Discussion and future research</i>	<i>211</i>
6.1 Assessment of project aims.....	211
6.1.1 RecA/Rad51 family proteins	211
6.1.2 Lhr family proteins	212
6.2 Evaluation of RecA/Rad51 protein phylogenetics	214
6.3 Lhr substrate preference.....	216
6.4 Lhr as a dual function protein	218
6.5 Lhr in a biological context	220
<i>Chapter 7 References.....</i>	<i>224</i>

Chapter 8 : Appendices	254
8.1 Appendix 1	254
8.1.1 <i>M. thermautotrophicus</i> Lhr protein sequence	254
8.1.2 <i>E. coli</i> Lhr protein sequences	255
8.1.3 <i>E. coli</i> RadA (Sms) protein sequence	258
8.1.4 <i>E. coli</i> RNaseT protein sequence	258
8.2 Distribution of Lhr among bacteria.....	259
8.3 H₂O₂ growth curves.....	265
8.4 <i>In vitro</i> analysis of <i>E. coli</i> Lhr gel examples	266
8.4.1 <i>E. coli</i> Lhr activity on d-uracil duplex DNA	266
8.4.2 <i>E. coli</i> Lhr-CTD activity on d-uracil flayed duplex and dsDNA.....	267
8.5 <i>E. coli</i> RadA (Sms) protein purification	268
8.6 <i>E. coli</i> RNaseT protein purification	269
8.7 <i>In vitro</i> analysis of <i>E. coli</i> RadA (Sms) and RNaseT	270
8.7.1 <i>E. coli</i> RNaseT is a DNA nuclease	270
8.7.2 <i>E. coli</i> RadA (Sms) EMSA	271
Chapter 9 Publications.....	272

List of Figures

Figure 1.1 Structural features of the DNA double helix.

Figure 1.2 Chemical structure of DNA.

Figure 1.3 Semi-conservative DNA replication resulting in daughter duplexes which contain a parental strand (black) and a newly synthesised strand (white).

Figure 1.4 Basic diagrammatic representation of the bacterial replisome, highlighting key components.

Figure 1.5 Example of an inhibitory obstacle poised downstream of the replisome complex.

Figure 1.6 DNA damage tolerance pathways alleviate replicative stress caused by inhibitory lesions.

Figure 1.7 Potential sources of damage (1), resulting DNA lesions (2) and most likely repair mechanism (3).

Figure 1.8 Characteristic domain organisation of HTH_42 superfamily proteins.

Figure 1.9 Examples of purine and pyrimidine lesions repaired by BER and potential resulting abasic sites.

Figure 1.10 Nucleotide excision repair pathway in *E. coli* following UV damage.

Figure 1.11 Simplified overview of double strand break repair via homologous recombination in *E. coli*.

Figure 1.12 Rescue of stalled replication forks may occur through distinct pathways dependent on lesion encountered.

Figure 1.13 The subfamilies of SF1 and SF2 helicases.

Figure 1.14 Domain organisation of Lhr proteins from archaea and bacteria.

Figure 1.15 Lhr-extended comprises of two distinct protein groups.

Figure 1.16 Lhr distribution in archaea and bacteria.

Figure 1.17 Lhr-core is located downstream of a metallophosphoesterase (MPE) in multiple bacteria.

Figure 3.1 Identification of highly conserved residues located within the Walker A/B active site using multiple sequence alignment.

Figure 3.2 Mapping of conserved residues identified by multiple sequence alignments.

Figure 3.3 Analysis of *Trypanosoma brucei* Rad51 paralogue candidates by phylogenetics.

Figure 3.4 Analysis of RecA/Rad51 family of proteins from a diverse set of organisms.

Figure 3.5 Lhr genomic context in multiple organisms.

Figure 3.6 Lhr expression may be controlled by a temperature sensitive promoter.

Figure 3.7 Analysis of *E. coli* *lhr* codon usage.

Figure 3.8 Identification of *lhr* rare codons within the first 313 codons.

Figure 3.9 Lhr presence in bacterial phyla.

Figure 4.1 *M. thermautotrophicus* Lhr interacts with stalled replication forks in *E. coli dnaE486 ΔrecQ* cells.

Figure 4.2 *M. thermautotrophicus* Lhr-core preferentially translocates and unwinds branched DNA substrates with a 3' to 5' polarity.

Figure 4.3 Lhr-core readily unwinds branched DNA substrates.

Figure 4.4 *M. thermautotrophicus* Lhr-core targets branched nucleic acids with a preference to fully base-paired DNA:DNA substrates.

Figure 4.5 *M. thermautotrophicus* Lhr-core binding and unwinding of DNA fork substrates giving clues on unwinding directionality.

Figure 4.6 Single-molecule FRET analysis of conformational changes induced by *M. thermautotrophicus* Lhr-core on fully base paired forked DNA.

Figure 4.7 Bioinformatic analysis of Lhr-core and extended C-terminus

Figure 5.1 AlphaFold predicted structure of *E. coli* Lhr.

Figure 5.2 Generation of knockout strains for use in genetic analysis.

Figure 5.3 Confirmation of *E. coli* Δ lhr Δ radA sensitivity to azidothymidine exposure highlighting potential roles in resolving DNA break associated replicative-stress.

Figure 5.4 *E. coli* Lhr is involved in oxidative damage repair.

Figure 5.5 *E. coli* RadA (Sms) is involved in ICL repair but does not function alongside Lhr.

Figure 5.6 *E. coli* Lhr and RadA (Sms) knockout cells show reduced acquired resistance to rifampicin as compared to wild type cells.

Figure 5.7 Overexpression optimisation of non-tagged full length *E. coli* Lhr.

Figure 5.8 *E. coli* Lhr elutes from a butyl sepharose column at a range of salt concentrations.

Figure 5.9 *E. coli* Lhr is purified using ion exchange chromatography.

Figure 5.10 Purification of *E. coli* Lhr.

Figure 5.11 Purification of *E. coli* Lhr C-terminus.

Figure 5.12 *E. coli* Lhr requires regions of ssDNA for stable binding.

Figure 5.13 EMSA of *E. coli* Lhr's ability to bind a d-uracil containing DNA substrate.

Figure 5.14 *E. coli* Lhr-CTD is unable to stably bind a ssDNA substrate.

Figure 5.15 *E. coli* Lhr-CTD glycosylase activity on 'undamaged' and d-uracil 'damaged' DNA substrates.

Figure 5.16 *E. coli* full length Lhr glycosylase activity in the presence of Mg^{2+} and Mn^{2+} .

Figure 5.17 Comparative analysis of *E. coli* full length Lhr and Lhr-CTD glycosylase activities on a ssDNA d-uracil containing substrate.

Figure 5.18 Comparative analysis of *E. coli* full length Lhr glycosylase activities on multiple d-uracil containing DNA substrates.

Figure 5.19 Analysis of *E. coli* full length Lhr glycosylase activity in the presence and absence of various metal coordination ions and ATP.

Figure 5.20 Analysis of *E. coli* full length Lhr glycosylase activity on d-uracil and 8-oxo-d-guanine containing DNA substrates as compared to commercial glycosylases.

Figure 5.21 Phyre² structural model of *E. coli* Lhr-CTD with proposed active site residues as annotated.

Figure 5.22 Purification of *E. coli* Lhr CTD D1536A mutant.

Figure 5.23 Purification of *E. coli* Lhr D1536A mutant.

Figure 5.24 Effect of D1536A mutation on *E. coli* Lhr glycosylase activity.

Figure 5.25 *E. coli* Lhr DNA binding and unwinding activity is unaffected by a D1536A mutation.

Figure 8.1 *MthLhr* amino acid sequence as used.

Figure 8.2 *EcoLhr* protein sequence untagged as used.

Figure 8.3 *EcoLhr* with cloned in His₆ tag and important residues highlighted.

Figure 8.4 *EcoLhr*-CTD protein sequence as used.

Figure 8.5 *E. coli* RadA (Sms) as used here.

Figure 8.6 *E. coli* RNaseT as used here.

Figure 8.7 Analysis of Lhr-extended proteins among bacterial phyla when present.

Figure 8.8 Effect of hydrogen peroxide on *E. coli* knockout cell strains full results.

Figure 8.9 *E. coli* Lhr glycosylase activity on a d-uracil duplex DNA substrate.

Figure 8.10 *E. coli* Lhr-CTD glycosylase activity on damaged flayed duplex and dsDNA substrates.

Figure 8.11 Purification of *E. coli* RadA (Sms).

Figure 8.12 Purification of *E. coli* RNaseT.

Figure 8.13 *E. coli* RNaseT acting as a potent nuclease.

Figure 8.14 *E. coli* RadA (Sms) EMSA in the presence and absence of ADP.

List of Tables

Table 2.1 Antibiotics used for cloning.

Table 2.2 List of cell strains used within this study.

Table 2.3 Cell strains produced by me during this work for genetic analysis.

Table 2.4 A list of plasmid vectors used within this study for cloning and overexpressions.

Table 2.5 A list of plasmid constructs used within this study.

Table 2.6 List of primers used for cloning.

Table 2.7 List of primers used for site directed mutagenesis (SDM).

Table 2.8 List of primers used for strain verification.

Table 2.9 List of oligonucleotides used for *in vitro* analysis.

Table 2.10 A list of media broths used within this study for *E. coli* cell culture cloning, protein overexpression and genetic analysis.

Table 2.11 A list of running buffers used for agarose gel electrophoresis and sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS PAGE) analysis for separation of DNA and protein respectively.

Table 2.12 Composition of buffers used in nucleic acid substrate preparation and in nucleic acid binding, unwinding and glycosylase assays.

Table 2.13 Composition of buffers used during protein purification.

Table 2.14 PCR programme for standard Vent PCR.

Table 2.15 PCR programme for standard Q5 hot start PCR.

Table 2.16 Proteins used from *Homo sapiens* with corresponding PDB IDs as described.

Table 2.17 Proteins used from *Trypanosoma brucei* with corresponding PDB IDs as described.

Table 2.18 Proteins used from *Methanothermobacter thermautotrophicus* with corresponding PDB IDs as described.

Table 2.19 Proteins used from *Escherichia coli* with corresponding PDB IDs as described.

Table 2.20 Proteins used from *Caenorhabditis elegans* with corresponding PDB IDs as described.

Table 2.21 Proteins used from *Synechococcus elongatus* with corresponding PDB IDs as described.

Table 2.22 Genotoxic agents used within this study and type of damage as indicated.

Table 8.1 Results from Blastp search of *E. coli* Lhr against each bacterial phyla.

List of Abbreviations

AP	abasic
A	adenine
AZT	azidothymidine
BER	base excision repair
BIR	break-induced repair
CTD	C-terminal domain
C	cytosine
D-loop	displacement loop
DTT	DNA damage tolerance
DSBR	double-strand break repair
dsDNA	double stranded DNA
ssDNA	single stranded DNA
G	guanine
HJ	Holliday junction
HR	homologous recombination
HTH	helix-turn-helix
ICL	interstrand crosslink
Lhr	large helicase-related

MMR	mismatch repair
MMC	mitomycin C
NER	nucleotide excision repair
ROS	reactive oxygen species
RDR	recombination dependent replication
SF	superfamily
T	thymine
TLS	translesion synthesis
U	uracil
WH	winged-helix

Chapter 1 : Introduction

1.1 Structure and function of DNA

The genetic code equips cells with tools to allow survival within a given environment through transcription of an RNA intermediate and then translation into functional proteins. Faithful replication of genomic DNA is of essential importance allowing coding sequences and regulatory elements to be passed on to subsequent generations². A cell's DNA is under constant threat from damaging sources which can alter its chemical structure. Persistent damage can have disastrous effects on the cell leading to genetic instability, mutation and if severe enough, cellular death³⁻⁵. The importance of maintaining genomic integrity is epitomised by the emergence of diverse, specialised DNA metabolism proteins. These proteins are involved in the maintenance of DNA including the synthesis and degradation reactions involved in DNA replication and repair. Many of these proteins have co-evolved between the two (or three) domains of life⁶⁻⁸. For this piece of work we will focus in on proteins which aid in replication-coupled DNA repair.

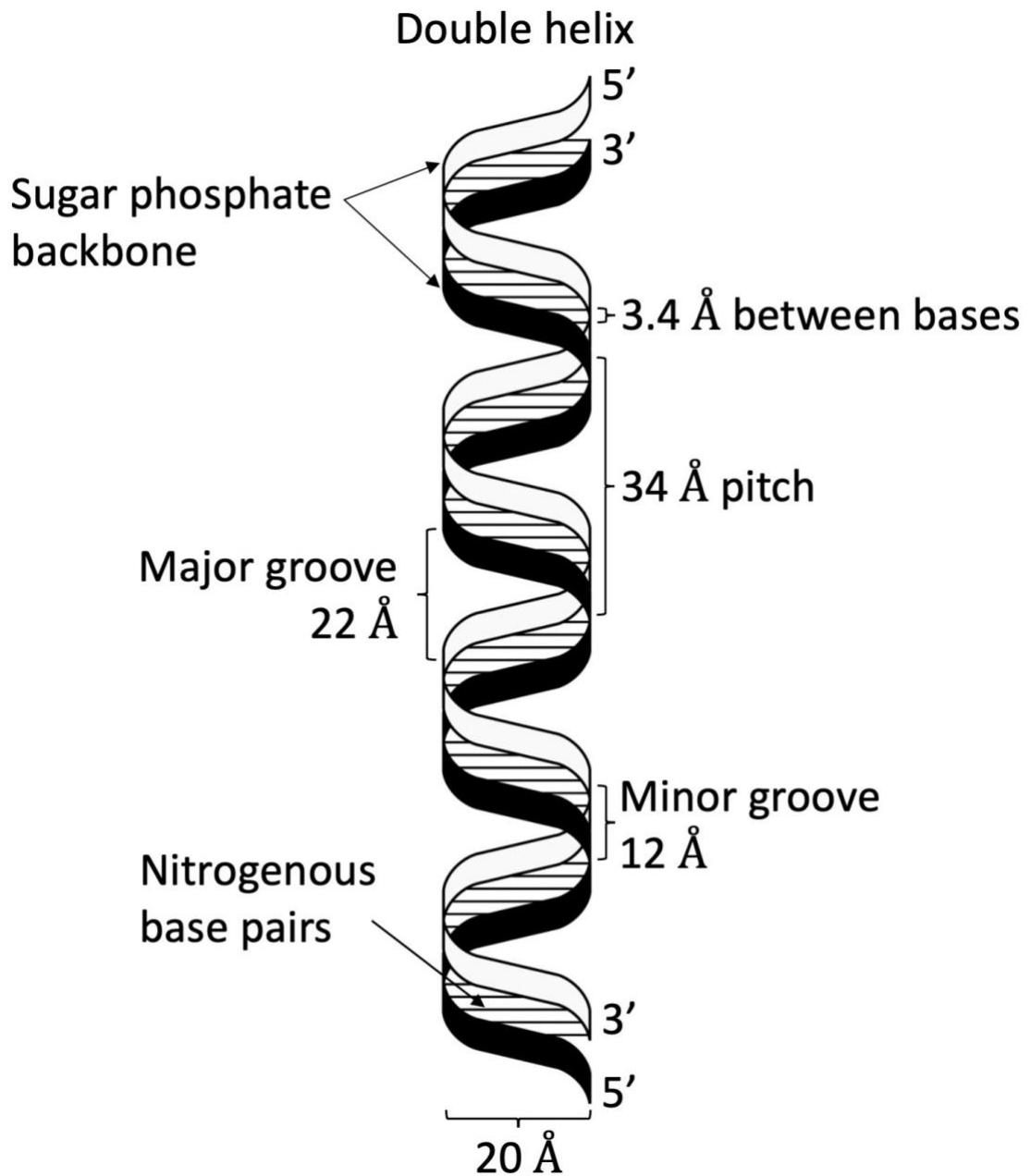


Figure 1.1 Structural features of the DNA double helix.

The genetic code is held within the sequence of nitrogenous bases.

DNA is a polymer comprising of two polynucleotide chains wrapped together, bound through hydrogen bonds. A hydrogen bond forms between a hydrogen atom covalently bonded to an electronegative donor and the lone pair of electrons of an electronegative acceptor, affording a weak interaction. Within the context of DNA, the accumulative hydrogen bonding between bases allow a strong interaction between

strands⁹. Each polynucleotide monomer consists of a deoxyribose sugar covalently linked to a phosphate group and a variable nucleobase (see Figure 1.2). Alternating sugar-phosphate linkages between adjacent nucleotides form a negatively charged phosphodiester backbone¹⁰. Sugar phosphate linkages occur between an oxygen atom covalently bonded to either the 3' or 5' carbon of the deoxyribose sugar. This gives the DNA a polarity which has important implications for its metabolism and the proteins which interact with it¹¹. Strand association is dependent upon the correct formation of weak hydrogen bond linkages between nucleobases of opposing strands. For a classical double helix adenine (A) must bind to thymine (T) through two hydrogen bonds and guanine (G) must bind to cytosine (C) through three¹² (Figure 1.2 C). DNA within this B-form is relatively stable however, it is liable to attack by reactive endogenous and exogenous agents which may lead to DNA damage¹³. Proteins which function on DNA are in-tune with its chemical properties, changing this through damage can lead to inhibition posing a big problem for the cell¹⁴, as will be discussed below.

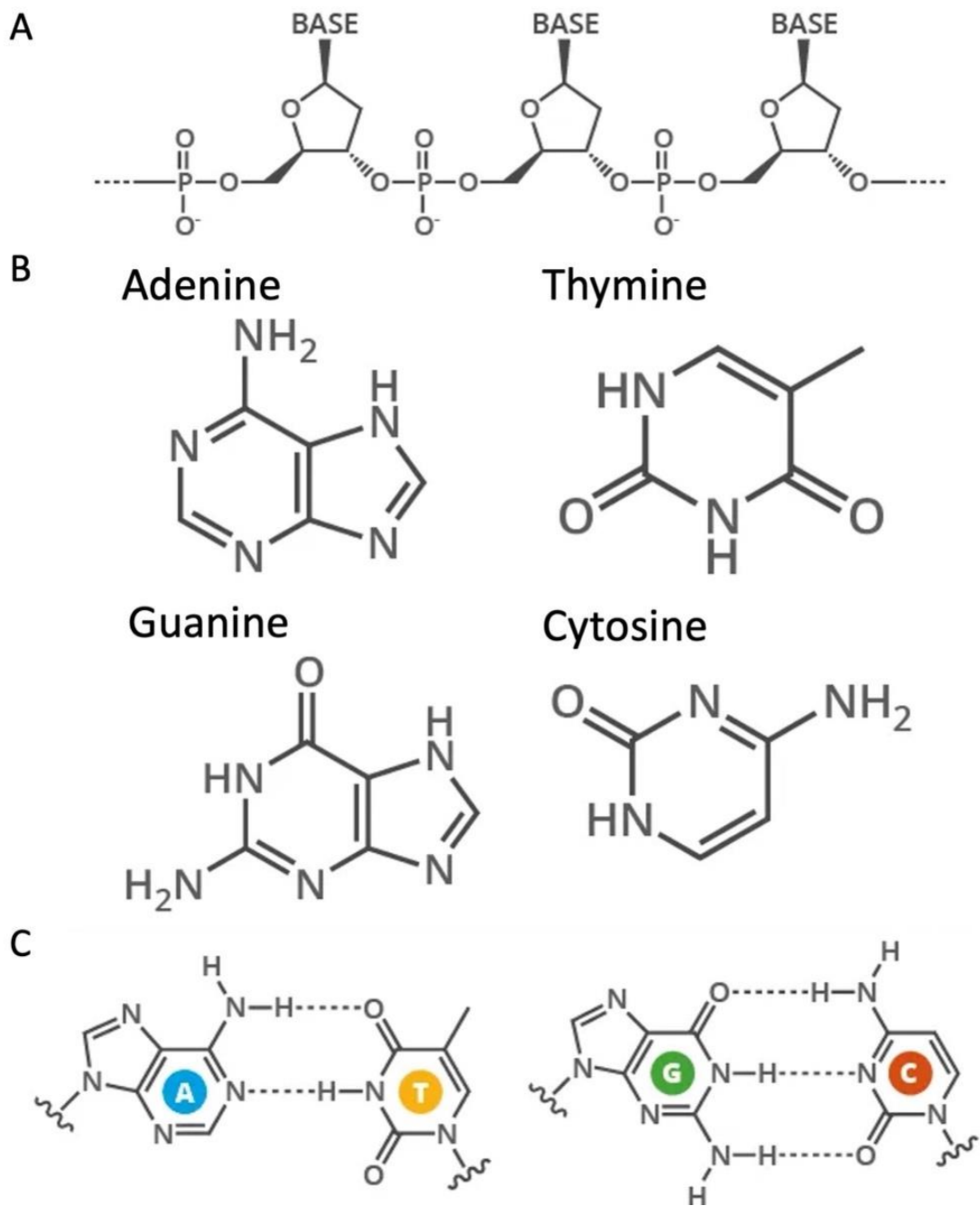


Figure 1.2 Chemical structure of DNA.

(A) DNA strands consist of an alternating pentose-sugar phosphate backbone and variable nitrogenous bases. (B) DNA consists of four possible bases, two of which are purines (left, adenine and guanine) and two are pyrimidines (right, thymine and cytosine). (C) Canonical Watson-Crick base pairing of DNA bases. Taken and adapted from¹⁵.

An organism's genome holds the information within the DNA sequence (or RNA for some classes of virus) to allow the cell to produce proteins needed for cellular survival. Functions include metabolic reactions for cellular maintenance and growth and DNA replication to accurately copy genomic DNA to pass on to progeny². In addition to gene coding regions, an organism will also have non-coding regulatory elements to ensure controlled gene expression. Regulation allows cells to acutely alter gene expression in response to both extra- and intra-cellular signals¹⁶.

Each species' genome will be tailored, through evolution, to allow the cell to survive within its particular niche. Genetic diversity within a population, potentially arising through mutation, may allow selection of new fitness traits in response to changes in the environment¹⁷⁻¹⁹. Here lies the delicate balance between preservation of genetic information and the ability to evolve to ever changing surroundings.

Genes conferring important function(s) are often conserved across multiple species. High levels of conservation with limited sequence difference suggest high importance in both protein function and pathways associated. Study of these highly conserved proteins can allow transferable knowledge to be obtained which can have beneficial implications in areas such as medicine and disease.

1.2 DNA replication

DNA replication is a conserved process which occurs across all forms of life. It is performed by actively dividing cells and comprises of three main stages initiation, elongation and termination²⁰. It is highly regulated in healthy cells, occurring only in circumstances of resource abundance²¹. DNA replication is semi-conservative resulting in two daughter duplexes containing a parental and a newly synthesised strand. To allow new strand synthesis to occur, double stranded DNA must first be prised apart. This poses a problem for the cell due to the exposure of single stranded DNA which may be bound inappropriately by other DNA binding proteins.

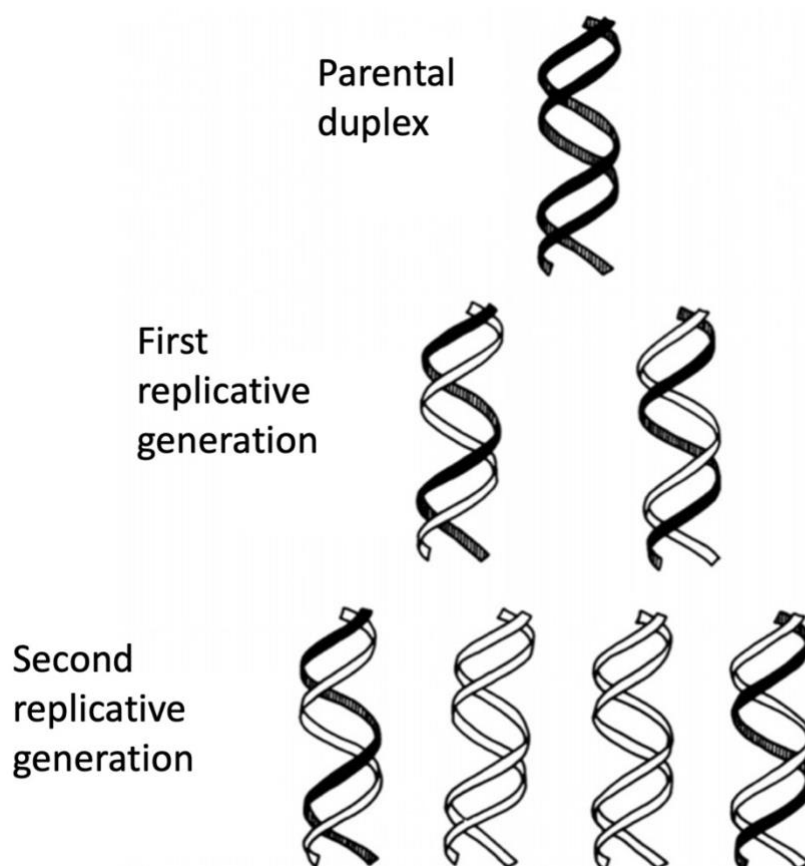


Figure 1.3 Semi-conservative DNA replication resulting in daughter duplexes which contain a parental strand (black) and a newly synthesised strand (white).

Taken and adapted from²².

DNA replication occurring semi-conservatively relies upon the presence of a contiguous parental backbone and complementarity between DNA bases²². Chemical and structural alterations due to exposure to damaging agents can impede metabolic processes such as DNA replication leading to cellular catastrophe. These alterations are major sources of DNA damage, repair of which will be discussed extensively below.

It is vitally important that replication occurs unhindered, ensuring faithful transmission of genetic material to daughter cells. Atypical cellular behaviour may be the result of 'fixed' alterations in the DNA sequence upon exposure to mutagens. This can have profound impacts, altering the phenotype of not only the cell but the whole organism²³. There are many overlapping repair pathways which act to resolve the DNA damage caused by mutagenic agents helping to maintain genomic stability²⁴. Inefficient coordination of these pathways can lead to the onset of diseases such as cancer in eukaryotes. Understanding DNA replication and its associated pathways is therefore key to combatting the development of disease and can allow early detection of aberrant cells²⁵.

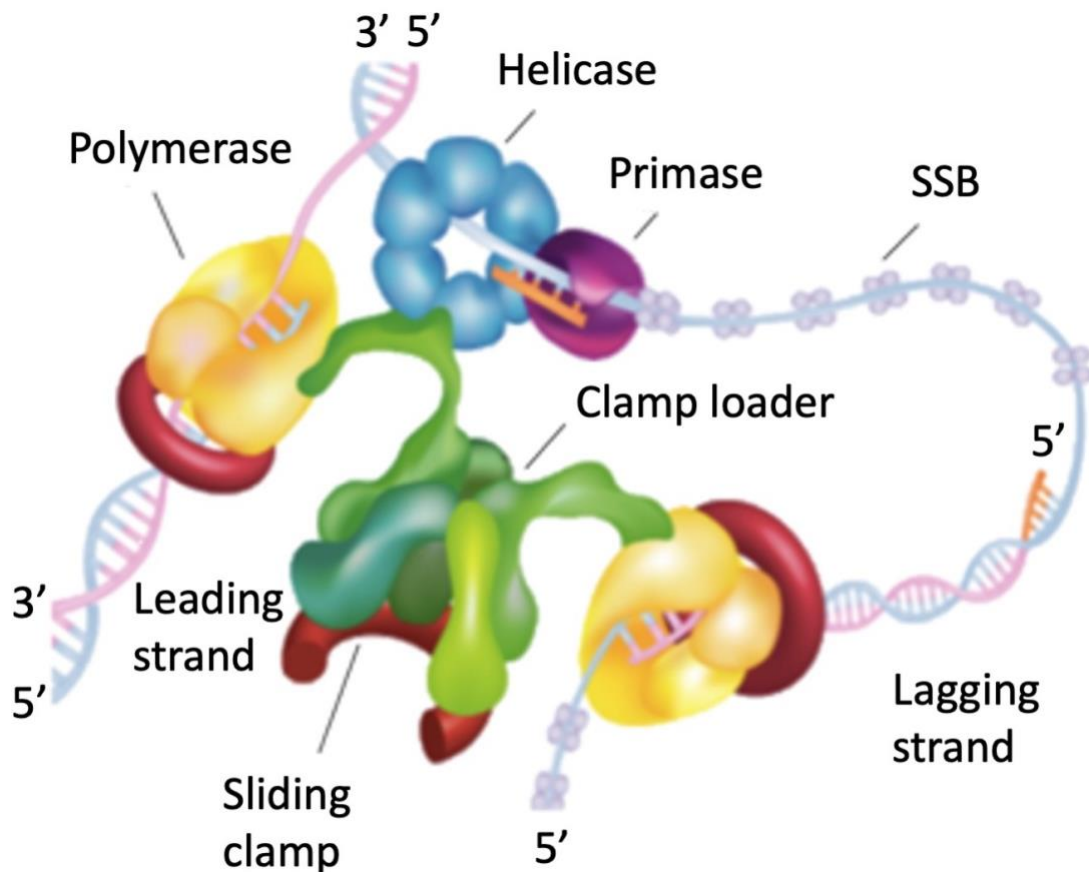


Figure 1.4 Basic diagrammatic representation of the bacterial replisome, highlighting key components.

Polymerase synthesis occurs with a 5' to 3' directionality. Taken and adapted from²⁶.

DNA replication is performed by the replisome, a large multi-protein complex (Figure 1.4). This protein complex contains six key components which are functionally conserved between all domains of life. Full assembly occurs after the initial loading of the replicative DNA helicase which unwinds double stranded (ds) DNA revealing exposed single stranded (ss) bases. Free bases of the leading strand are acted upon by the replicative DNA polymerase which synthesises nascent DNA in a 5'-3' fashion. Association between the polymerase and helicase is maintained by the clamp loader protein. The lagging strand is synthesised in a step-wise fashion due to the strict chemical restriction of DNA polymerases²⁷. Okazaki fragment synthesis is facilitated by the actions of the DNA primase, whilst single stranded binding proteins (SSB)

protect the DNA from nucleases and reduce inhibitory ssDNA secondary structures from forming^{20,28}. The interactions between proteins of the replisome and that of the individual protein complexes (such as that of the hexameric replicative helicase) increase processivity allowing rapid DNA replication to occur²⁹.

1.3 Replication stress

The processivity of the replicative polymerase and helicase are continually challenged during replication. Strong inhibitory obstacles can cause replicative stress, slowing the replisome (also known as replication fork stalling)³⁰. DNA lesions are particularly detrimental to fork progression as well as tightly bound nucleoprotein complexes, such as transcribing RNA polymerases, which can act as roadblocks to other DNA binding proteins³¹. Replication stress may also be a result of a dNTP pool imbalance causing the replicative polymerase to stall and potentially uncouple its activity for the helicase. Uncoupling in this manner may result in extensive stretches of ssDNA ahead of the polymerase, promoting instability and uncontrolled inappropriate access by other DNA interacting proteins³². Stalled replication forks, if left unresolved, are a major source of genomic instability³³. Collapsed forks may lead to the formation of double strand breaks, a particularly lethal form of damage³⁴. Recovery of stalled replication forks is a vital step for dividing cells due to the strict regulation of replisome assembly during the cell cycle. Resumption of DNA replication is also important for prokaryotic cells to ensure daughter cells contain the full complement of genomic DNA^{34,35}. The genomic instability caused by replicative stress and replisome stalling may lead to mutations in essential genes such as those involved in replication and DNA repair. This can lead to the manifestation of numerous diseases in eukaryotic cells such as cancer, cause developmental defects and be a contributing factor in aging³⁶.

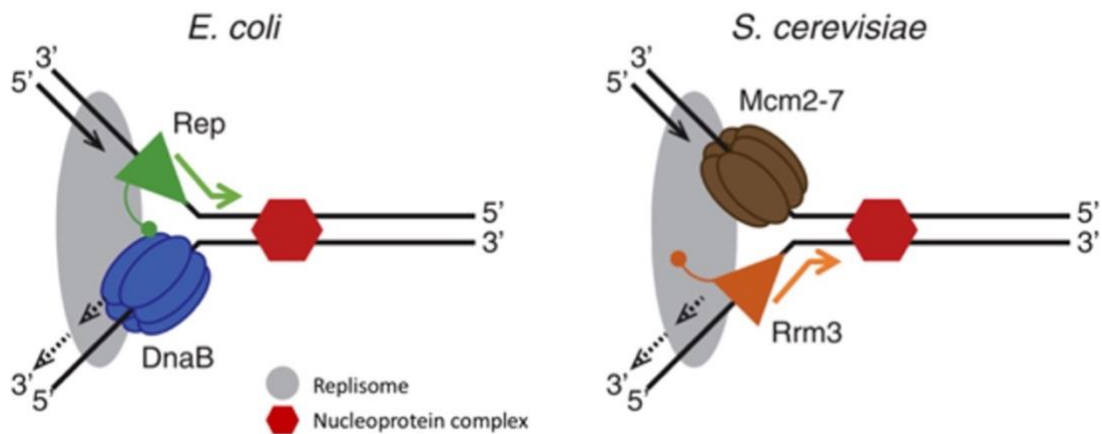


Figure 1.5 Example of an inhibitory obstacle poised downstream of the replisome complex.

Additional accessory helicases are described (Rep and Rrm3). Coloured arrows indicate directionality of accessory helicases. Taken and modified from³¹.

Additional proteins such as ‘accessory helicases’, which translocate alongside the replisome, can aid replicative progression by disrupting high-affinity noncovalent nucleoprotein complexes ahead of the replication fork. *E. coli* possess multiple redundant helicases to help alleviate replicative stress such as Rep and UvrD. The importance of this role is highlighted by the synthetic lethality of *repΔ uvrΔ* double mutant³⁷.

Due to the risk of genomic instability caused by fork stalling and collapse, pathways exist to allow the replisome to bypass certain types of damage. These responses are collectively known as ‘DNA damage tolerance’ (DTT). These pathways include template switching and translesion synthesis (TLS)³⁸.

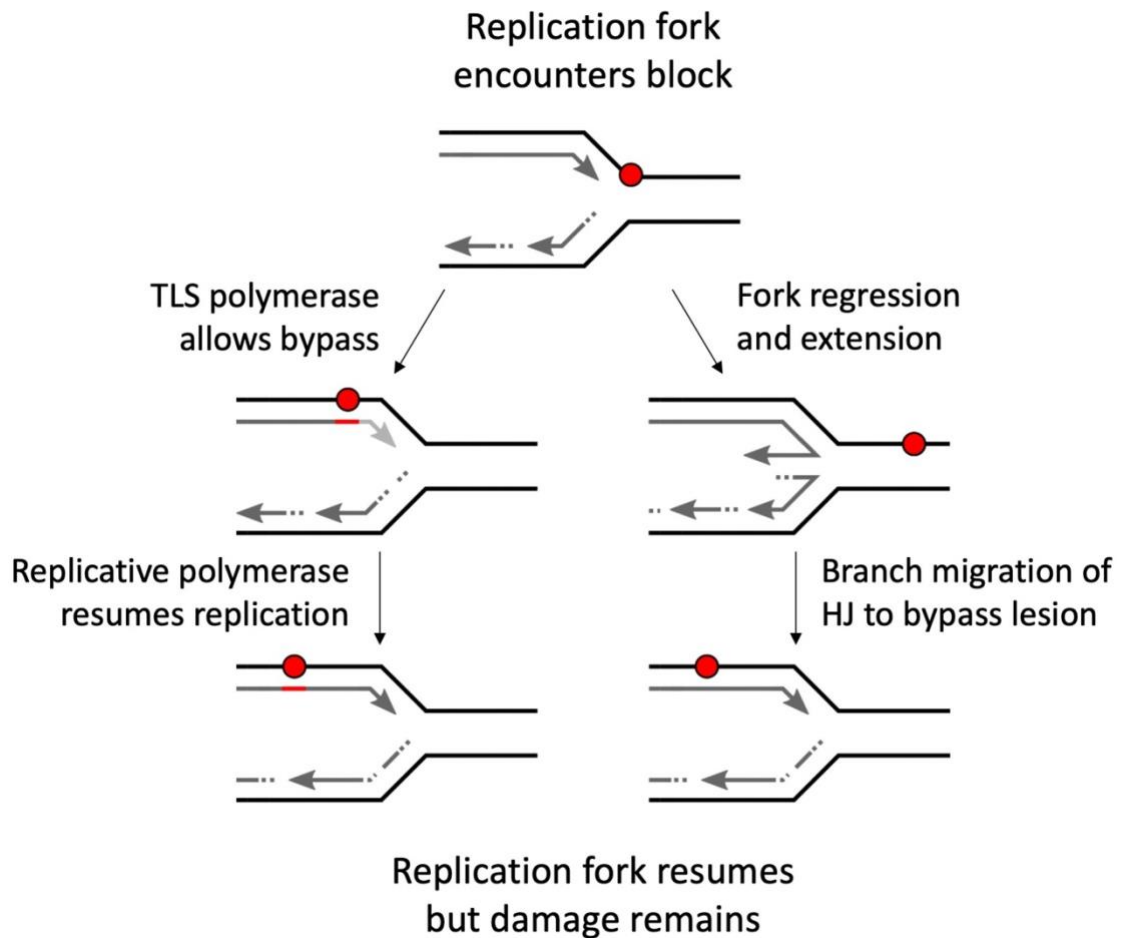


Figure 1.6 DNA damage tolerance pathways alleviate replicative stress caused by inhibitory lesions.

DNA damage (red circle) is bypassed using translesion synthesis (TLS, left) or by fork remodelling and branch migration (right). DNA repair pathways are needed to repair damage after the replication fork has progressed passed the inhibitory lesion.

For template switching to occur, the forked DNA structure must first be remodelled to allow loading of the replisome onto undamaged DNA strands and access to the damage by repair proteins. In *E. coli* template switching is performed by a myriad of repair proteins which may include RecG, PriA and the RecBCD and RecFOR protein complexes^{35,39-42}. These proteins make up specific DNA repair pathways which are deployed dependent on the type of damage encountered. DNA damage bypass by

template switching relies on the activities of helicase proteins to regress and remodel DNA as shown in Figure 1.6.

In contrast, translesion synthesis utilises alternative polymerases which are able to accommodate damaged DNA bases within their active site. These polymerases, which are lower in processivity and lack proofreading functions, allow damage bypass by temporarily replacing the replicative polymerase³⁸.

Although these systems allow replication fork progression, DNA damage is still often present. Additional repair pathways exist which act to repair these lesions post replication, some of which will be described below.

1.4 DNA damage

DNA within cells is continually exposed to mutagenic agents which can cause abnormal chemical alterations to DNA structure. Mutagens can originate from a variety of sources including reactive oxygen species (ROS, exogenous) or from radiation such as X-rays and UV light³. The resulting lesion is dependent on the type of exposure, as can be seen in Figure 1.7. DNA damage can act as a structural obstacle for the replisome which may lead to replication fork slowing/stalling or incorporation of wrongly paired bases⁴³. Cells possess multiple overlapping repair pathways to act upon sites of DNA damage to limit the chances of persisting mutations and to aid not only in DNA replication but other processes too, such as transcription. These repair systems are highly conserved and span across the prokaryotic/eukaryotic evolutionary border⁴³.

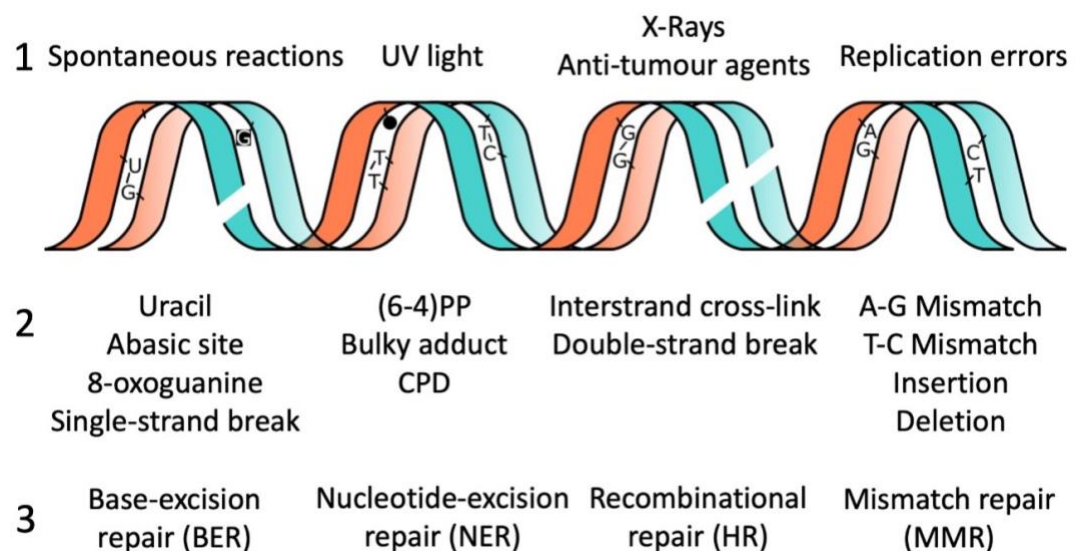


Figure 1.7 Potential sources of damage (1), resulting DNA lesions (2) and most likely repair mechanism (3).

Of particular interest to this work is DNA damage resulting in covalent linkages to DNA bases (DNA adducts), chemical alterations to the DNA bases (such as cytosine deamination) and DNA strand breaks. Repair is elicited by highly specialised DNA repair pathways, all of which utilise the activities of DNA helicase and glycosylase repair proteins which are the main focus of this study.

1.5 DNA repair systems

1.5.1 A brief history of DNA glycosylases with mechanistic insights

The importance of DNA repair and glycosylase enzymes is exemplified by the winners of the 2015 Nobel prize for chemistry. This was awarded to three scientists who led the work in identifying the enzymatic mechanisms of three DNA repair pathways as I will discuss below. Tomas Lindahl demonstrated the instability of DNA under physiological conditions leading to 'decay' (damage) and identified the first DNA glycosylase, the *E. coli* uracil-DNA glycosylase UNG, which repairs damage caused by cytosine deamination^{44,45}. Both Aziz Sancar and Paul Modrich further expanded our understanding of orchestrated DNA repair processes, detailing the steps and enzymes involved in repair of UV-induced thymidine dimers and the targeted removal of mismatched DNA bases respectively^{46,47}.

Since their initial discovery, a plethora of DNA glycosylases have been discovered which can be subdivided into distinct functional superfamilies (SFs). The SF which is most relevant to this study is the uracil DNA glycosylase (UDG) superfamily.

As the name suggests, the UDG family are a group of monofunctional glycosylases which specifically target and remove uracil (and uracil analogs) from DNA^{48,49}. Uracil in DNA can occur through spontaneous cytosine deamination, the misincorporation of dUMP during replication, or due to exposure to ROS^{48,50-52}. UDG family proteins can be divided further into six subfamilies dependent on substrate specificity, conservation of active site residues, and enzymatic efficiency. The most highly processive subfamily, to which *E. coli* UNG is a part of, breaks the glycosidic bond via

nucleophilic attack of an activated water molecule. This occurs through a conserved aspartic acid residue, residing within its 'motif I', acting as a general base⁵³⁻⁵⁵.

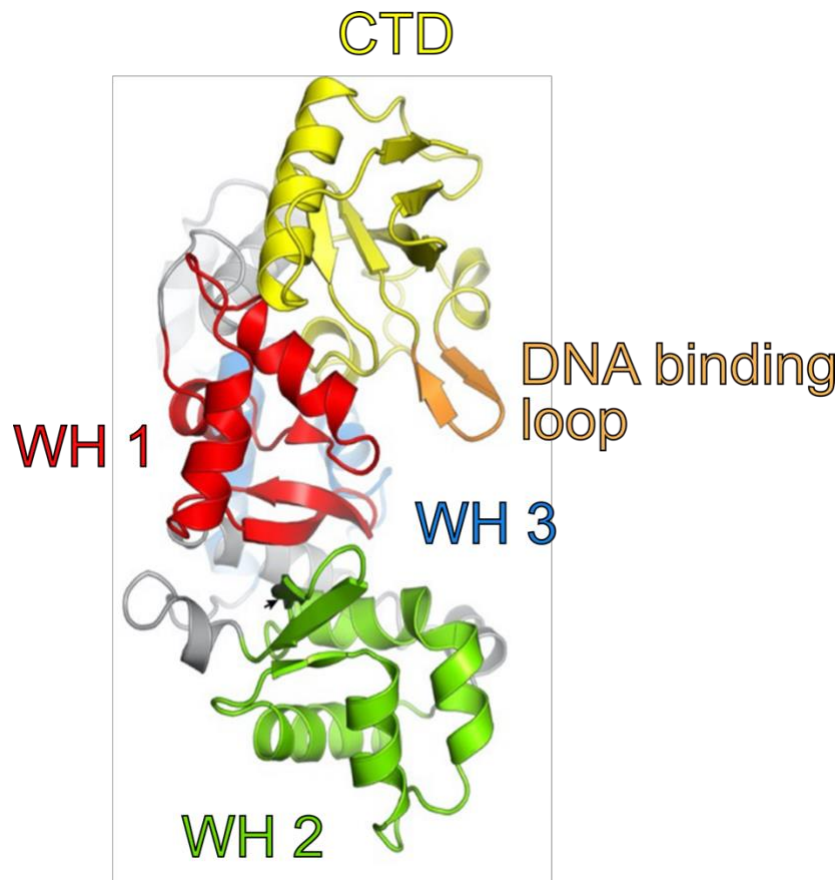


Figure 1.8 Characteristic domain organisation of HTH_42 superfamily proteins.

Domain structure of *Streptomyces sahachiroi* AlkZ highlighting C-shape topology. AlkZ consists of a series of winged-helix (WH) motifs (**red, green and blue**), a C-terminal domain (CTD, **yellow**) containing a DNA binding β -loop (**orange**) and connecting helices (**grey**). Take and adapted from⁵⁶.

Recent discovery and characterisation of glycosylases belonging to a novel superfamily of DNA binding proteins, the helix-turn-helix (HTH_42) superfamily, are of particular interest to this study. This SF can be subdivided into proteins which are located within bacterial biosynthetic gene clusters (BGCs) and those which provide a general role in genome protection against exogenous DNA damaging agents⁵⁷. The prior subfamily, termed 'AZL' (AlkZ-like), provide self-resistance to genotoxic alkylating agents which

may be produced by the cell to gain a competitive advantage over adjacent microbes⁵⁶. The latter group of proteins are conserved in genomic context but lie distinct from BGCs, potentially providing roles in a yet to be identified repair pathway. These YcaQ-like (YQL) proteins may have a wider range of damaged DNA substrates and may have roles in wider cellular processes such as cell wall biosynthesis and transformation competency⁵⁸.

The structure of the HTH_42 superfamily is highly conserved, depicting a characteristic 'C-shape' formed by tandem winged helix-turn-helix motifs (Figure 1.8). Enzymatic function relies upon an essential QΦQ motif/QΦD motif (with Q representing glutamine, D representing aspartic acid and Φ, an aliphatic residue). This catalytic motif is located within 'WH 1' and is aided by the DNA binding β-hairpin which together orient the DNA substrate. This allows activation of a water molecule by the proximal glutamine and subsequent nucleophilic attack, similar to UDG super family glycosylases⁵⁶.

1.5.2 Repair of DNA chemical alterations

Lesions which are the result of small chemical alterations to DNA are repaired by the Base Excision Repair pathway (BER)^{59,60}. Damage often occurs through endogenous sources or as a result of the spontaneous decay of DNA manifesting as deamination, oxidation or methylation of DNA bases⁴⁵. The BER pathway was first discovered through the identification of an *E. coli* uracil-DNA glycosylase (Ung) which acted upon genomic uracil as a result of the deamination of cytosine⁴⁴. Glycosylases excise damaged bases through base flipping and cleavage of the *N*-glycosidic bond between the sugar and base of DNA. This leaves behind an abasic site (AP-site)⁵⁹. This 'lesion intermediate' then requires further processing by additional repair enzymes such as an AP-endonuclease, an exonuclease, a DNA polymerase and a ligase⁵⁹. These proteins in combination form the BER pathway.

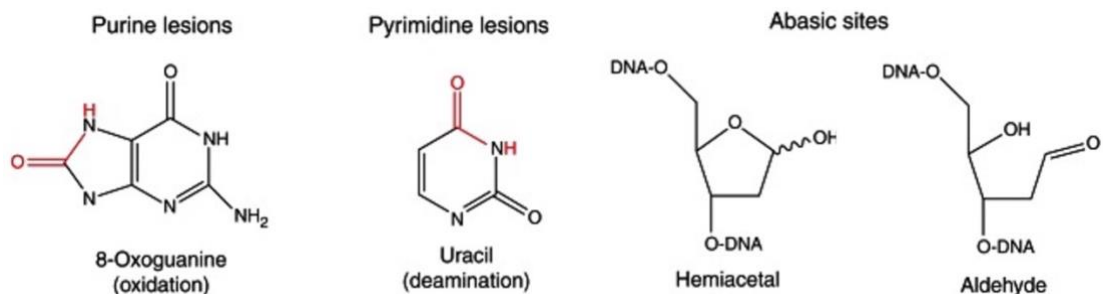


Figure 1.9 Examples of purine and pyrimidine lesions repaired by BER and potential resulting abasic sites.

Taken and adapted from⁵⁹.

Our understanding of DNA glycosylases and the BER pathway has since expanded dramatically detailing it as a well conserved form of repair. Glycosylases' role in maintaining genomic stability is exemplified through the synthetic lethality and predisposition to cancer presented through gene knockout studies in mice^{61–63}.

DNA damage through cytosine deamination or guanine oxidation (see Figure 1.9) pose little threat to components of the replisome^{59,64}. These types of damage alter the bases hydrogen bonding capacity and if left unrepaired, mutations can arise through mispairing of incorporated bases by the replicative polymerase. Mutations manifest as C:G to T:A transitions (U:A mismatch) and G:C to T:A transversions (G:A mismatch)⁵⁹. It is therefore vitally important that proteins are able to seek out and repair damaged bases without needing recruitment by stalled replisome associated DNA damage repair signals.

1.5.3 Repair of bulky DNA adducts

Damages which cause distortion of the DNA helix are repaired by the **Nucleotide Excision Repair** pathway (NER). These lesions, such as pyrimidine dimers (due to UV exposure), pose a greater challenge to components of the replisome as well as other DNA metabolism complexes. The NER pathway is able to resolve a myriad of diverse forms of DNA damage, activated by damage seeking proteins such as UvrA and UvrB in *E. coli*, or in response to RNA polymerase stalling during transcription⁶⁵⁻⁶⁹. Due to the focus of this thesis being replication-coupled DNA repair, NER pathways in the context of transcription-coupled repair will not be discussed. NER is a functionally conserved process across bacteria and eukaryotes, with pathways in archaea sharing homologs to both^{65,70}. Repair pathway steps echo those performed in BER however, repair involves the incision of the DNA backbone adjacent to damage and subsequent endonuclease activity. This results in DNA gaps as opposed to an AP site meaning full repair requires 'gap filling' protein functionality.

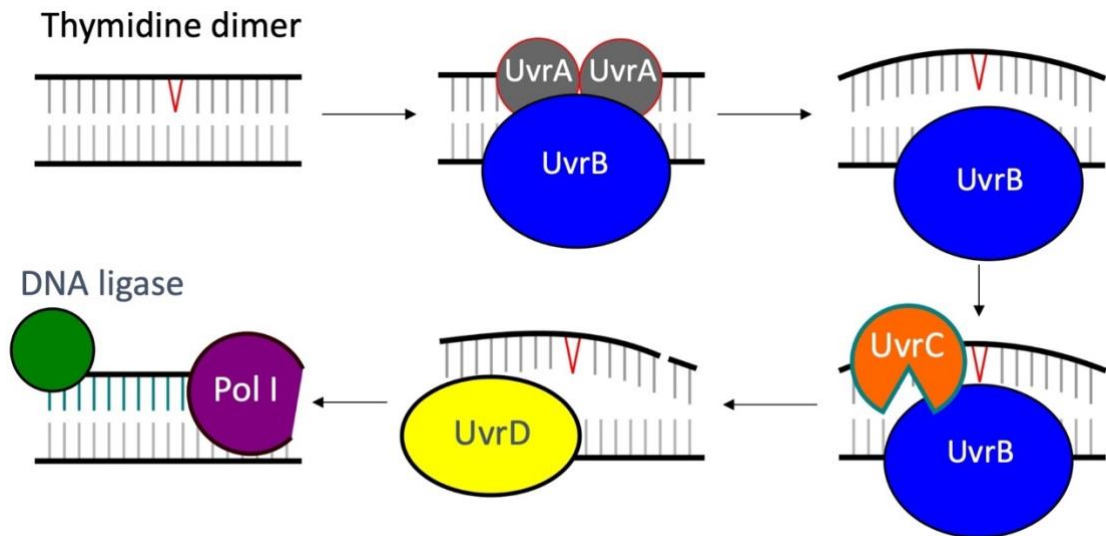


Figure 1.10 Nucleotide excision repair pathway in *E. coli* following UV damage.

Damage is bound by UvrAB, UvrA dissociates and UvrB causes local unwinding. UvrC is recruited and cuts the DNA backbone. UvrD displaces UvrBC and removes damaged DNA. The gap is filled by DNA Pol I and repair is completed by DNA ligase.

NER in *E. coli* is performed by the UvrABC system. DNA damage is first recognised by a UvrA dimer which recruits the DEAD-box superfamily 2 (SF2) helicase UvrB⁶⁵. UvrB displays poor unwinding potential and instead causes local unwinding following DNA binding and UvrA dissociation^{71,72}. This allows access and activity by UvrC which incises nucleotides either side of the DNA lesion⁶⁵⁻⁶⁷. Removal of UvrBC and the damaged DNA is achieved through the action of the SF1 DNA helicase UvrD. Following removal, DNA polymerase I and a DNA ligase complete the repair process^{65-67,73}.

1.5.4 Repair of mismatched bases

The **Mismatch Repair (MMR)** system is a post-replicative repair pathway which seeks out erroneously paired bases and small insertion/deletion loops caused during DNA replication⁷⁴⁻⁷⁶. Disruption in genes associated in MMR is directly linked to an accumulation of mutations which can lead to the development of diseases such as cancer in higher order organisms⁷⁴⁻⁷⁷. The steps involved for repair follow closely to that used in NER. In *E. coli*, damage recognition is performed by a MutS homodimer which scans along dsDNA seeking mismatched bases. Binding occurs asymmetrically, with one MutS binding the mismatch and the other stacking into the DNA duplex to elicit a 60° kink⁷⁸. MutL is then recruited and forms further contacts to the DNA, promoted by MutS. In some organisms MutL is able to encircle DNA and produce an incision to allow removal of the mismatched base^{77,79}. In *E. coli*, DNA nicking activity is performed by the endonuclease MutH^{76,79}. After MutH recruitment by MutL, MutH bidirectionally scans the DNA seeking a hemi-methylated d(GATC) site. During replication the new daughter strand is void of adenine methylation allowing distinction between strands. Following incision, UvrD is recruited to displace the damaged strand and repair is completed similarly to NER^{76,79}. This highlights the overlap and often redundancy of DNA repair proteins which are able to be deployed in a variety of pathways.

1.5.5 Homologous recombination (HR)

HR events rely on the presence of a repair template to occur and are employed in a variety of circumstances such as DNA repair, DNA replication, and telomere maintenance⁸⁰. In eukaryotic cells these events are limited to the synthesis (S) and growth 2 (G2) phases of the cell cycle due to the presence of the second newly replicated DNA repair template⁸¹. Genetic recombination is also present in eukaryotic cells, involving a slightly different protein network used to promote genetic diversity during crossover events in meiosis⁸². HR is the preferred process to an opposing repair system known as non-homologous end joining (NHEJ) as HR is less-mutagenic and so helps to preserve genetic information⁸¹. HR may occur through classical **double-strand break repair (DSBR)**, or through **synthesis-dependent strand annealing (SDSA)** and **break-induced replication (BIR)** sub-pathways⁸³.

HR is utilised to repair DNA lesions such as **interstrand crosslinks (ICL)**, a particularly toxic type of damage, which may form after exposure to ionising radiation (e.g. X-rays) or through the action of certain anti-tumour reagents⁴. The resulting covalent structural bridge has the ability to inhibit essential processes such as DNA replication and transcription⁸⁴. Double strand breaks may also arise as a result of replication through a single-strand nick. This form of repair requires additional replisome repriming/reactivation steps^{40,85,86}.

Repair through HR, as displayed in Figure 1.11, can be divided into three major steps, pre-synapsis (end resection and strand invasion), synapsis (strand exchange and branch migration) and post synapsis (Holliday junction (HJ) resolution).

Initial processing is performed by the RecBCD complex in *E. coli* to reveal regions of ssDNA^{87,88}. In eukaryotes and archaea end resection is performed by a myriad of proteins initiated by a conserved MRN-CtIP complex. This is further processed by EXO1, DNA2 and BLM repair proteins in eukaryotic cells or NurA-HerA in archaea^{70,89}.

The exposure of ssDNA allows binding by repair proteins such as SSB in *E. coli* and RPA in eukaryotes which serve to protect ssDNA, limiting secondary structure formation and elicit a wider recruitment signal for additional repair proteins⁹⁰⁻⁹². Recruited proteins may then amplify the DNA damage-repair signal further, leading to cell cycle arrest in eukaryotes and activation of the SOS response in bacteria⁹³⁻⁹⁵. SSB and RPA protective proteins are then replaced by recombinases which oligomerise onto exposed ssDNA forming a nucleoprotein filament. This catalyses homology search and strand invasion events and is performed by RecA in bacteria, RadA in archaea and Rad51 in eukaryotes^{96,97}.

Recombinase protein loading is aided by the Rad51 paralogue proteins which are structurally similar but have developed distinct functions⁹⁸. These proteins promote genomic integrity through roles in DNA damage signalling⁹⁹, HR, replication fork restart, mitochondrial genome stability and in telomere maintenance¹⁰⁰. Strand invasion orchestrated by these proteins results in the formation of a displacement loop (D-loop) which migrates in search of homologous regions within the invaded DNA template¹⁰¹⁻¹⁰³. The exact involvement of Rad51 paralogues and how D-loop migration

is controlled are still not fully resolved. We speculate that additional helicase motor proteins such as Lhr, introduced extensively in section 1.7, may confer an important role.

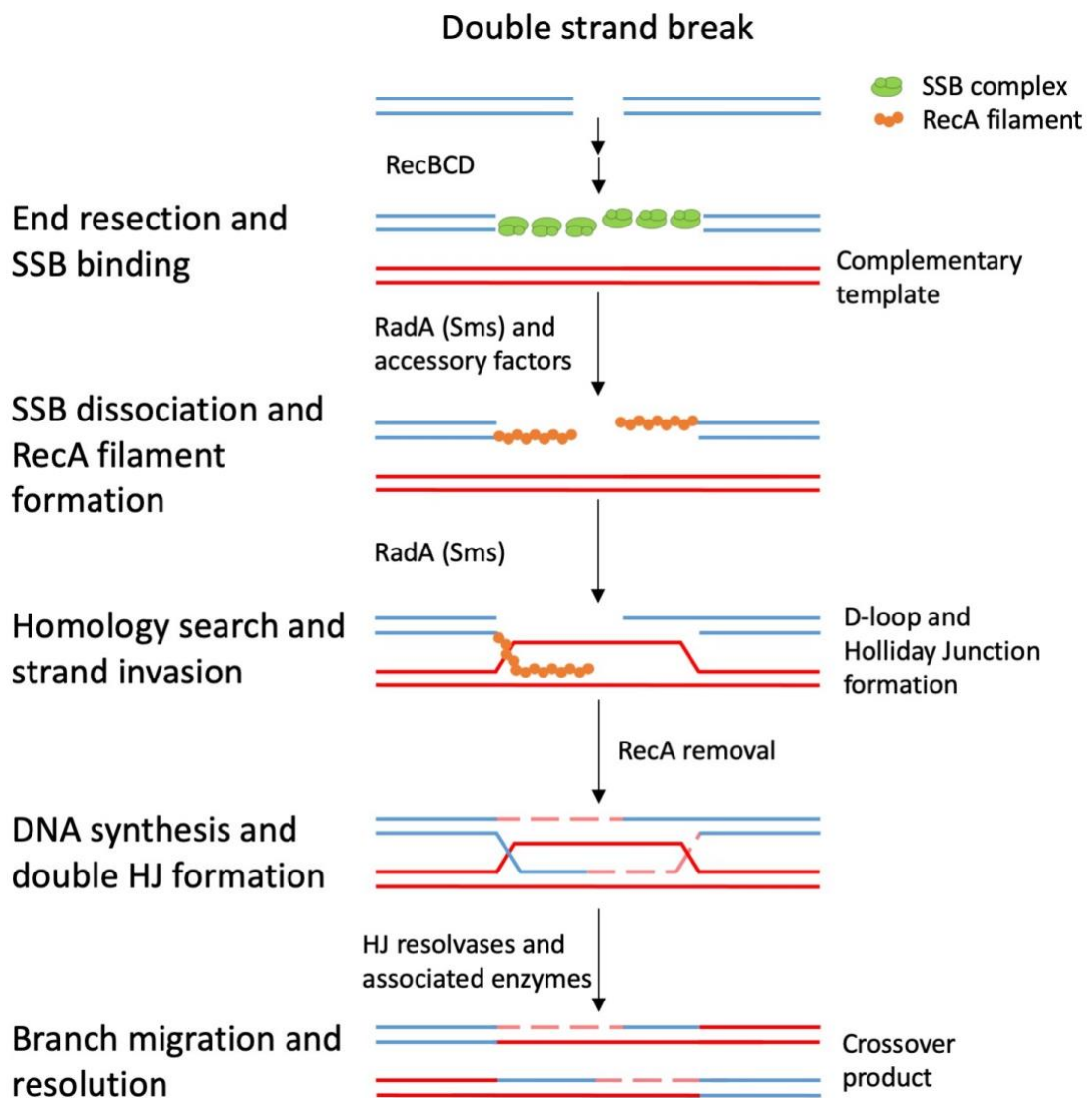


Figure 1.11 Simplified overview of double strand break repair via homologous recombination in *E. coli*.

Protein key players highlighted and possible DNA products suggested. Non-crossover products also possible.

Upon reaching homologous regions, the D-loop is extended by a DNA polymerase to form a HJ which can then be resolved (see Figure 1.11)^{80,104}. Resolution in *E. coli*

requires the action of the RuvABC resolvosome. RuvAB-mediated HJ branch migration allows scanning and cleavage by RuvC at preferred sites which may result in crossover or non-crossover products¹⁰⁴.

1.5.6 Rescue of stalled replication forks

For recovery of stalled replication forks, extensive remodelling is required using the functionalities of accessory DNA repair helicases. This may result in fork reversal into a 'chicken foot' intermediate allowing access to the inhibitory lesion for repair as displayed in Figure 1.12, left³⁵. In bacteria PriA, a DEXH-type 3' to 5' helicase, is essential for replication fork reassembly and restart required for repair by BIR and RDR^{33,103,105,106}. These examples highlight the fundamental importance of DNA helicase enzymes in maintaining genome stability and ensuring cellular survival.

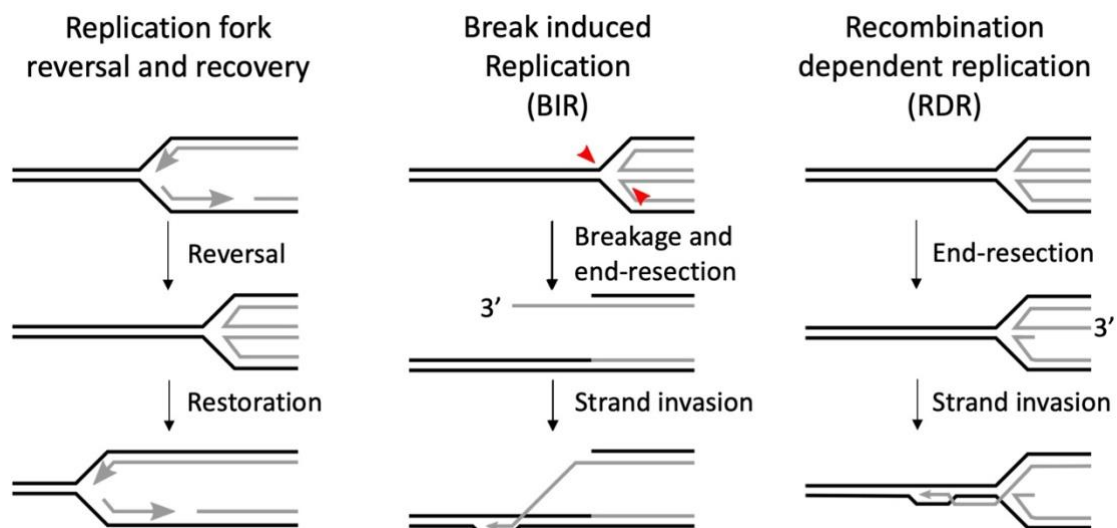


Figure 1.12 Rescue of stalled replication forks may occur through distinct pathways dependent on lesion encountered.

Examples of replication fork recovery. Each example requires DNA remodelling and processing, highlighting the essential function of DNA helicases.

Numerous Ski-2 like DNA helicases have been shown to play important roles in alleviating replicative stress such as Hel308 in archaea, its human homolog HelQ and the bacterial protein Lhr¹⁰⁷⁻¹¹⁰. It is known that these and other helicase proteins promote genome stability and fork restart but their actions are yet to be refined biochemically. By purifying examples of these repair proteins along with interacting protein partners, we aim to shed light on the functions of these proteins to promote fork restart and in homologous recombination.

1.6 DNA Helicases

Helicases and nucleic acid translocases play a central role in all aspects of DNA metabolism. These proteins belong to a larger AAA+ superfamily of protein molecular motors, assigned by their functional RecA-like domains¹¹¹. These domains couple the energy released through adenosine triphosphate (ATP) hydrolysis to conformational changes in protein structure to allow translocation along DNA, often unwinding or remodelling the DNA as they move¹¹². Helicase unwinding occurs through disruption of hydrogen bonds between paired bases¹¹. Movement typically occurs in a directional manner (5' to 3' or 3' to 5') which is referred to as the proteins 'polarity'. Not all nucleic acid translocases have strand separation functionality but they do share NTP-dependent biased directionality along ss- or ds- nucleic acid substrates¹¹³. In addition to nucleic acid remodelling, helicases are capable of displacing strongly inhibitory nucleoprotein complexes as mentioned in section 1.3^{31,114}. Helicase function is strongly dictated by additional domains which fold around the 'helicase unwinding core'. Conservation of core ATPase and helicase domains as well as functional appendages allow further classification into distinct super-family groups¹¹⁵⁻¹¹⁷.

1.6.1 Superfamily-1 and superfamily-2 helicases

Helicase proteins can be subdivided into 6 SF groups dependent on conservation within the Walker A and Walker B motifs responsible for ATP binding and hydrolysis^{11,118}. Subdivision can be extended further between toroidal ring forming (SFs 3-6) and non-ring forming (SF1/2) and then into subfamilies through specific traits such as polarity, nucleotide triphosphate usage (such as ATP), substrate preference or mechanistic features¹¹⁵.

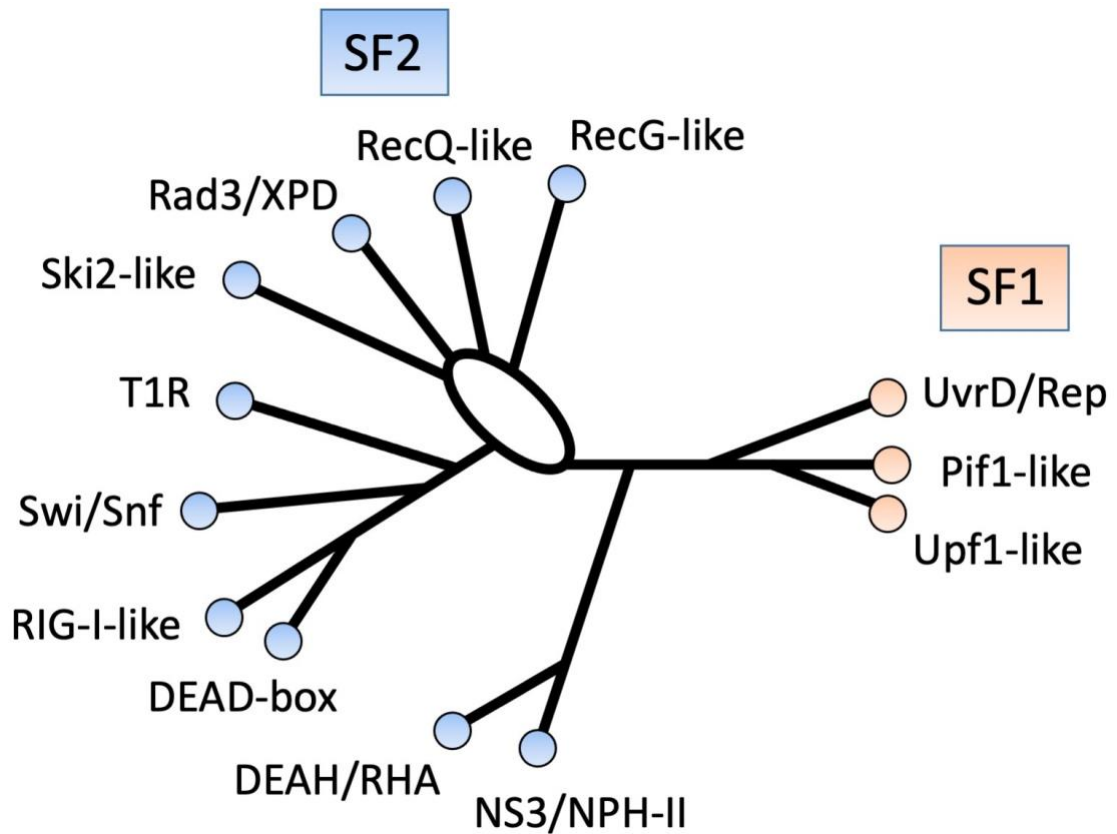


Figure 1.13 The subfamilies of SF1 and SF2 helicases.

SF1 (right, **orange**) can be subdivided into three subfamilies, whilst SF2 (left, **blue**) can be divided into 9. Each subfamily is distinct in function and grouped through sequence homology.

When present, SF1 and SF2 families share considerable conservation between twelve identified signature motifs located within the RecA1 and RecA2 helicase-core domains. These features include residues which coordinate ATP binding and hydrolysis. Great diversity can be seen in the proximal domains situated around the helicase-core which allows further subdivision as shown in Figure 1.13.

1.7 Lhr

Lhr (Large helicase-related) is an SF2 ATP-dependent DNA helicase which was first discovered in bacteria due to it being the longest known protein in *E. coli*^{119,120}. It is highly conserved among bacteria and archaea^{116,121} and has sequence homologues in eukaryotes including humans^{122–124}. Recent work has identified sequence and structural homology of Lhr's extended C-terminal domain (CTD) to that of a newly discovered superfamily of proteins, the helix-turn-helix (HTH_42) superfamily (as introduced in 1.5.1)^{110,125}. Lhr belongs to 'Class II' of this superfamily and does not appear to be associated with any BGCs. This HTH_42 extension, as seen in many Lhr proteins, is yet to be characterised biochemically and its biological role remains in speculation⁵⁷.

1.7.1 Initial discovery

Initial study of Lhr in *E. coli* yielded little insight into protein activity or an associated repair pathway. Reuven *et al.* reported no detectible protein ATPase activity and were unable to show sensitivity in *lhr* knockout strains when exposed to UV or H₂O₂. Additionally, growth was unaffected when *lhr* was knockout out in combination with *recA*, *recB*, *recD*, *uvrB*, *uvrD*, or *rep* repair proteins¹¹⁹. *Lhr* was determined to be a non-essential gene situated downstream of *rnt* coding for the tRNA-processing enzyme RNaseT, potentially sharing the same promoter. RNaseT may also have a hand in DNA repair pathways¹¹⁹.

1.7.2 Genetics give clues to function

Genetic studies of *lhr* by Susan Lovetts' group reported the first genetic phenotype. *Lhr radA* dual knockout strains showed sensitivity to azidothymidine (AZT), a chemical which causes replication strand termination leading to ssDNA gaps and DSB, suggesting synergistic function¹. AZT has been shown to promote the template switching repair pathway in *E. coli*, but no link was discussed here¹²⁶. Cooper *et al.* also reconfirmed the little observed sensitivity when Δlhr cells were grown after UV irradiation¹. *E. coli* RadA (Sms) is a Rad51/RecA family protein which will be presented briefly in Chapter 3. An additional genetic phenotype has been observed in *Sulfolobus islandicus* when *lhr* (*SiRe_1605*) knockout strains were grown in the presence of methyl methanesulfonate (MMS)¹²⁷. MMS is a potent DNA alkylating agent which causes DNA replication inhibition through production of *N*³-methyladenine¹²⁸.

Further development of Lhr's role within the cell was achieved through monitoring its changes in expression level during exposure to multiple genotoxic agents. *Mycobacterium tuberculosis* Lhr expression levels were shown to increase at least 2-fold in response to UV irradiation and when grown in the presence of mitomycin C¹²⁹. Lhr levels were largely unaffected in *recA* deficient cells in *M. tuberculosis*¹³⁰ or in *Sulfolobus acidocaldarius* (*Saci_1500*, protein now termed 'Lhr1'¹³¹) upon exposure to UV. However, cell survival rates were greatly reduced in $\Delta saci_1500$ and $\Delta saci_1500 \Delta uspE$ strains suggesting a potential role in downstream DNA repair pathways following damage by UV¹³². Song *et al.* highlighted increased expression of Lhr from *Sulfolobus islandicus* (*SiRe_1605*) when grown in the presence of MMS. They further suggest a possible role for Lhr in controlling gene transcription due to the down-regulation of many genes involved in nucleotide metabolism, DNA repair, and cell

division in MMS exposed $\Delta SiRe_{1605}$ cells¹²⁷. This relationship may be due to the inactivation of the DNA repair signals associated with Lhr's repair pathway(s) causing indirect reduction of gene transcription.

More recently a second Lhr (aLhr1) protein was discovered in *S. acidocaldarius* (Saci_0814). Genetic study showed a 5-fold decrease in HR frequency suggesting Lhr's direct involvement¹³¹.

1.7.3 Protein structure and domain organisation

Variable phenotypes between organisms suggest the Lhr family of helicases are a diverse group of proteins with multiple roles in DNA damage repair and are able to influence a wide range of metabolic pathways. Functional variability is supported by its highly interchangeable C-terminus as highlighted in Figure 1.14.

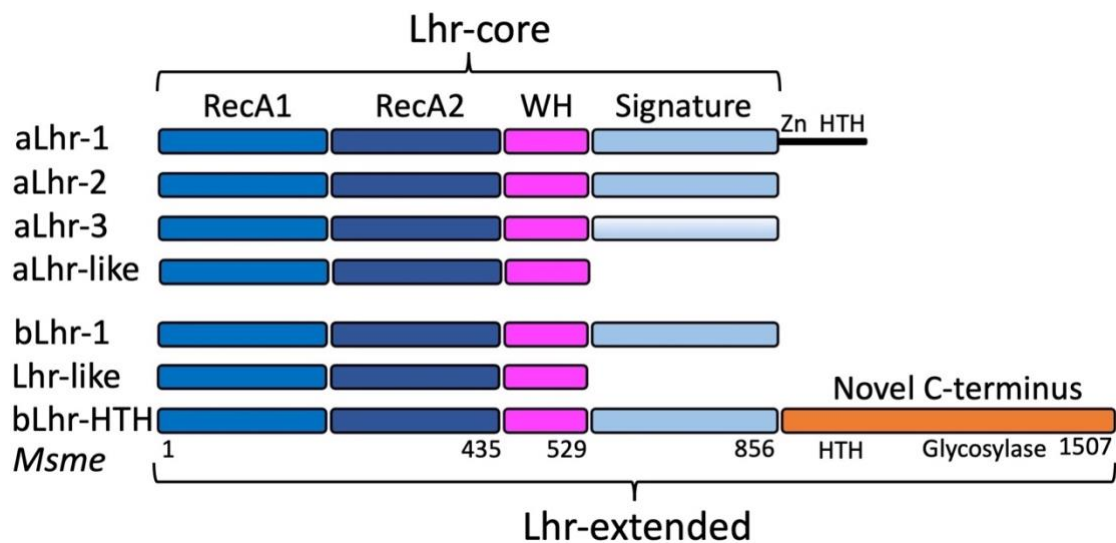


Figure 1.14 Domain organisation of Lhr proteins from archaea and bacteria.

Lhr proteins contain an 'Lhr-core' consisting of two RecA-like domains (RecA1 and RecA2), a winged helix domain (WH) and a 'signature domain' of unknown function. aLhr-3 displays a deteriorated signature domain whilst aLhr-1 contains proximal Zn-finger like and helix-turn-helix (HTH) motifs. Further diversity is seen with bLhr-HTH or 'Lhr-extended' which possess a yet to be characterised novel C-terminal domain. *Mycobacterium smegmatis* (*Msme*) bLhr-HTH annotation gives context of domain length.

All Lhr family proteins contain an 'Lhr-core' consisting of two RecA-like domains responsible for DNA binding and ATP hydrolysis, a winged helix domain (WH) reminiscent of Hel308/HelQ helicases, and a signature 4th domain of unknown function. The WH domain from *MthHel308* stabilises DNA binding through interactions with duplex unwound DNA and promotes ATPase/helicase activity via

contacts to the RecA-like domains¹³³. It is possible that the WH in Lhr serves a similar purpose but it is as yet unconfirmed. As reported by Hajj *et al.*, a great degree of diversity is seen in domain structure after the WH domains¹²¹. This has allowed further classification of Lhr and Lhr-like proteins (see Figure 1.14). Of particular interest is the novel C-terminus of bLhr-HTH 'Lhr extended'. *Mycobacterium smegmatis* Lhr's CTD was recently determined using cryo-EM (PDB: 7LHL) depicting a C-shaped structure consisting of six tandem WH's and a proximal β -barrel (shown in Figure 1.15)¹²⁵.

M. smegmatis 'Lhr-core' was also determined (PDB: 5V9X) showing similarities to *Archaeoglobus fulgidus* Hel308¹³⁴. Here the WH domain lies distinct between the two structures however, different DNA substrates are bound by each protein which may invalidate direct comparison. Mutational analysis showed residues Arg279 (RecA2) and Trp597 (domain 4) couple ATPase hydrolysis to mechanical translocase activity, and Thr145 (RecA1) and Ile538 (domain 4) are responsible for duplex DNA unwinding¹³⁴.

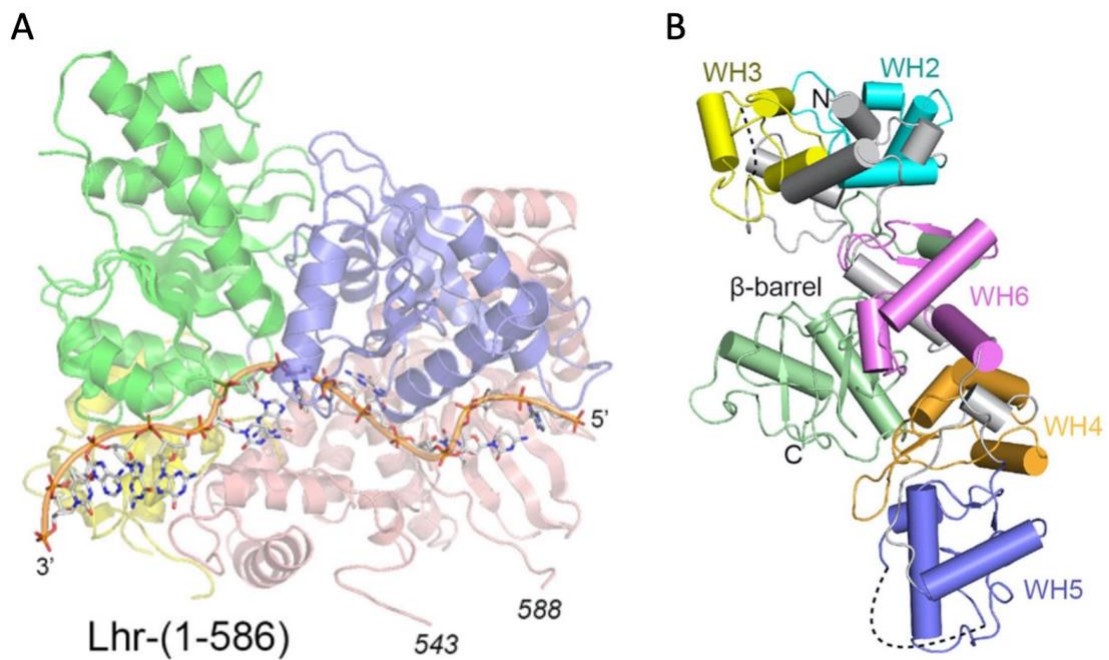


Figure 1.15 Lhr-extended comprises of two distinct protein groups.

Mycobacterium smegmatis Lhr-core crystal structure bound to ssDNA (PDB: 5V9X) in (A) and cartoon representation of its extended C-terminus determined by CryoEM (PDB: 7LHL) in (B). Domain colours in (A) show RecA1 (green), RecA2 (blue), WH (yellow) and signature domain (salmon) taken and adapted from¹³⁴. Colours in (B) depict each distinct winged helix motive (WH2 – cyan, WH3 – yellow, WH4 – orange, WH5 – blue) and an antiparallel β sheet (green) forming a characteristic 'C' shape. Taken and adapted from¹²⁵.

Lhr's extended CTD has been directly compared to AlkZ, the founding member of the HTH_42 superfamily proteins, and a DNA glycosylase enzyme responsible for the repair of interstrand crosslinks caused by azinomycin B^{56,110,125}.

1.7.4 Biochemical characterisation

Lhr-core has been biochemically interrogated in multiple bacteria and archaeal species. It is exclusively a 3' to 5' ATP-dependent ssDNA helicase able to unwind a variety of nucleic acid substrates.

Lhr from *Sulfolobus solfataricus* (protein named Hel112) was initially characterised showing unwinding activity on forked and 3'-partial duplex (PD) DNA substrates but no activity on duplex or bubbled DNA¹³⁵. *Sulfolobus acidocaldarius* Lhr1 was unable to unwind fully base paired duplex DNA and unwound a HJ substrate non-canonically resulting in single stranded DNA products¹³². Suzuki *et al.* suggest a role for *S. acidocaldarius* aLhr1 in double HJ migration to increase the frequency of crossover recombination. *Saca*Lhr1 displayed binding to variety of DNA substrates including a three-way DNA junction and duplex DNA although, duplex DNA unwinding was again not observable¹³¹. A DNA substrate preference is also seen in *Pseudomonas putida* where Lhr-core was able to unwind DNA:DNA and RNA:DNA 3'-PD when DNA was the loading/tracking strand¹³⁶.

Lhr-core from *Mycobacterium smegmatis*, *Escherichia coli* and *Thermococcus barophilus* showed a higher level of unwinding activity on RNA:DNA 3'-PD substrates with DNA as the loading/tracking strand as compared to an equivalent DNA:DNA substrate^{120,121,125}. *M. smegmatis* Lhr-core was also able to unwind a forked DNA substrate¹²⁰. Increased unwinding may be due to the relative stabilities of RNA:DNA duplexes allowing unwinding to occur more readily. Analysis on a variety of substrates across multiple Lhr types would allow pinpointing of Lhr-cores preferred nucleic acid combination.

1.7.5 Bioinformatic analysis

Lhr distribution among bacteria and archaea was extensively studied building upon the relative abundance of bacterial Lhr as noted during its initial discovery^{119,121}. This study identified multiple subgroups of Lhr proteins as shown in Figure 1.14.

Of the organisms included in the study, bacterial Lhr appeared the most divergent when comparing Lhr-core domains, attributed to an increased rate of evolution. Lhr-like proteins were also identified which may have been acquired by archaea from bacteria through horizontal gene transfer¹²¹.

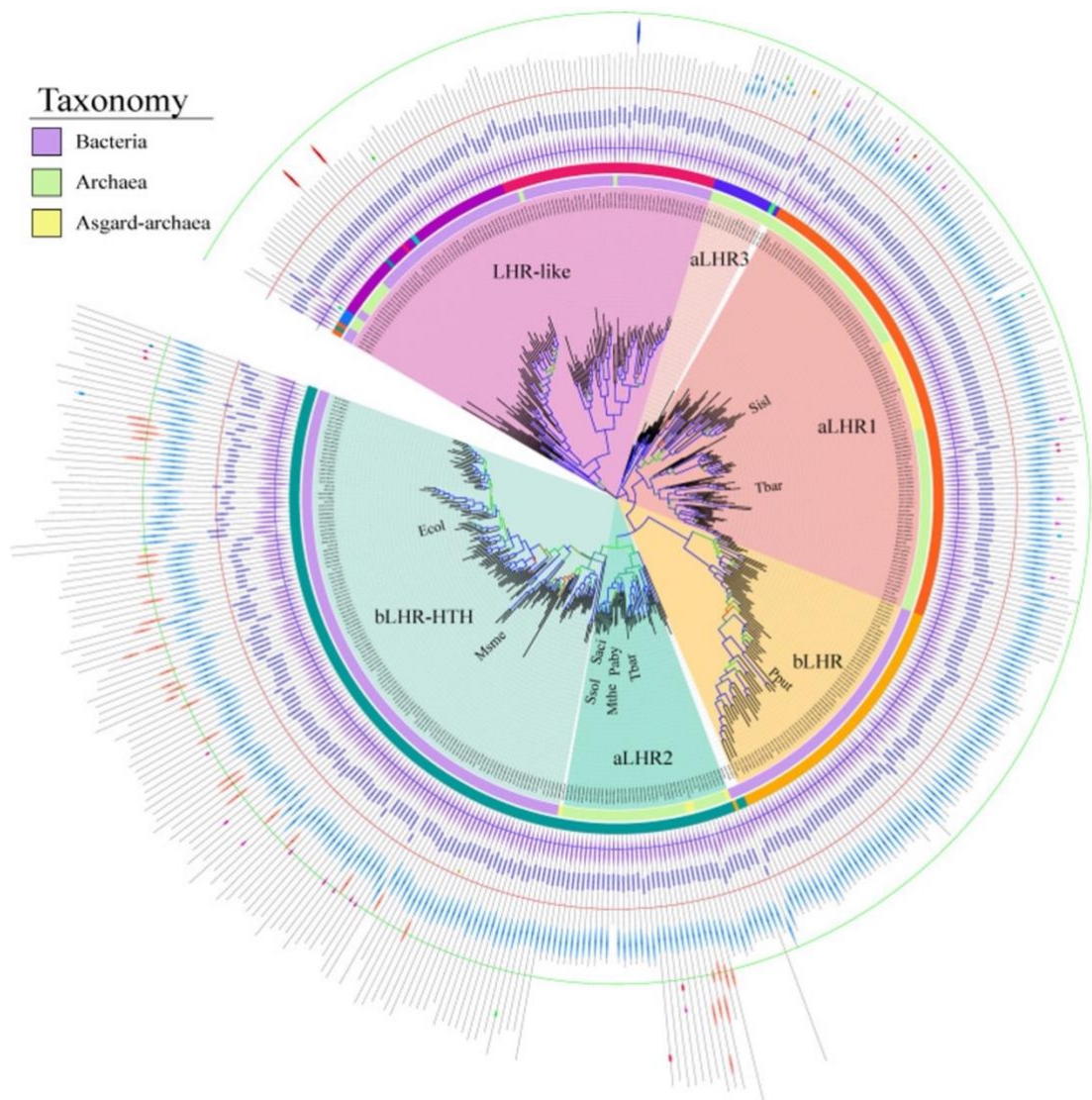


Figure 1.16 Lhr distribution in archaea and bacteria.

Extensive study and identification of multiple Lhr and Lhr-like family proteins distributed in archaea (**yellow** and **green** rings) and bacteria (**purple** ring). Reference sequences are annotated with organism abbreviation (*Msme*, *Ecol*, *Pput*, *Ssol*, *Saci*, *Mthe*, *Paby*, *Tbar* and *Sisl*). Full detailed explanation can be obtained from figure source¹²¹.

Lhr (in the form of aLhr1 and aLhr2) is highly conserved in archaea, with only 4 out of 219 studied genomes showing gene loss of both forms. Limited study was conducted in Asgard archaea due to the limited sequence data readily available¹²¹. Asgard

archaea represent the closest evolutionary link to eukaryotes so further study of Lhr in these organism may allow identification of eukaryotic homologues.

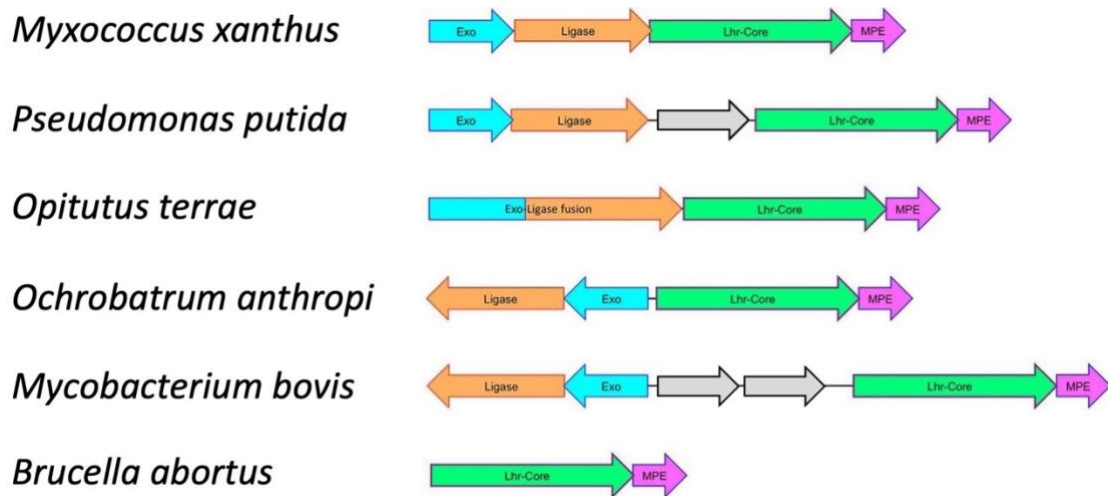


Figure 1.17 Lhr-core is located downstream of a metallophosphoesterase (MPE) in multiple bacteria.

Genetic clustering of Lhr-core (**green**) and other DNA repair enzymes (MPE – **magenta**, ATP dependent DNA ligase – **orange**, exonuclease – **cyan**). Taken and adapted from¹³⁶.

Lhr's genomic context was also investigated. No conservation of gene clusters was identified when comparing *aLhr1* and *aLhr2* contexts however, a metallophosphatase (MPP) superfamily gene was found in variable locations adjacent to *aLhr1/2* genes in a large proportion of archaeal genomes¹²¹. This data is in strong agreement with that identified for bacterial *Pseudomonas putida* Lhr-core, which is conserved in close proximity to a metallophosphoesterase¹³⁶. Further conserved genes were identified within a similar context of *aLhr2* in genomes from the Sulfolobales, Desulfurococcales and Thermococcaceae. This gene was thought to be involved in cell wall biogenesis.

1.8 Project aims

Identification and characterisation of *Trypanosoma brucei* Rad51 paralogues as a model for human homologues using structural modelling and biochemical protein analysis.

Phylogenetic analysis of Rad51 paralogue proteins across bacteria, archaea and eukaryotes to determine highly conserved regions and display ancestral relationships which may give clues to conserved protein function(s).

Outcome: Phylogenetic analysis displayed *T. brucei*'s limited suitability as model, biochemical analysis was abandoned. Structural modelling highlighted the extensive variability in Rad51 paralogue protein accessory domains. Further phylogenetic analysis highlighted the potential emergence of Rad51/RecA protein 'Rad51 C' and 'XRCC3' subfamilies.

Investigate Lhr family protein abundance in both bacteria and archaea and characterise 'Lhr-core' substrate preference and mechanistic action.

Outcome: Lhr is highly abundant in both 'Lhr-core' and 'extended' forms. *MthLhr* shows preference to forked DNA substrates. Lhr-mediated fork branchpoint distortion allows protein loading. Subsequent translocation through the parental duplex results in strand separation.

Identify *Escherichia coli* Lhr DNA repair phenotypes using genetic techniques with investigation into repair pathway protein partners.

Characterisation of *Escherichia coli* Lhr's extended C-terminus using biochemical techniques and identify catalytic residues.

Outcome: *EcoLhr* has a role in relieving replicative stress and in oxidative damage repair, often in conjunction with RadA (Sms). Lhr-CTD displayed d-uracil DNA glycosylase activity mediated through a catalytic aspartic acid residue, with strong preference to forked DNA substrates.

Chapter 2 Materials and Methods

2.1 Antibiotics

Ampicillin and kanamycin powder was dissolved into sterile distilled water, passed through a 0.2 μm filter, decanted into 1 ml aliquots and stored at -20°C .

Chloramphenicol powder was dissolved in absolute ethanol (Fisher) and stored at -20°C as a 10 ml stock solution.

Table 2.1 Antibiotics used for cloning.

Antibiotic	Supplier	Stock concentration	Working concentration
Ampicillin	Invitrogen	100 mg/ml	50 $\mu\text{g}/\text{ml}$
Chloramphenicol	Fisher Scientific	35 mg/ml	35 $\mu\text{g}/\text{ml}$
Kanamycin	Alfar Aesar	50 mg/ml	40 $\mu\text{g}/\text{ml}$

2.2 Bacterial strains and cell lines

Table 2.2 List of cell strains used within this study.

Strain glycerol stocks containing 40% glycerol v/v were stored at -80°C and streaked out before use.

<i>E. coli</i> strain	Supplier	Genotype	Use
DH5α	New England BioLabs	F ⁻ Φ80 <i>lacZ</i> ΔM15 Δ(<i>lacZYA-argF</i>)U169 <i>recA1 endA1 hsdR17</i> (<i>r_K⁻</i> , <i>m_K⁺</i>) <i>phoA supE44 thi-1 gyrA96 relA1 λ⁻</i>	Molecular cloning
BL21 AI	ThermoFisher	F ⁻ <i>ompT hsdS_B (r_B⁻ m_B⁻) gal dcm araB::T7RNAP-tetA</i>	Protein overexpression
BL21 Codon plus (C+)	Agilent technologies	F ⁻ <i>ompT hsdS(r_B⁻ m_B⁻) dcm⁺ Tet^r gal endA Hte [argU ileY leuW Cam^r]</i>	Protein overexpression
Rosetta 2 (DE3)	Sigma-Aldrich	F ⁻ <i>ompT hsdSB(r_B⁻ m_B⁻) gal dcm (DE3) pRARE2 (CamR)</i>	Protein overexpression
MG1655	ATCC	F ⁻ <i>lambda- ilvG- rfb-50 rph-1</i>	Genetic analysis
<i>dnaE486 ΔrecQ</i> (OH1000) ¹³⁷	Dr Takashi Hishida (Osaka University, Japan)	<i>thr-1, araC14, leuB6(Am), Δ(gpt-proA)62, lacY1, tsx-33, qsr⁻0, glnV44(AS), galK2(Oc), LAM-, Rac-0, hisG4(Oc), rfbC1, mgl-51, rpoS396(Am), rpsL31(strR), kdgK51, xylA5, mtl-1, argE3(Oc),</i>	Genetic analysis

		<i>thi-1, rec⁺, ruv⁺, gyr⁺, dna,</i> <i>dnaE486, ΔrecQ::Cam^r</i>	
--	--	---	--

Table 2.3 Cell strains produced by me during this work for genetic analysis.

<i>E. coli</i> strain	Genotype	Method created
RB001a	Δ/hr Kan ^R	P1 transduction from Keio collection 'CGSC# - 9400' into MG1655.
RB002a	Δ/hr	RB001a pCP20 treated to remove Kan ^R FRT scar.
RB003a	$\Delta/hr \Delta/radA$ Kan ^R	P1 transduction from Keio collection 'CGSC# - 11107' into RB002a.
RB004a	$\Delta/radA$ Kan ^R	P1 transduction from Keio collection 'CGSC# - 11107' into MG1655.
RB005a	$\Delta/radA$ Kan ^R	P1 transduction from Keio collection 'CGSC# - 11107' into BL21 AI.
RB006a	$\Delta/hr \Delta/radA$	RB003a pCP20 treated to remove Kan ^R FRT scar.
RB007a	$\Delta/radA$	RB004a pCP20 treated to remove Kan ^R FRT scar.

2.3 Plasmids and DNA substrates

Table 2.4 A list of plasmid vectors used within this study for cloning and overexpressions.

Plasmid	Size	Description	Resistance
pET22b	5493	ColE1 replicon, C-terminal His ₆ -tag, MCS ¹ , T7 promoter/terminator, lac operator.	Ampicillin
pT7-7	2473	ColE1 replicon, MCS, T7 promoter.	Ampicillin
pET14b	4671	ColE1 replicon, N-terminal thrombin cleavable His ₆ -tag, MCS, T7 promoter/terminator.	Ampicillin
pACYC Duet	4008	P15A replicon, dual MCS, T7 promoter/terminator, lac operator.	Chloramphenicol
pNH-TrxT	7602	ColE1 replicon, N-terminal His ₆ -Trx, LIC ² site, <i>sacB</i> for negative selection on 5% sucrose. Gift from Dr. Christopher Cooper.	Kanamycin
pRARE2	4965	Contains 7 rare tRNAs controlled under native promoters.	Chloramphenicol
pCP20	≈9400	oriV _{pSC101} (TS ³ replicon), <i>S. cerevisiae</i> Flp ¹³⁸⁻¹⁴⁰ , λ repressor (TS).	Ampicillin, Chloramphenicol

¹ Multiple cloning site, contains multiple restriction sites allowing targeted insertion of foreign DNA.

² Ligation independent cloning.

³ Temperature sensitive.

Table 2.5 A list of plasmid constructs used within this study.

Plasmid	Gene	Parent Vector	Description	Resistance
pEB352	<i>lhr</i>	pET22b	<i>Methanothermobacter</i> <i>lhr</i> cloned as described ¹¹⁰ .	Ampicillin
pEB353	<i>lhr</i>	pT7-7	<i>Methanothermobacter</i> <i>lhr</i> cloned as described ¹¹⁰ .	Ampicillin
pEB695	<i>rNase T</i>	pET14b	<i>Escherichia coli rNaseT</i> cloned NdeI/BamHI with N-terminal His ₆ - thrombin site.	Ampicillin
pRJB14	<i>radA (sms)</i>	pACYC Duet	<i>Escherichia coli radA</i> cloned BamHI/HindIII with N-terminal His ₆ - tag.	Chloramphenicol
pRJB15	<i>lhr</i>	pT7-7	<i>Escherichia coli lhr</i> cloned NdeI/HindIII untagged.	Ampicillin
pRJB23	<i>lhr</i> CTD	pNH-TrxT	<i>Escherichia coli lhr</i> nucleotides 876-1538 with His ₆ -Trx tag cloned via LIC by	Kanamycin

			Nadia Ahmed in Dr Christopher Cooper's lab (University of Huddersfield).	
pRJB28	<i>lhr</i>	pT7-7	pRJB15 with His ₆ -tag inserted between codons 1 and 2, created by Gibson assembly.	Ampicillin
pRJB29	<i>lhr</i> CTD	pNH-TrxT	pRJB23 with D1536A mutation created by Gibson assembly.	Kanamycin
pRJB32	<i>lhr</i>	pT7-7	pRJB28 with D1536A mutation created by Gibson assembly.	Ampicillin

Oligonucleotides listed in Table 2.6 through to Table 2.9 were ordered and synthesised by Merck (Sigma-Aldrich) unless stated otherwise.

Table 2.6 List of primers used for cloning.

Underlined characters indicate respective restriction endonuclease consensus sequence, characters in bold correspond to respective gene sequence.

Gene target	Primer use	Oligonucleotide sequence 5' to 3'
Mth_1802 (<i>lhr</i>)	PCR amplification from <i>Mth</i> genomic DNA for cloning into pET22b (pEB352) and pT7-7 (pEB353) using NdeI site.	CATGCATATGATAAAGAAACAGGAGAGG
	PCR amplification from <i>Mth</i> genomic DNA for cloning into pET22b (pEB352) and pT7-7 (pEB353) using EcoRI site.	CATGGGATCCCTACCTTTTTATTCATC
Eco_b1653 (<i>lhr</i>)	PCR amplification from <i>Eco</i> genomic DNA for cloning into pT7-7 (pRJB15) using NdeI site.	GCGCATATGGCAGATAATCCAGACCCCTC

	PCR amplification from <i>Eco</i> genomic DNA for cloning into pT7-7 (pRJB15) using HindIII site.	ATCG <u>AAGCTT</u> CTATCCCAATCCAGCCCTTG
Eco_b4389 [<i>rada (sms)</i>]	PCR amplification from <i>Eco</i> genomic DNA for cloning into pACYC Duet (pRJB14) using BamHI site.	CATGGGATCCGCAAAGCTCCAAAACGCG
	PCR amplification from <i>Eco</i> genomic DNA for cloning into pACYC Duet (pRJB14) using HindIII site.	CATGAAGCTTTTATAAGTCGTCGAACACGC
Eco_b1652 (<i>rNaseT</i>)	PCR amplification from <i>Eco</i> genomic DNA for cloning into pET14b (pEB695) using NdeI site.	GCGCATATGTCGGATAACGCTCAACTTACC
	PCR amplification from <i>Eco</i> genomic DNA for cloning into pET14b (pEB695) using BamHI site.	GCGGGATCCTTACACCTCTTCGGCGGCAG

Table 2.7 List of primers used for site directed mutagenesis (SDM).

Gene target	Primer use	Oligonucleotide sequence 5' to 3'
C-terminal <i>lhr</i> in pNH-TrxT (pRJB23). Contains nucleotides 876-1538.	Gibson assembly primer pair for D1536A mutation 'GAT' to 'GCG'.	AGGGCTGGCGTGGGGATAGCAGTAAAGGTGGATACGGATCC
		GTAGGACAGGTGCCGGCAGCGCTCTGGG
	Gibson assembly primer pair for D1536A mutation 'GAT' to 'GCG'.	GCTGCCGGCACCTGTCCTACGAGTTGCATG
		GCTATCCCCACGCCAGCCCTTGTGGCGAACTTG
Full length <i>lhr</i> in pT7-7 (pRJB28).	Gibson assembly primer pair for D1536A mutation 'GAT' to 'GCG'.	AGGGCTGGCGTGGGGATAGAGCTTATCGATGATAAGCTG
		CGGTTAGAGGTTGCGCCGGAGGTCGACT
		TCCGGCGCAACCTCTAACCGTGTACAAAGTAGC

	Gibson assembly primer pair for D1536A mutation 'GAT' to 'GCG'.	TCTATCCCCACGCCAGCCCTTGTGGCGAACTTG
Full length <i>lhr</i> in pT7-7 (pRJB15).	Gibson assembly primer pair for insertion of His ₆ -tag between 1 st and 2 nd codons	GATTATCTGCGTGGTGATGATGGTGATGCATATGTATATCTCCTTCTAAAG
		TTAAACAAAATTATTTCTAGAGGGAAACCG
	Gibson assembly primer pair for insertion of His ₆ -tag between 1 st and 2 nd codons	CGTTCACTCCCGCCGAAGCGGATCAGGC
		TATACATATGCATCACCATCATCACCACGCAGATAATCCAGACCCTTCATC
		CGCTTCGGCGGGAGTGAACGATTCTGCAGC

Table 2.8 List of primers used for strain verification.

Gene target	Primer use	Oligonucleotide sequence 5' to 3'
<i>Ihr</i> ⁴	Binds start codon of gene.	GCGCATATGGCAGATAATCCAGACCCTTC
	Binds to terminus of gene.	GCCGCTCGAGCTATCCCCAATCCAGCCCTTG
<i>radA (sms)</i>	Binds 100 base pairs upstream of gene.	TGACCTGATGGGGTATTCTGC
	Binds 100 base pairs downstream of gene.	ATACCGCTGGCATCAGCTACCTGC

⁴ Primers used here also used for cloning so contain NdeI and XhoI sites respectively.

Table 2.9 List of oligonucleotides used for *in vitro* analysis.

Strand labelling is indicated with ●- for ³²P radioactive labelled substrates and ●- for Cy5, ■- for ATTO 674N, ■- for ATTO 532 fluorescently labelled substrates. ATTO 674N and ATTO 532 oligos were synthesised by IDT. **Bold underlined** shows location of ‘damaged’ DNA base. Damaged bases located within the duplex region of substrates base pair as if ‘in a cell’ before repair to mimic any changes to DNA topology (underlined).

Substrate name	Oligonucleotide name	Oligonucleotide sequence 5' to 3'
Linear duplex	RGL16	●ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	ELB37	ACGCTGCCGAATTCTGGCTTGCTAGGACATGCTGTCTAGAGACTATCGAT
3'-tailed duplex	RGL16	●ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	PM2	GGACATGCTGTCTAGAGACTATCGAT
5'-tailed duplex	RGL16 5' ³² P	●ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	ELB38	ACGCTGCCGAATTCTGGCTTGCTAGG
Gapped 70mer duplex	ELB41	GCAGGATCCGATCCGTAAGGAGCTCTCGAAGGCCATCGTCGCGAACG ATCCTGCCTAGGGAGCTCC

	ELB42 5' ³² P	●GGAGCTCCCTAGGCAGGATCG
	ELB43 5' ³² P	●CGAAGAGCTCCAGTTACGGATACGGATCCTGC
J12 – Mobile Holliday junction 'HJ1' ¹¹⁰	RGL16	ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	RGL13 5' ³² P	●GACGCTGCCGAATTCTGGCTTGCTAGGACATCTTTGCCACGTTGACCC
	RGL14	TGGGTCAACGTGGGCAAAGATGTCCTAGCAATGTAATCGTCTATGACGTT
	RGL15	CAACGTCATAGACGATTACATTGCTAGGACATGCTGTCTAGAGACTATCGA
J6 – Static Holliday junction 'HJ2' ¹¹⁰	RGL16	ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	ELB21 5' ³² P	●GACGCTGCCGAATTCTGGCTTGCTAGGACATTCTTTGCCACGTTGACCC
	ELB23	GGGTCAACGTGGGCAAAGAATGTCCTACGTCCGATACGGATAATCGCCAT
	ELB22	ATGGCGATTATCCGTATCGGACGTAGGACATGCTGTCTAGAGACTATCGA
Fork 4a – Breathable 'Fork1' ¹¹⁰	RGL16	ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	ELB37	ACGCTGCCGAATTCTGGCTTGCTAGGACATGCTGTCTAGAGACTATCGAT
	RGL13 5' ³² P	●GACGCTGCCGAATTCTGGCTTGCTAGGACATCTTTGCCACGTTGACCC

	PM3	TGGGTCAACGTGGGCAAAGATGTCC
Fork 4b – Static 'Fork2' ¹¹⁰	RGL16	ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	ELB37	ACGCTGCCGAATTCTGGCTTGCTAGGACATGCTGTCTAGAGACTATCGAT
	RGL13 5' ³² P	●GACGCTGCCGAATTCTGGCTTGCTATGTA ACTCTTTGCCACGTTGACCC
	ELB30	GGGTCAACGTGGGCAAAGAGTTACA
Fork2a – Flayed duplex 'FD' ¹¹⁰	RGL16	ATCGATAGTCTCTAGACAGCATGTCCTAGCAAGCCAGAATTCGGCAGCGT
	RGL13 5' ' ³² P	●GACGCTGCCGAATTCTGGCTTGCTATGTA ACTCTTTGCCACGTTGACCC
Fork2a – Flayed duplex RNA	RGL16 RNA	AUCGAUAGUCUCUAGACAGCAUGUCCUAGCAAGCCAGAAUUCGGCAGCGU
	RGL13 5' ' ³² P	●GACGCTGCCGAATTCTGGCTTGCTATGTA ACTCTTTGCCACGTTGACCC
Fork2b – Flayed duplex	MW12 5' Cy5	●TCGGATCCTCTAGACAGCTCCATGATCACTGGCACTGGTAGAATTCGGC
	MW14	CAACGTCATAGACGATTACATTGCTACATGGAGCTGTCTAGAGGATCCGA
Fork 2 FRET– Fully base paired fork	RGL16-Biotin	Biotin- ATCGATAGTCTCTAGACAGTATGTCCTAGCAAGCCAGAATTCGGCAGCGT

	PM2-ATTO 647N	GGACA T ACTGTCTAGAGACTATCGAT
	RGL13	GACGCTGCCGAATTCTGGCTTGCTATGTAAATCTTTGCCACGTTGACCC
	ELB30-ATTO 532	GGGTCAACGTGGGCAAAGAT T TACA
MW12 37	MW12 37 5' Cy5	●GTCGGATCCTCTAGACAGGCTCCATGCGTAGTACTCG
MW12 d-U	MW12 37 d-Uracil 5' Cy5	●GTCGGATCCTCTAGACA <u>U</u> GCTCCATGCGTAGTACTCG
MW12 oxo-d-G	MW12 37 8-oxo-d-Guanine 5' Cy5	●GTCGGATCCTCTAGACA <u>G</u> GCTCCATGCGTAGTACTCG
MW12 37 duplex	MW12 37 5' Cy5	●GTCGGATCCTCTAGACAGGCTCCATGCGTAGTACTCG
	MW14 37C Duplex	CGAGTACTACGCATGGAGCCTGTCTAGAGGATCCGAC
MW12 d-U duplex	MW12 37 d-Uracil 5' Cy5	●GTCGGATCCTCTAGACA <u>U</u> GCTCCATGCGTAGTACTCG
	MW14 37G Duplex	CGAGTACTACGCATGGAGC <u>G</u> TGTCTAGAGGATCCGAC
MW12 oxo-d-G duplex	MW12 37 8-oxo-d-Guanine 5' Cy5	●GTCGGATCCTCTAGACA <u>G</u> GCTCCATGCGTAGTACTCG

	MW14 37C Duplex	CGAGTACTACGCATGGAGC <u>C</u> TGTCTAGAGGATCCGAC
MW12 37 fork	MW12 37 5' Cy5	●GTCGGATCCTCTAGACAGGCTCCATGCGTAGTACTCG
	MW14 37C Fork	ATTACATTGCTACATGGAGCCTGTCTAGAGGATCCGAC
MW12 d-U fork	MW12 37 d-Uracil 5' Cy5	●GTCGGATCCTCTAGACA <u>U</u> GCTCCATGCGTAGTACTCG
	MW14 37G Fork	ATTACATTGCTACATGGAGC <u>G</u> TGTCTAGAGGATCCGAC
MW12 oxo-d-G fork	MW12 37 8-oxo-d-Guanine 5' Cy5	●GTCGGATCCTCTAGACA <u>G</u> GCTCCATGCGTAGTACTCG
	MW14 37C Fork	ATTACATTGCTACATGGAGC <u>C</u> TGTCTAGAGGATCCGAC

2.4 Solution composition

2.4.1 Media

Table 2.10 A list of media broths used within this study for *E. coli* cell culture cloning, protein overexpression and genetic analysis.

Media name	Composition
LB Broth	10 g/L Tryptone (BD) 10 g/L NaCl 5 g/L Yeast Extract (BD) adjusted to pH 7 using NaOH
LB Broth agar plates	10 g/L Tryptone (BD) 10 g/L NaCl 5 g/L Yeast Extract (BD) 1.5% w/v Agar powder (VWR) adjusted to pH 7 using NaOH
P1 agar plates	10 g/L Tryptone (BD) 10 g/L NaCl 5 g/L Yeast Extract (BD) 1.5% w/v Agar powder (VWR) 0.13% w/v D-Glucose (Fisher) 5 mM CaCl ₂ adjusted to pH 7 using NaOH
MC Buffer	100 mM MgSO ₄ 5 mM CaCl ₂

M9 Minimal media	1x M9 minimal Salts 200 μ M Adenine HCl 2 mM CaCl ₂ 2 mM ZnSO ₄ 10 μ g/ml Thiamine HCl 20 mM Glucose
------------------	---

2.4.2 Electrophoresis running buffers and stains

Table 2.11 A list of running buffers used for agarose gel electrophoresis and sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS PAGE) analysis for separation of DNA and protein respectively.

Sample loading buffers and SDS PAGE stains also listed, used for separated protein visualisation.

Stock solution	Composition
10x Tris-borate-EDTA (TBE) running buffer	1 M TRIS (Sigma-Aldrich) 1 M Boric acid (Fisher) 20 mM EDTA (Fisher)
10x DNA loading buffer	5% SDS (Sigma-Aldrich) 10 mM EDTA 40% Glycerol (Sigma-Aldrich) 4.64 mM Xylene glycol 3.73 mM Bromophenol blue
10x SDS PAGE running buffer	250 mM TRIS 1.92 M Glycine (Sigma-Aldrich) 1% (v/v) SDS
5x SDS PAGE loading buffer	200 mM Tris-HCl pH6.8 8% SDS 0.4% Bromophenol blue 40% Glycerol
Coomassie Brilliant blue stain	40% Methanol 10% Glacial acetic acid

	0.05% Coomassie Brilliant Blue R-250
Destain	20% Ethanol 10% Acetic Acid
Formamide Stop	79% Formamide 20% Glycerol 20 mM EDTA

2.4.3 Solutions used for substrate preparation and *in vitro* protein analysis

Table 2.12 Composition of buffers used in nucleic acid substrate preparation and in nucleic acid binding, unwinding and glycosylase assays.

Buffer name	Composition
10x SSC buffer	1.5 mM NaCl 150 mM Na Citrate pH 7
10x Annealing Buffer	100 mM Tris-HCl pH 7.5 500 mM NaCl 10 mM EDTA
5x OG dye	80% Glycerol A pinch of Orange G dye (use a spatula)
Fork elution buffer	20 mM Tris-HCl pH8 50 mM NaCl
Helicase buffer (HB)-DTT (HB buffer without DTT)	20 mM Tris-HCl pH7.5 10% glycerol 100 µg/ml BSA (Sigma-Aldrich)
Stop solution	2 mg/ml proteinase K (Invitrogen) 2.5% SDS 200 mM EDTA

2.4.4 Buffers used during protein purification

Table 2.13 Composition of buffers used during protein purification.

Buffer name	Composition
Ni-NTA buffer A	20 mM Tris-HCl pH8 20 mM Imidazole (Sigma-Aldrich) 500 mM NaCl 10% Glycerol
Ni-NTA buffer B	20 mM Tris-HCl pH8 500 mM Imidazole 500 mM NaCl 10% Glycerol
Low salt buffer A	20 mM Tris-HCl pH8 150 mM NaCl 10% Glycerol
High salt buffer B	20 mM Tris-HCl pH8 1.5 M NaCl 10% Glycerol
Hydrophobic buffer A	20 mM HEPES pH8 (Sigma-Aldrich) 1.5 M Ammonium sulphate (Sigma-Aldrich) 10% Glycerol
No Salt buffer B	20 mM HEPES pH8 10% Glycerol

Ni-NTA buffer A His-Lhr	20 mM HEPES pH8 20 mM Imidazole 1.5 M NaCl 10% Glycerol
Ni-NTA buffer B His-Lhr	20 mM HEPES pH8 500 mM Imidazole 1.5 M NaCl 10% Glycerol
Low salt buffer 2A	20 mM HEPES pH8 150 mM NaCl 10% Glycerol
High salt buffer 2B	20 mM HEPES pH8 1.5 M NaCl 10% Glycerol

2.5 General Microbiology

2.5.1 Protocol for making competent cells

Escherichia coli cell strains listed in Table 2.2 and Table 2.3 were used to inoculate baffled conical flasks containing LB broth in a 1:100 dilution. Cells were grown in a Fisher Scientific shaking water at 37°C until they reached an optical density (OD)₆₀₀ of 0.5, monitored by spectrophotometry. Cells were then harvested by centrifugation (5 minutes, 1800 RCF using an Eppendorf Centrifuge 5430R) and resuspended into 1/8th volume of ice cold 0.1 M CaCl₂. Cells were incubated on ice for an hour before being pelleted again using a pre-chilled rotor. Cells were resuspended in the same volume of fresh 0.1 M CaCl₂ as before. Ice cold glycerol was added to 30% before aliquoting and flash freezing for storage at -80°C.

2.5.2 Protocol for transforming competent cells

Competent cells were thawed on ice before use. 100 µl of cells were added to 5 ng of plasmid, mixed by gentle pipetting. Cell/DNA mix were incubated on ice for 20 minutes before being heat shocked for 2 minutes at 42°C and incubated on ice for a further 2 minutes. LB broth was added and cell/DNA mix was incubated at 37°C for 1 hour in a shaking water bath. Cells were pelleted via centrifugation (1 minute at maximum speed), resuspended in ½ volume LB broth and 200 µl was plated onto appropriate LB agar plates using the spread plate technique¹⁴¹. Plates were transferred to a 37°C incubator and grown overnight.

2.5.3 Polymerase chain reaction (PCR)

50 µl PCR reaction mixtures were assembled containing 50 ng of template DNA, 0.5 µM upstream primer, 0.5 µM downstream primer, 1 mM dNTP mix (reaction mixture contains 200 µM of each dNTP), 1x ThermoPol buffer, 2 mM MgSO₄ and 1 µl of Vent polymerase (NEB).

For SDM, Q5 hot start polymerase was used instead of Vent. SDM reactions were run in 1x Q5 reaction buffer and MgSO₄ was omitted.

For colony PCR, DNA template was prepared by resuspending a single colony in 20 µl of sterile distilled water and boiled at 95°C for 5 minutes. 2 µl was then added to the 'Vent PCR reaction mix' as stated above.

The following PCR programmes were used for DNA amplification using a Veriti Thermal Cycler (Applied Biosystems).

Table 2.14 PCR programme for standard Vent PCR.

Cycle step	Temperature	Time	Number of cycles
Initial denaturation	95°C	5 minutes	1x
Denaturation	95°C	30 seconds	35x
Annealing	65°C	30 seconds	
Extension	72°C	1 minute per kb	
Final extension	72°C	5 minutes	1x

Table 2.15 PCR programme for standard Q5 hot start PCR.

Cycle step	Temperature	Time	Number of cycles
Initial denaturation	98°C	30 seconds	1x
Denaturation	98°C	10 seconds	35x
Annealing	65°C	30 seconds	
Extension	72°C	30 seconds per kb	
Final extension	72°C	2 minutes	1x

Successful PCR amplification was confirmed via analysis by agarose gel electrophoresis as described in section 2.6.1.

2.5.4 Site directed mutagenesis

Interesting conserved residues located in Lhr's 'glycosylase pocket' were targeted for mutagenesis. Primers listed in Table 2.7 were designed using NEBaseChanger. Mutagenesis was performed first in a plasmid containing only C-terminal *E. coli* Lhr (pRJB23), interesting mutants were then translated to the full length protein for further analysis (into plasmid pRJB28). Amino acid substitutions were achieved using a make shift 'Q5 site directed mutagenesis kit'. In summary pRJB23 containing WT C-Lhr was subject to PCR with Q5 hot start polymerase. A SDM mix was made using 1 µl PCR product, 1 µl of T4 DNA Ligase, 1 µl of DpnI, 1 µl of T4 PNK, 1 µl of T4 DNA Ligase buffer and 4 µl of SDW. Enzymes quoted were sourced from NEB. The SDM mix was incubated at room temperature for 1 hour. The total resulting reaction mixture was transformed into *E. coli* DH5alpha cells using the transformation protocol previously

described in section 2.5.2. Colonies were picked and plasmids were purified using the 'Wizard® Plus SV Miniprep DNA Purification System'¹⁴² (Promega) before being sent for Sanger sequencing by GeneWiz.

Certain mutations proved difficult using the above method so where stated, a Gibson assembly method was used.

PCR primers were generated using NEBuilder to give two PCR fragments containing homologous overlap and the desired mutation(s). PCR samples were run on a 0.5% w/v agarose gel, bands of the correct size were excised and extracted using the 'agarose gel extraction protocol' as described in section 2.6.2. For assembly, the NEBuilder HiFi Assembly Master Mix (NEB) was used with a few slight alterations from the standard protocol¹⁴³. 0.2 pmol of each fragment was added to the mix which was incubated at 50°C for 1 hour. The total resulting mix was transformed into *E. coli* DH5alpha cells using the transformation protocol previously described in section 2.5.2. Colonies were picked and plasmids were purified using the 'Wizard® Plus SV Miniprep DNA Purification System'¹⁴² (Promega) before being sent for Sanger sequencing by GeneWiz.

2.5.5 Production of knockout cell lines for genetic analysis

Knockout strains shown in Table 2.3 were produced by P1 transduction of an FRT (FLP recognition target) flanked Kan^R marker from respective 'Keio collection' knockout strain into MG1655. P1 transduction and removal of Kan^R-FRT insert was performed as follows:

Preparation of Phage P1 lysate from Keio collection knockout cell strain

0.5 ml of Keio collection strain overnight culture was used to inoculate 8 ml of LB Broth containing 6 mM CaCl₂. Cells were grown at 37°C to OD₆₀₀ 0.8-1.0 in a shaking water bath. Upon reaching OD, 0.1 ml of cells were added to four small overlay tubes containing 3 ml of 0.4% w/v LB broth agar and held at 42°C to prevent premature setting. P1 phage stock was diluted 10-100-fold in MC buffer to produce ≈10⁸ pfu/ml (plaque-forming units). 0.05 ml, 0.1 ml and 0.2 ml of diluted phage was added to cell/agar containing overlay tubes as a titre and gently vortexed to mix. The remaining tube was left without phage as a control. The contents of each overlay tube was poured onto P1 agar plates and left to set. Lid condensation was wiped and plates were grown overnight, lid up, at 37°C for no longer than 18 hours. Two successful phage-lysed plates were scraped into the same falcon tube using a glass rod. 1 ml of MC buffer and 0.5 ml of chloroform was added and mixed vigorously by vortex before centrifugation at 5752 RCF for 20 minutes at 4°C. The supernatant was collected and 0.5 ml of chloroform was added, mixed by inversion and stored at 4°C as new P1 phage stock.

Transduction of antibiotic resistance markers into desired cell strain

0.5 ml of MG1655 overnight culture was added to 8 ml of LB broth and grown to OD₆₀₀ 0.8 in a shaking water bath. Upon reaching OD, cells were pelleted, resuspended into 1 ml MC buffer and left at room temperature for 10 minutes. 0.2 ml of cells were added into 3 overlay tubes containing 0 ml, 0.05 ml and 0.2 ml of P1 lysate produced previously, and incubated for 30 minutes at 37°C. 2.5 ml of 0.6% agar was added to cell/P1 lysate containing overlay tubes, mixed gently and poured onto LB broth agar plates containing 30 µg/ml kanamycin and left to set. Plates were grown for 1-2 days,

lid-up, at 37°C to allow colonies to develop. Colonies were then picked and purified by streaking onto LB broth agar plates containing no antibiotic. This was repeated 3 times before plating again onto agar containing 30 µg/ml kanamycin for confirmation of gene knockout and Kan^R-FRT insertion.

Removal of Kan^R-FRT gene inserted by P1 transduction

Successful colonies were made competent and transformed with pCP20 using methods described in sections 2.5.1 and 2.5.2, with transformed plates being grown at 30°C overnight. A colony was picked and used to inoculate 8 ml of LB broth containing no antibiotic. Culture was grown overnight at 45°C in a shaking water bath then streaked onto LB broth agar plates to produce single colonies. Plates were grown overnight at 37°C. Colonies were re-streaked a further 3 times using this method. Multiple colonies were replica-streaked to confirm loss of pCP20 plasmid and Kan^R-FRT insertion. This entailed streaking onto LB broth agar plates containing 50 µg/ml ampicillin, 30 µg/ml kanamycin and then no antibiotic. Isolates which only grew on the no antibiotic agar plates were grown overnight for glycerol stock production and streaked a further time for colony PCR diagnostic confirmation as stated in section 2.5.3.

2.6 Gel electrophoresis

2.6.1 Agarose gel electrophoresis

DNA samples were separated and visualised using 0.5% agarose gels stained with 0.2 µg/ml ethidium bromide (Sigma-Aldrich) in 1x TBE buffer. 10x DNA loading buffer was added to each sample. A 1 kb ladder (NEB) was run alongside for band sizing. Sufficient migration was achieved after 60 minutes at 120 V using a PowerPac Basic power supply (BioRad). Gels were imaged using a U:Genius³ Bio-imaging system (Syngene) with UV exposure.

2.6.2 Agarose gel extraction protocol

DNA products were excised from agarose gels in the dark room using an LED Blue/White Light Transilluminator (Thermo Scientific). Excised bands were subject to the Wizard[®] SV Gel and PCR Clean-Up System using the protocol as stated¹⁴² (Promega). DNA was eluted into sterile distilled water.

2.6.3 SDS PAGE analysis

Protein containing samples were added to 4x SDS loading buffer and 30 mM dithiothreitol (DTT) before being boiled at 95°C for 8 minutes using a Dri-block DB-2D (Techne) heat block. Samples were run on either a standard 8 or 10% SDS gel at 120 V for 1 hour to allow full migration. To visualise proteins present gels were stained using 'Coomassie Brilliant blue stain' and de-stained using 'Destain buffer' as described in Table 2.11.

2.7 Phylogenetic analysis

Phylogenetic trees were generated using the “One Click” mode on ‘phylogeny.fr’. This analysis, needing only FASTA sequences to begin, chains together multiple programs for simple and effective phylogenetic tree generation¹⁴⁴. Processes are as follows.

2.7.1 MUSCLE multiple sequence alignment

Multiple Sequence Comparison by Log-Expectation (MUSCLE) allows fast and effective alignment of multiple protein sequences. Exact parameters can be found here¹⁴⁵ and additional comparative testing here¹⁴⁶.

2.7.2 Gblocks alignment check

Gblocks software highlights poorly aligned sequences or ‘blocks’ of sequence and eliminates them from the overall alignment. Although this step essentially reduces information from the analysis, it bolsters the phylogenetic signal to give a more accurate picture of the relationship between highly conserved regions. Extensive discussion and justification can be found here ^{147,148}.

2.7.3 PhyML tree generation and TreeDyn visualisation

PhyML allows fast and accurate Maximum Likelihood tree generation¹⁴⁹ which can be visualised and annotated using the tethered TreeDyn software¹⁵⁰.

2.7.4 List of proteins used for phylogenetic analysis

Tables describing the proteins used for the phylogenetic analysis of RecA/Rad51 family proteins. Protein names as presented in section 3.1 and PDB ID are shown.

Species, species code and – Phylum, as described

***Homo sapiens*, Hsa** – Chordata

Table 2.16 Proteins used from *Homo sapiens* with corresponding PDB IDs as described.

Protein name	PDB ID
Rad51_A	RAD51_HUMAN
Rad51_B	RA51B_HUMAN
Rad51_C	RA51C_HUMAN
Rad51_D	RA51D_HUMAN
XRCC2	XRCC2_HUMAN
XRCC3	XRCC3_HUMAN
DMC1	DMC1_HUMAN

***Trypanosoma brucei*, Tbr** – Euglenozoa

Table 2.17 Proteins used from *Trypanosoma brucei* with corresponding PDB IDs as described.

Protein name	PDB ID
p1	Q38E34_TRYB2
p2	Q384W8_TRYB2
p3	Q384K0_TRYB2
p4	Q580V2_TRYB2
p5	Q386Q5_TRYB2
p6	Q389E0_TRYB2

***Methanothermobacter thermautotrophicus*, Mth** – Euryarchaeota

Table 2.18 Proteins used from *Methanothermobacter thermautotrophicus* with corresponding PDB IDs as described.

Protein name	PDB ID
RadA	RADA_METTH
RadB	RADB_METTH

***Escherichia coli*, Eco** – Proteobacteria

Table 2.19 Proteins used from *Escherichia coli* with corresponding PDB IDs as described.

Protein name	PDB ID
RecA	RECA_ECOLI
RadA_Sms	RADA_ECOLI

***Caenorhabditis elegans*, Cel** – Nematoda

Table 2.20 Proteins used from *Caenorhabditis elegans* with corresponding PDB IDs as described.

Protein name	PDB ID
RFS-1	RFS1_CAEEL

***Synechococcus elongatus*, Syf** – Cyanobacteria

Table 2.21 Proteins used from *Synechococcus elongatus* with corresponding PDB IDs as described.

Protein name	PDB ID
KaiC	KAIC_SYNE7

2.8 *In vivo* experimentation

Escherichia coli cells used for genetic analysis were freshly streaked to single colonies from -80°C stored glycerol stocks before each experiment. Overnights were set up from single colonies without the plate entering the fridge to limit occurrence of suppressor mutations due to low temperature shock. Overnights, outgrowth and plating were performed using LB Broth media (see Table 2.10) unless stated otherwise. All steps were performed under a roaring Bunsen burner using aseptic technique.

Table 2.22 Genotoxic agents used within this study and type of damage as indicated.

Chemical	Supplier	Stock concentration	Genotoxic effect
Azidothymidine (AZT)	Bio-technie	10 mg/ml	Chain-terminating nucleoside ¹ .
Hydrogen Peroxide (H ₂ O ₂)	Sigma- Aldrich	3% (w/w) or 0.98 M	Oxidative damage ¹⁵¹ .
Mitomycin C (MMC)	Sigma- Aldrich	2 mg/ml	DNA crosslinker ¹⁵² .
Rifampicin	Sigma- Aldrich	10 mg/ml	Transcription inhibitor ¹⁵³ .

2.8.1 Hydrogen peroxide viability spot assay

Overnight cultures were used to inoculate sterile lidded test tubes containing 8 ml of media to a 1:100 dilution. Cultures were allowed to grow to an OD₆₀₀ of 0.3-0.4 in a shaking water bath (Grant) at 37°C. Growth took between 1.5-2 hours monitored by spectrophotometry (Thermo Spectronic 20+ Photospectrometer). Once at OD, cells were placed on ice for 5 minutes to limit further growth. Variable growth rates was observed between cell lines so this allowed pausing until all cells had reached an appropriate OD. Once growth was completed, H₂O₂ was added at varying concentrations directly to the test tube. Initial concentrations varied between 0 and 25 mM H₂O₂, 12.5 mM was used in subsequent experiments. Following H₂O₂ addition, cells were allowed to grow for a further 30 minutes before being serial diluted into 1x M9 minimal salt to arrest growth. 10 µl of each serial dilution was spotted on agar plates, left to dry and grown overnight in a 37°C incubator. Plates were photographed on a white light pad (Clever Scientific).

2.8.2 Hydrogen peroxide growth curves

Cells were grown to OD₆₀₀ 0.3-0.4 using the method as stated in section 2.8.1. Cells were pipetted into a 24-well flat-bottomed plate (Falcon) and H₂O₂ was added to appropriate wells to a total volume of 500 µl. The 24-well plate was shifted to a FLUOstar microplate reader (BMG Labtech) where growth was monitored over 16 hours at 37°C with double orbital shaking at 200 rpm. OD₆₀₀ readings were taken every 30 minutes using a discrete wavelength without pathlength correction. After completion, data was extracted and analysed using Prism (GraphPad) software¹⁵⁴.

2.8.3 Mitomycin C viability spot assay

Cells were grown to OD₆₀₀ 0.3-0.4 using the method as stated in section 2.8.1 and serially diluted into 1x M9 salts. 10 µl of each serial dilution was spotted onto agar plates containing 0.25 µg/ml, 0.5 µg/ml and 1 µg/ml mitomycin C. Agar plates containing no mitomycin were also spotted as a control of normal viability. All plates were grown overnight at 37°C as stated in section 2.8.1. Following growth, colonies were counted using a digital colony counter (Stuart). Colony forming units were calculated and plotted using Prism (GraphPad) software¹⁵⁴.

2.8.4 Azidothymidine viability spot assay

Cells were treated as stated in section 2.8.3. 10 µl of each serial dilution were spotted onto agar plates containing 2.5 ng/µl, 5 ng/µl, 7.5 ng/µl, 10 ng/µl and 25 ng/µl azidothymidine and grown overnight. Colonies were counted and analysed as stated in section 2.8.3.

2.8.5 Rifampicin viability assay

Broth media was inoculated from overnight cultures as described in section 2.8.1. Cell growth was monitored by spectrophotometry as stated before, with 1 ml samples being taken at OD₆₀₀ 0.4 and 0.6 or after one or two days of growth. 1 ml samples were spread onto agar plates containing 10 µg/ml or 20 µg/ml rifampicin where stated. 1 ml samples were spread onto agar containing no rifampicin after one day of growth as a control. Spread agar plates were then grown overnight at 37°C in an incubator. Plates were imaged using a white light pad.

2.8.6 Genetic analysis of *Methanothermobacter thermautotrophicus* *lhr*

Genetic analysis was performed by Dr. Edward Bolt using methodology as detailed¹⁰⁷.

In summary, transformed *E. coli dnaE486 ΔrecQ* cells were grown in broth containing 50 µg/ml ampicillin to OD₆₀₀ 0.5 before plating 100 µl of culture onto each sector of an agar ampicillin plate. Plating was performed on three plates to allow overnight incubation at 30°C, 37°C and 45°C.

2.9 *In vitro* experimentation

2.9.1 Preparation of 5'-end labelled ³²P DNA substrates

DNA substrates containing ³²P end labelling were prepared by Dr. Edward Bolt as stated in¹¹⁰. In summary, 900 ng of one strand for each substrate was 5'-end labelled with ³²P using T4 polynucleotide kinase (NEB) and γ -³²P-ATP (20 μ l reaction volume). Reactions were loaded onto a BioSpin 6 column (BioRad) to allow separation of labelled DNA. DNA substrates were prepared by mixing ³²P labelled DNA with 900 ng of each appropriate unlabelled oligo in 1x SSC buffer, heated to 95°C for 5 minutes, and slowly cooled overnight from 95°C to room temperature to allow annealing. 'OG dye' was added to 1x before sample loading onto a standard 10% native acrylamide TBE gel. Samples were run for 2 hours at 150 V to allow ample migration. The gel was then exposed to autoradiography film and the developed film revealed the positions of the desired substrates for excision from the gel. Gel slices were soaked overnight at 4°C in 20 mM Tris HCl pH 7.5 for DNA elution. DNA in buffer was pipetted for separation from gel debris and quantified by scintillation counting using standard scintillation counts of samples taken throughout the procedure of known DNA mass (ng). This established the final yield of substrate DNA in ng that was converted to a final concentration of DNA (nM) for use in assays.

2.9.2 Preparation of 5'-end labelled Cy5 DNA substrates

DNA substrates listed in Table 2.9 containing a 5'-end labelled Cy5 moiety (●) were prepared by slow annealing overnight from 95°C to room temperature in 1x Annealing Buffer. 'OG dye' was added to 1x before sample loading onto a standard 10% native

acrylamide TBE gel. Samples were run for 3 hours at 140 V to allow ample migration. A sample of the single stranded 5' Cy5 DNA oligo used to make the DNA substrate was run alongside to allow distinction between fully formed substrate bands. Correctly sized bands were cut out and eluted into 'Fork elution buffer' at 4°C overnight. DNA in buffer was pipetted for separation from gel debris and quantified using the absorption reading at 260 nm, determined by a DeNovix DS-11 spectrophotometer, and the substrates extinction coefficients. These values were applied to the Beer-Lambert law to calculate μM concentration of DNA for use in assays.

2.9.3 DNA binding assays

Concentration gradient DNA binding assays were performed at 37°C in reaction mixtures containing HB-DTT, 12.5 nM Cy5-fluorescently labelled DNA substrate, 25 mM DTT and 5 mM EDTA. Tubes containing reaction mixture were pre-incubated at 37°C for 5 minutes before addition of protein. Reactions were held at 37°C for 20 minutes before being shifted on ice for a further 10 minutes. 'OG dye' was added to 1x before sample loading. Products were analysed using a standard 5% native acrylamide TBE gel ran for 1 hour 30 minutes at 140 V to allow sufficient migration. Gels were imaged using a Typhoon phosphor-imager (Amersham) at 633 nm using a R765 filter for Cy5 detection.

2.9.4 DNA unwinding assays

For *Methanothermobacter thermautotrophicus* Lhr.

DNA unwinding assays were performed at 45°C in reaction mixtures containing HB-DTT, variable ³²P labelled DNA substrate (concentrations as stated in figure legends), 2 mM DTT, 2 mM ATP and 1 mM MgCl₂. Reactions were initiated upon addition of *MthLhr* (or *EcoRuvAB* as control) and held at 45°C for 20 minutes for protein concentration reactions and for up to 5 minutes during time course assays. Protein concentration titration assays were stopped via de-proteinisation upon addition of stop solution (4 µl per 20 µl reaction). Time course assays were stopped by addition of 18 µl of reaction to pre-aliquoted stop solution at each time point. 'OG dye' was added to 1x before sample loading. Products were analysed using a standard 10% native acrylamide TBE gel ran for 1 hour at 150 V to allow sufficient migration. Gels were dried under a vacuum on a flatbed dryer before being imaged on a Storm scanner (Amersham). TIFF images were analysed using GelAnalyzer 19.1 (Lazar) software¹⁵⁵. Graphs of % unwinding were generated using Prism (GraphPad) software¹⁵⁴.

For *Escherichia coli* Lhr.

Concentration gradient DNA unwinding assays were performed at 37°C in reaction mixtures containing HB-DTT, 12.5 nM Cy5-fluorescently labelled DNA substrate, 25 mM DTT, 1.25 µM unlabelled 'trap' DNA, 5 mM ATP and 5 mM CaCl₂. Tubes containing HB-DTT, Cy5-fluorescently labelled DNA substrate, DTT, CaCl₂ and protein were pre-incubated at 37°C for 5 minutes before reactions were initiated upon addition of unlabelled trap and ATP. Reactions were held at 37°C for 30 minutes before being

stopped through de-proteination by addition of stop solution (4 μ l per 20 μ l reaction). 'OG dye' was added to 1x before sample loading. Products were analysed using a standard 10% native acrylamide TBE gel ran for 45 minutes at 150 V to allow sufficient migration. Gels were imaged using a Typhoon phosphor-imager at 633 nm using a R765 filter for Cy5 detection.

2.9.5 DNA glycosylase assays

Time course and protein concentration titration assays were performed at 37°C in reaction mixtures containing HB-DTT, 12.5 nM Cy5-fluorescently labelled DNA substrate, 25 mM DTT, 5 mM ATP, 4 mM MnCl₂ and 4 mM CaCl₂. Tubes containing reaction mix were pre-incubated at 37°C before being initiated by addition of Lhr protein. For protein concentration titrations, reactions were held at 37°C for 30 minutes before being stopped through de-proteination by addition of stop solution (4 μ l per 20 μ l reaction) and 4 μ l of 1 M NaOH. Reaction samples were boiled for 5 minutes before 6 μ l of Formamide stop was added. 30 μ l of reaction sample was loaded onto a 15% denaturing acrylamide TBE gel and ran for 4 hours at 5 watts per gel to allow sufficient migration for analysis. Denaturing gels were set overnight, wells washed with copious amounts of distilled water and pre-ran for 1 hour at 5 watts before loading. Each well was washed again prior to sample addition to allow even migration of DNA samples for accurate sizing determination. A Cy5-fluorescently labelled DNA ladder was ran alongside reaction samples to aid in DNA size determination. Gels were imaged using a Typhoon phosphor-imager at 633 nm using

a R765 filter for Cy5 detection. TIFF images were analysed using GelAnalyzer 19.1 software¹⁵⁵. Graphs of glycosylase activity were generated using Prism software¹⁵⁴.

2.10 Protein overexpression and purification

2.10.1 Obtaining *M. thermautotrophicus* Lhr protein

Methanothermobacter thermautotrophicus gene 1802 (pEB352) was previously amplified using polymerase chain reaction (PCR) and cloned into pET22b using NdeI and BamHI restriction sites by Dr. Edward Bolt. pEB352 was transformed into competent *E. coli* BL21 C+ cells using the protocol described in section 2.5.2.

Pilot overexpression and purification methods were optimised during my MRes. Protocol is as quoted and presented in¹⁵⁶ with minor tweaks.

‘Successful transformants were subjected to a pilot overexpression. Single colonies were grown overnight in 5 ml LB broth with 50 µg/ml ampicillin and 35 µg/ml chloramphenicol at 37°C in a test tube rotator. Overnight cultures were used to inoculate fresh LB broth in a 1:100 dilution and grown until OD₆₀₀ 1.1 monitored by spectroscopy. Expression of *MthLhr* was induced upon addition of L-arabinose to 0.2% and isopropyl β-D-1-thiogalactopyranoside (IPTG) (VWR) to 0.8 mM. Growth was allowed to continue overnight at 28°C. Successful overexpression was confirmed by SDS PAGE analysis. Cells from 1 ml samples were pelleted through centrifugation and resuspended in 150 µl SDW, 4x SDS loading buffer and 30 mM DTT. Samples were heated at 95°C for 10 minutes, aliquots were run on a standard 8% SDS gel before being stained with SimplyBlue™ SafeStain (Invitrogen) using the microwave protocol¹⁵⁷. BL21 C+ cells transformed with pBAD/HisA was subjected to the same overexpression protocol, SDS samples of which were run alongside as a negative control. A blue protein standard ladder (NEB) was used to confirm sizes of overexpressed protein.

A 36x upscale of this method was then performed to obtain sufficient cell biomass for protein purification. After overnight overexpression, 1 ml samples were taken for analysis, the remaining culture was centrifuged using an Avanti J-26 XP centrifuge (Beckman Coulter) using a JLA 10.500 rotor for 12 minutes at 1800 RCF. Cell pellets were resuspended in 15 ml of 'Low salt buffer A' with half of a cComplete™, EDTA-free Protease Inhibitor Cocktail to inhibit unwanted protease activity and flash frozen using dry ice for storage at -80°C.

To purify, frozen biomass was thawed on ice and sonicated using a Vibra cell sonicator (Jencons). Samples were then clarified by centrifugation using an Avanti J-26 XP centrifuge (Beckman Coulter) with a JA 25.50 rotor at 21400 RCF for 25 minutes. *MthLhr* containing supernatant (S1) was decanted and loaded onto a 5ml Heparin column (GE Healthcare Life Sciences) which had been pre-equilibrated with 'Low salt buffer A'. Flow-through (eluate from loading S1) and wash-through (eluate containing weakly bound protein) were collected for analysis. *MthLhr* was eluted over a gradient of increasing ionic strength, achieved using an increasing concentration of 'High salt buffer B'. *MthLhr* containing fractions (determined by SDS PAGE analysis) were pooled for dialysis. Dialysis was performed overnight at 6°C on a magnetic stirrer using Dialysis Tubing D104 (BioDesign) against 'Low salt buffer A'. Dialysed solution was then loaded onto a 1 ml HiPrep Q sepharose column (GE Healthcare Life Sciences) pre-equilibrated with Low salt buffer A. Flow- and wash-through were collected as before. *MthLhr* was again eluted over a gradient of increasing ionic strength and *MthLhr* containing fractions were pooled and dialysed overnight against 'Low salt buffer A' containing 35% glycerol. Dialysed purified protein was then aliquoted and flash frozen using dry ice for storage at -80°C. A sample of purified protein was tested using a DeNovix DS-

11 spectrophotometer to give a 280nm absorption value. This value, along with *MthLhrs* extinction coefficient (93100), were applied to the Beer-Lambert law to calculate approximate protein concentration. Lhr was confirmed following trypsin digestion and mass spectrometry carried out by S. Ashra (Core Biotechnology Services, University of Leicester). This purification method typically yielded 628 µg of purified Lhr protein per L of cell culture.'

2.10.2 Obtaining *E. coli* full length Lhr protein and D1536A mutant

Escherichia coli gene b1653 (pRJB15) was amplified from *E. coli* MG1655 genomic DNA using polymerase chain reaction (PCR) and cloned into pT7-7 using NdeI and HindIII restriction sites for this work. pRJB15 was transformed into competent *E. coli* Rosetta 2 cells using the protocol described in section 2.5.2.

Successful transformants were subjected to a pilot overexpression. Single colonies were grown overnight in 5 ml LB broth with 50 µg/ml ampicillin and 35 µg/ml chloramphenicol at 37°C in a test tube rotator. Overnight cultures were used to inoculate fresh LB broth (containing only ampicillin) in a 1:100 dilution and grown until OD₆₀₀ 1.2 monitored by spectroscopy. Cultures were transferred to an ice slurry for cooling before addition of 0.8 mM IPTG. Growth was allowed to continue overnight at 18°C, samples were taken and analysed as described in section 2.10.1 with pT7-7 empty vector as a negative control.

A 24x upscale of this method was then performed to obtain sufficient cell biomass for protein purification. Cells were harvested as stated in 2.10.1 and resuspended into 15

ml of 'Ni-NTA buffer A' with 0.1 mM phenylmethylsulfonyl fluoride (PMSF) to inhibit unwanted protease activity.

EcoLhr was purified in both a non-tagged (pRJB15) and a hexa-His tagged (pRJB28) form. The two purification protocols are as follows:

For non-tagged: frozen biomass was thawed on ice, sonicated (Vibra cell- Jencons) and clarified by centrifugation as described for *MthLhr*. *EcoLhr* containing supernatant (S1) was decanted before loading onto a 5 ml Butyl sepharose column which had been equilibrated with 'Hydrophobic salt buffer A'. *EcoLhr* was eluted upon decreasing [ammonium sulphate] using 'No salt buffer B'. Fractions containing Lhr were pooled and dialysed overnight into 'Low salt buffer 2A' and loaded onto a 1 ml Heparin column. *EcoLhr* eluted early during an increasing gradient of potassium acetate using 'High salt buffer 2B'. Appropriate factions were pooled and dialysed overnight again against 'Low salt buffer 2A' for loading onto a Q-sepharose column. *EcoLhr* was again eluted using an increasing gradient of potassium acetate using the same buffers, appropriate factions were pooled and dialysed overnight into 'Low salt buffer 2A' with 35% glycerol for aliquoting. Protein was flash frozen and stored at -80°C. A sample of purified protein was tested using a DeNovix DS-11 spectrophotometer to give a 280 nm absorption value. This value, along with *EcoLhr*'s extinction coefficient (178105), were applied to the Beer-Lambert law to calculate approximate protein concentration.

For hexa-His tagged Lhr wild type and D1536A mutant: frozen biomass was thawed on ice, sonicated (vibra cell- Jencons) and clarified by centrifugation as described for

MthLhr. *EcoLhr* containing supernatant (S1) was decanted before loading onto a 5 ml Butyl sepharose column which had been equilibrated with 'Hydrophobic salt buffer A'. The column was washed with 60% 'No salt buffer B' to remove contaminating proteins. A 5 ml Ni-NTA column pre-charged (using 2x column volumes of 0.1 M NiCl₂) and pre-equilibrated with 'Ni-NTA buffer A His-Lhr' was attached in tandem and both columns were washed with 100% 'No salt buffer B'. After sufficient washing (AKTA UV trace bottomed off), the 5 ml Butyl sepharose column was removed and *EcoLhr* was eluted from the Ni-NTA column over an increasing [imidazole] using 'Ni-NTA buffer B His-Lhr'. *EcoLhr* containing fractions were pooled and dialysed overnight into 'Low salt buffer 2A' and loaded onto a 1 ml Q sepharose column. *EcoLhr* was eluted using an increasing gradient of potassium acetate using 'High salt buffer 2B'. Appropriate fractions were pooled and dialysed overnight into 'Low salt buffer 2A' with 35% glycerol for aliquoting. Protein was flash frozen and stored at -80°C. A sample of purified protein was tested using a DeNovix DS-11 spectrophotometer to give a 280 nm absorption value. This value, along with *EcoLhrs* hexa-his extinction coefficient (178105), were applied to the Beer-Lambert law to calculate approximate protein concentration.

2.10.3 Obtaining purified *E. coli* C-terminal Lhr protein and D1536A mutant

Escherichia coli gene b1653 C-terminus (amino acids 876-1538) was amplified and cloned by ligation independent cloning (LIC) by Nadia Ahmed in Dr. Christopher Cooper's lab (University of Huddersfield). The resulting pNH-TrxT plasmid based

vector (named pRJB23 for this work) was transformed into competent *E. coli* Rosetta 2 cells using the protocol described in section 2.5.2.

Overexpressions of *Eco* C-terminal Lhr and D1536A mutant were carried out as stated in section 2.10.2 using 40 µg/ml kanamycin instead of ampicillin and 0.1 mM IPTG to induce expression.

Once pilot overexpression was confirmed a 20x upscale was performed with cells harvested and resuspended into 'Ni-NTA buffer A' with 0.1 mM PMSF and flash frozen for storage at -80°C.

To purify, frozen biomass was thawed on ice, sonicated (vibra cell- Jencons) and clarified by centrifugation as described previously. *Eco* C-Lhr containing supernatant (S1) was decanted before loading onto a 5 ml Ni-NTA column which had been pre-charged (using 2x column volumes of 0.1 M NiCl₂) and equilibrated with 'Ni-NTA buffer A His-Lhr'. *Eco* C-Lhr was eluted upon increasing [imidazole] using 'Ni-NTA buffer B His-Lhr'. Fractions containing wild type *Eco* C-Lhr were pooled and dialysed overnight into 'Low salt buffer A' containing 35% glycerol for storage.

Purification of mutant *Eco* C-Lhr D1536A included an extra step as compared to WT C-Lhr. For this, Ni-NTA fractions were pooled and dialysed into 'Low salt buffer A' and loaded onto a 1 ml Q sepharose column. Pure *Eco* C-Lhr D1536A was collected by an increasing gradient of potassium acetate. Appropriate fractions were pooled and dialysed overnight against 'Low salt buffer A' with 35% glycerol.

Protein concentrations were calculated using the method previously described using an extinction coefficient of 88475.

2.10.4 Obtaining purified *E. coli* RNaseT protein

Escherichia coli gene b1652 (pEB695) was amplified from *E. coli* MG1655 genomic DNA using polymerase chain reaction (PCR) and cloned into pET14b using NdeI and BamHI restriction sites by Dr. Edward Bolt. pEB695 was transformed into competent *E. coli* BL21 AI cells using the protocol described in section 2.5.2.

Overexpressions of *Eco*RNaseT were carried out in a similar fashion to that stated in section 2.10.1. *E. coli* BL21 AI cells transformed with pEB695 were grown with 50 µg/ml ampicillin until OD₆₀₀ 0.6 before cooling in an icy slurry and induction by addition of L-arabinose to 0.2% and IPTG to 0.8 mM. Growth was allowed to continue overnight at 18°C before testing.

Once pilot overexpression was confirmed a 20x upscale was performed with cells harvested and resuspended into 'Ni-NTA buffer A' with 0.1 mM PMSF and flash frozen for storage at -80°C.

Purification was performed as stated in section 2.10.3, using the C-Lhr D1536A protocol. Protein concentration was determined using a Bradford Protein assay using a standard protocol. This method was key to accurately determining protein concentration due to the proteins lack of tryptophan and tyrosine residues.

2.10.5 Obtaining purified *E. coli* RadA (Sms)

Escherichia coli gene b4389 (pRJB14) was amplified from *E. coli* MG1655 genomic DNA using polymerase chain reaction (PCR) and cloned into pACYC Duet using BamHI and

HindIII restriction sites for this work. pRJB14 was transformed into competent *E. coli* BL21 AI cells using the protocol described in section 2.5.2.

Overexpression of *EcoRadA* (Sms) was performed as stated in section 2.10.4 using *E. coli* BL21 AI cells transformed with pRJB14 and 35 µg/ml chloramphenicol instead of ampicillin.

Purification was performed as stated in section 2.10.3, using the C-Lhr D1536A protocol. Protein concentration was determined by testing the 280nm absorption using the method previously described and 22960 as extinction coefficient.

Chapter 3 Bioinformatic study of RecA/Rad51 family proteins and bacterial Lhr helicase

3.1 Phylogenetic analysis of RecA/Rad51 family of proteins

3.1.1 Introduction to RecA/Rad51 family of proteins

The RecA/Rad51 family are a conserved set of proteins which confer multiple important roles in genome maintenance and DNA damage repair^{98,158}. The family can be subdivided into proteins which promote DNA strand exchange and their paralogue regulators. All proteins within this family, although often diverse in function, share a high level of conservation within their active site, situated within the 'RecA-like core'^{7,96}. Protein function often depends upon the proteins ability to bind and hydrolyse ATP, using the Walker A and Walker B motifs¹⁵⁹. The relatedness of the Walker A/B motifs allow phylogenetic mapping and can potentially indicate function. The Rad51 paralogues, displaying 20-30% homology to their recombinase counterparts, have seen extensive study in eukaryotes, owing to the synthetic lethality shown in human knockout cells¹⁵⁸. The Rad51 paralogues are thought to originate through gene duplication events and have since developed specific key roles. Eukaryotes often have multiple Rad51 paralogue proteins per species. Limited study in archaea has identified a handful of paralogues, with RadB being the most well-known^{160,161}. Rad51 paralogues appear starkly absent from bacteria although candidates are beginning to be identified for study^{1,162-165}.

3.1.2 A brief introduction to *E. coli* RadA (Sms)

Bacterial RadA, distinct from the archaeal recombinase of the same name, has been identified in several organisms namely *E. coli*, *Bacillus subtilis*, and *Deinococcus radiodurans*¹⁶⁴. It has gathered recent interest due to it being one of the only known bacterial Rad51 paralogue proteins and the novelty of its zinc-finger motif and Lon protease-like domains, which are not found in its eukaryotic or archaeal counterparts^{164–167}.

RadA (Sms) has been shown to be involved in recombinational repair through the sensitivities of knockout strains to various DNA damaging agents^{1,163,168}. This is further supported by the synergistic phenotypes seen when knocked out in combination with other known HR repair proteins such as *ruvABC* and *recG* in *E. coli*^{163,169}. This involvement has also been seen in *B. subtilis*¹⁷⁰. Further expression analysis has linked *E. coli* RadA (Sms) with a circadian rhythm function similar to the 'KaiC' RecA/Rad51 paralogue protein found in cyanobacteria^{98,171,172}. Recent biochemical study has shown the ability of RadA (Sms) to form hexameric rings facilitated through interactions of the Lon protease-like domain, allowing action as a DnaB-type DNA helicase^{164,165,167}. Further characterisation showed RadA (Sms)'s ability to promote RecA recombinase activity onto a D-loop substrate, remodelling this HR intermediate and mediating branch migration^{165,167}. RadA (Sms) as yet has not been investigated within the context of the RecA/Rad51 family as a whole or its biological relationship with Lhr, a protein which is the main focus of this study.

3.1.3 Construction of phylogenetic trees

Phylogenetic analysis was performed using “One Click” mode on ‘phylogeny.fr’ which optimises alignments using software, only requiring the user to input a list of query sequences. This is as opposed to other options on ‘phylogeny.fr’ which require a greater level of user input for tailored interrogation. To generate phylogenetic trees, homologous regions need to be identified, aligned, and given phylogenetic scores which depict ‘relatedness’. This data is then processed and represented visually.

The ‘phylogeny.fr’ website is aimed at the biologist with limited or no prior experience in creating phylogenetic trees, only requiring BLAST sequences for initial input. This website chains multiple programs together allowing fast and accurate tree generation.

Protein sequence alignment is performed by ‘MUSCLE 3.8.31’, generating homologous region hits¹⁴⁵, these are then refined by ‘Gblocks 0.91b’ which eliminates any poorly aligned regions. Although this step essentially reduces information from the analysis, it bolsters the phylogenetic signal to give a more accurate picture of the relationship between highly conserved regions. Extensive discussion and justification can be found here ^{147,148}. Tree generation is performed by ‘PhyML 3.1/3.0 aLRT’. This software is based on the maximum likelihood principle¹⁴⁹, this is then visualised using ‘TreeDyn 198.3’.

3.1.4 Conservation of Walker A/B active sites

MUSCLE alignment identifies key conserved residues located within the Walker A and Walker B active sites. Conserved residues confer both structural and enzymatic functions allowing effective ATP orientation, binding and hydrolysis. Degree of conservation considering all query sequences used allows protein pairing and phylogenetic relationships to be displayed (see Figure 3.3 and Figure 3.4).

As shown in Figure 3.1, considerable conservation is found within the P-loop of the Walker A motif (red), presenting as 'GX₄GK(S/T)' where x is any residue. The P-loop forms contacts to the γ -phosphate of ATP through hydrogen bonding. This action is complemented by the hydrophobic β -sheet of the Walker B (' ϕ_4 D', where ϕ is hydrophobic) coordinating a Mg²⁺ ion and activates the attacking water molecule for hydrolysis^{173,174}. Interestingly, *E. coli*'s RadA (Sms) does not have a catalytic aspartate residue within the Walker B and is instead replaced by an inert glycine. This may explain why its ATP hydrolysis activity is markedly absent¹.

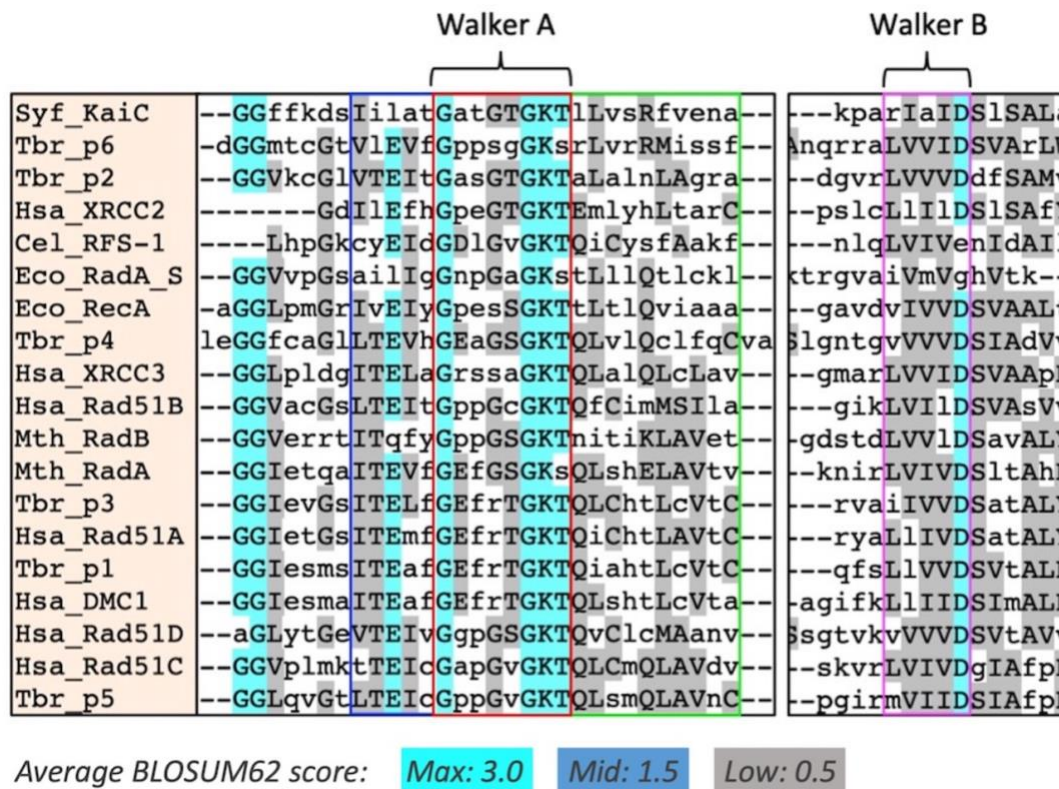


Figure 3.1 Identification of highly conserved residues located within the Walker A/B active site using multiple sequence alignment.

Multiple protein alignment of RecA/Rad51 family proteins generated by MUSCLE and refined by Gblocks. Conserved residues identified by average BLOSUM62 score, graded on amount of sequence similarity. Walker A/B conservation shown here as an example. Generated during phylogenetic interrogation.

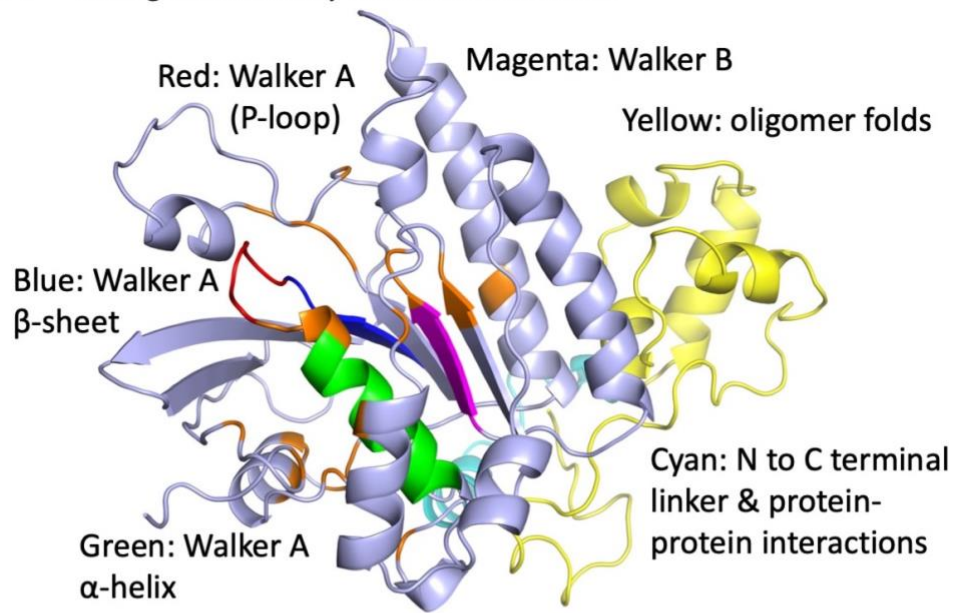
Phyre² is a protein modelling prediction software which identifies homologous proteins with known structures to allow 3D visualisation of the query amino acid sequence. Results include a generated '.pdb' file which can be opened using software such as 'PyMOL'. Figure 3.2 shows a Phyre² predicted model of *H. sapiens* Rad51. Initially, this structure was used to compare and contrast the *T. brucei* Rad51 paralogue proteins with their human counterparts aiming to identify novel or conserved structures. This figure now serves to highlight individual functional domains

and to visually map conserved residues as identified through phylogenetic analysis (Figure 3.1).

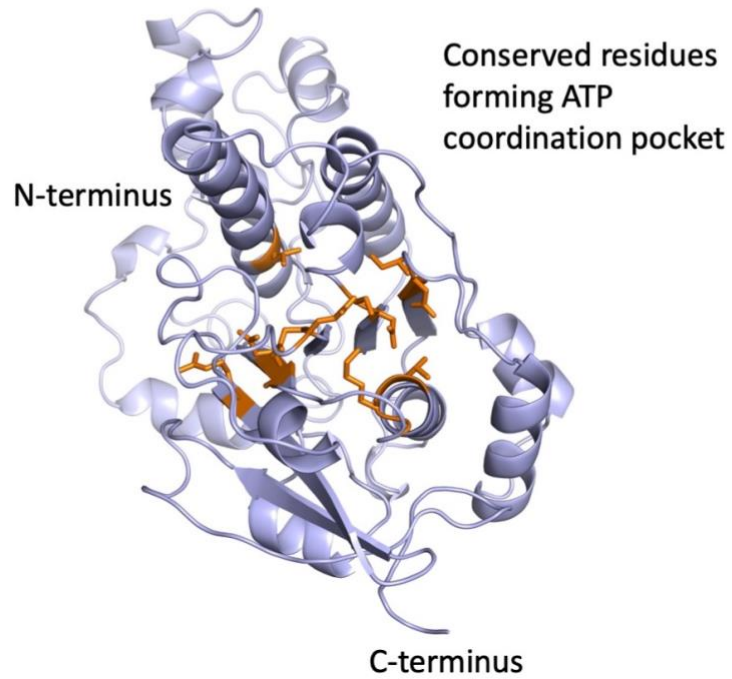
Panel **A** of Figure 3.2 depicts the typical architecture of a Rad51/RecA family protein. Protein-protein interaction and oligermsation motifs are coloured in **cyan** and **yellow** respectively. These fringe domains present the most diversity in amino acid composition. Highly conserved residues are highlighted in **orange** as identified by MUSCLE and Gblocks.

Figure 3.2 **B** identifies the key residues which form the ATP coordination pocket. These amino acids serve as both structural anchors and coordination/active site residues. Conserved residues which lie outside of the Walker A and Walker B motifs make distinct contacts between the α -helices and β -sheets surrounding the ATP binding site for correct orientation.

A Orange: absolutely conserved residues



B



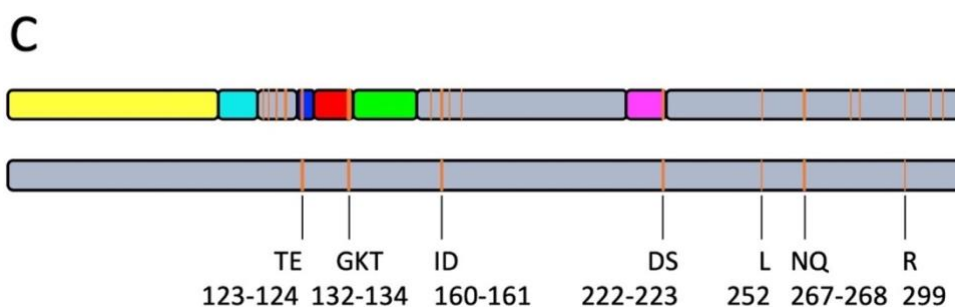


Figure 3.2 Mapping of conserved residues identified by multiple sequence alignments.

Phyre² annotated structural model of *H. sapiens* Rad51 (**A**) Blue, red, green and magenta highlights conserved residues identified within the functional Walker A and Walker B motifs, orange shows conserved residues throughout the remaining structure. Cyan and yellow labelling highlight additional domains. (**B**) Side chain orientation of orange conserved residues which form the ATP binding pocket. (**C**) Gene profile of Rad51 showing highly conserved amino acids (orange bands) as relating to structure in **A** (top bar) and which of those form the ATP binding pocket (**B**, bottom bar). Colour labelling consistent with that as annotated in **A**. Models visualised and annotated using PyMOL¹⁷⁵.

Just under half of all identified highly conserved residues make up the Walker A and Walker B ATP binding and hydrolysis active site (Figure 3.2 **C**). The presence, or absence of these residues may allow proteins of unknown function to be paired with characterised Rad51/RecA paralogues enabling guided study. A large majority of protein structure remains highly changeable, owing to the diversity of proteins and pathways that the 'RecA-like' domain is found in.

An interesting highly conserved residue which is distinctly removed from the ATP binding site is R299. Upon further investigation using PyMOL¹⁷⁵, R299 represents the 'arginine finger' residue. Recombinogenic Rad51/RecA family proteins rely on an arginine from the adjacent protein within the oligomer to initiate ATP hydrolysis^{176,177}.

In other AAA+ superfamily proteins this role may be carried out by small effector proteins. The arginine finger is an essential residue allowing catalytic acceleration through interactions with the γ -phosphate. This interaction may be through polarisation or stabilisation of a transition state¹⁷⁸.

R299 is aided in its positioning by TE 123-124, which forms contacts between two posterior β -sheets (using Figure 3.2 **A** as reference) to extend R299 away from the surrounding structure for contacts with the adjacent oligomer subunit.

3.1.5 Analysis of *T. brucei* as a model for Rad51 paralogue proteins

Phylogenetic trees are generated by PhyML which is based on the maximum likelihood principle. Tree generation using this principle evaluates the data set to determine the outcome with the highest statistical probability. This analysis can be processed much faster than other estimates such as maximum parsimony, whilst still maintaining a high degree of accuracy allowing efficient, large scale phylogenetic tree generation^{149,179}. Phylogenetic iteration is then displayed using TreeDyn which allows extensive editing and annotation¹⁵⁰.

Phylogenetic trees displayed below (Figure 3.3 and Figure 3.4) are rooted phylograms meaning branch lengths reflect estimated evolutionary change, this relationship is rooted by a predicted common ancestor (furthest left of each figure). Each node (branch point) represents a predicted common ancestor between the two corresponding proteins, these two proteins together form a clade^{180,181}. Each branchpoint or clade is assigned a Bootstrap value (highlighted in red). This value represents their statistical reliability, i.e. 'if the alignment was repeated 'x' amount of times, does the outcome/relationship still hold true?'. Reliability increases as the Bootstrap value tends towards 1. The scale indicates the number of changes occurring along the branch point, '0.3' as in Figure 3.3, denotes a change of 3 bases in every 10 per given scale length¹⁸². This number may seem high for a conserved family of proteins but it may also reflect the high degree of variation seen in Rad51/RecA family proteins outside the active conserved sites. These changes often dictate specialised function within repair pathways through differing protein-protein interactions.

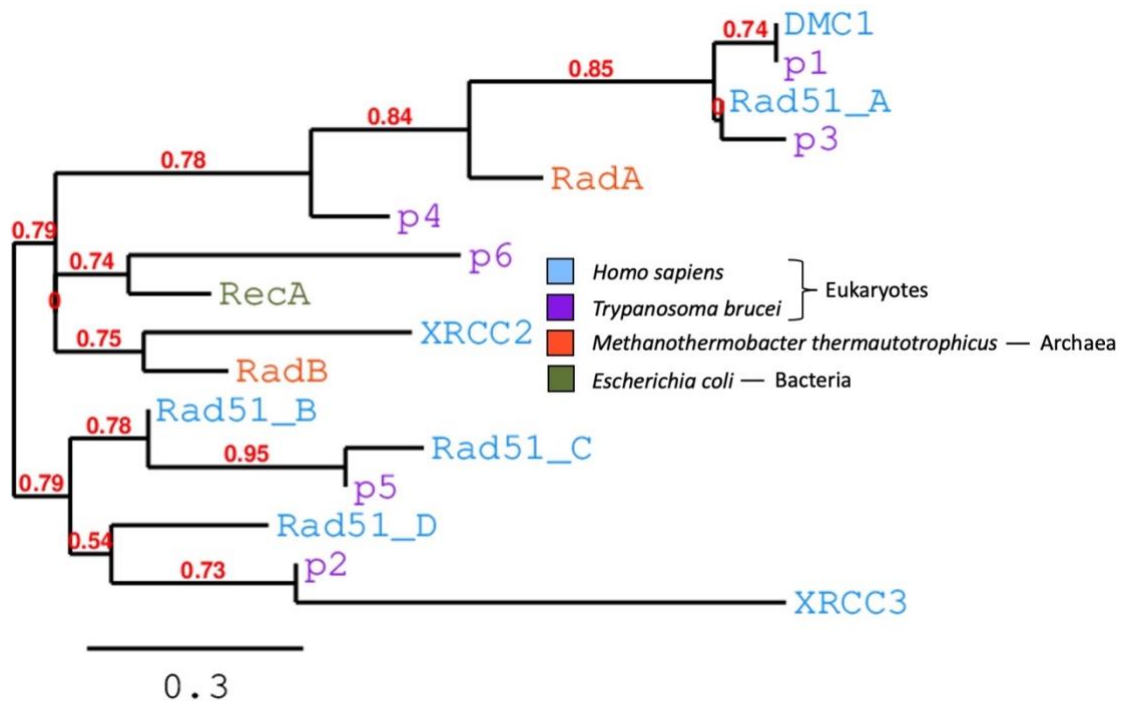


Figure 3.3 Analysis of *Trypanosoma brucei* Rad51 paralogue candidates by phylogenetics.

Blastp identified *T. brucei* proteins of interest were subject to “one click” phylogenetic analysis on ‘phylogeny.fr’¹⁴⁴. Other RecA/Rad51 family proteins selected for sufficient phylogenetic context. Tree generation facilitated by PhyML¹⁴⁹ and annotated by TreeDyn¹⁵⁰.

T. brucei was initially chosen to study as a potential model for *H. sapiens* Rad51 paralogues due to the presence of 6 known Rad51/RecA like proteins. Our theory was, one would be a bona fide Rad51 recombinogenic protein and the remaining five would be equivalent to the five human Rad51 paralogues. As shown in Figure 3.3, these *T. brucei* proteins are limited in their suitability.

‘p1’ (paralogue 1), ‘p3’ and ‘p4’ show a close evolutionary relationship with *H. sapiens* Rad51 and DMC1, as well as *M. thermautotrophicus* RadA. ‘p6’ is more closely related to bacterial RecA. This data is supported with Blastp results and can be justified when

considering the complex life cycle of trypanosomes as they move through transmission and into the multiple stages of infection¹⁸³.

An interesting relationship can be seen between 'p5' and Rad51 C and 'p2' and XRCC3. These two proteins in *H. sapiens* form an important complex with multiple key functions such as in promoting repair via recombination¹⁸⁴.

It must be noted that in Figure 3.3, a few of the branch points show Bootstrap values of '0' (such as between Rad51A and 'p3'). This may be due to the promiscuity of proteins involved, i.e. they align well with more than one query protein but they are paired and displayed in this way due to the strength of the node before. Rad51 and 'p3' may be the 'frame work' proteins from which the paralogues are based so would justifiably match multiple other proteins. It is also interesting that none of the *T. brucei* candidate proteins match well with each other meaning they are not just a result of gene duplication and that each protein has a defined, specific role to play within the cell.

3.1.6 Identification of 'Rad51 C' and 'XRCC3' subfamilies

Further phylogenetic iterations were performed to interrogate the emergence of potential Rad51 C and XRCC3 subfamilies as seen in Figure 3.3. This was achieved by adding more examples of Rad51/RecA paralogues as identified from other model organisms. New paralogue proteins were selected to diversify the functional properties of the queried protein list instead of adding, for example, multiple more recombinogenic proteins.

It may be argued that to fully understand the phylogenetic relationship you will need to add all examples of Rad51/RecA proteins to the study. However, when adding more proteins to the initial query (data not shown), it was noticed that MUSCLE and Gblocks began to align additional structural elements which only muddied the water away from the important catalytic sites. The resulting trees served as only a visual representation of what may be obtained from a simple Blastp.

Figure 3.4 represents a snapshot of the Rad51/RecA family of proteins providing hints into a new angle of studying these paralogue candidates by classing them under subfamilies to enable protein characterisation in a more targeted manner.

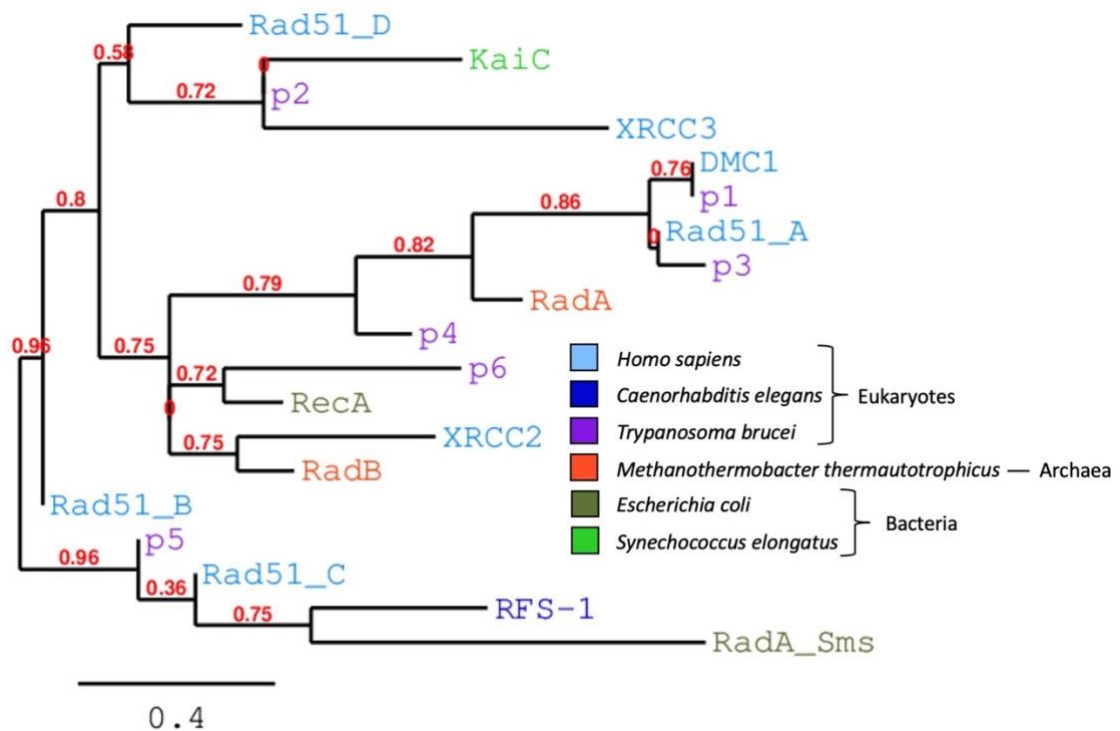


Figure 3.4 Analysis of RecA/Rad51 family of proteins from a diverse set of organisms.

Suggestion of 3 subfamily trees, Rad51-like (**central**), XRCC3-like (**top**) and Rad51C-like (**bottom**). Protein sequences queried using “one click” phylogenetic analysis on ‘phylogeny.fr’¹⁴⁴. Tree generation facilitated by PhyML¹⁴⁹ and annotated by TreeDyn¹⁵⁰.

Rad51/RecA family proteins can be separated into three distinct branches, Rad51-like (middle), XRCC3-like (top) and Rad51 C-like (bottom). Each species listed appears to have differing numbers of Rad51/RecA family proteins, this may be due to the complexity of each organism and the most common form of DNA damage encountered. Additional proteins from this family may be yet to be discovered.

E. coli RadA (Sms) shows the highest level of evolutionary change within its branch and in Figure 3.4 as a whole. This protein has been shown to have an additional helicase function, interacting with RecA and promoting branch migration in HR^{165,167}. As this protein forms part of the Rad51 C subfamily, studying it and its protein partners may

shed light onto its human counterpart. Data gathered may allow study into Rad51 C from a previously unknown way.

3.2 Distribution of Lhr in bacteria and archaea

3.2.1 Introduction to *Large Helicase Related (Lhr)* proteins

Lhr is a highly conserved DNA repair helicase found throughout archaea and in bacteria^{110,121,134}. Bacterial Lhr shows conservation within its genomic context, situated adjacent to additional DNA repair enzymes but this relationship is not conserved between all archaeal counterparts^{121,134}. *E. coli* Lhr is a non-essential protein yielding little observable phenotypes when knockout alone^{1,119}. *E. coli* Lhr is situated with an operon, downstream of RNaseT, an enzyme thought to be involved in tRNA maturation but no functional link has been described^{119,185}.

3.2.2 Study of the genomic context of Lhr across multiple species

Lhr has been identified and studied across a range of organisms¹²¹. Its genomic context is intriguing, often forming part of an operon with other DNA modifying enzymes¹³⁶.

To investigate if this relationship is conserved, we visualised Lhr and adjacent proteins using the genome browser feature on ‘biocyc.org’¹⁸⁶.

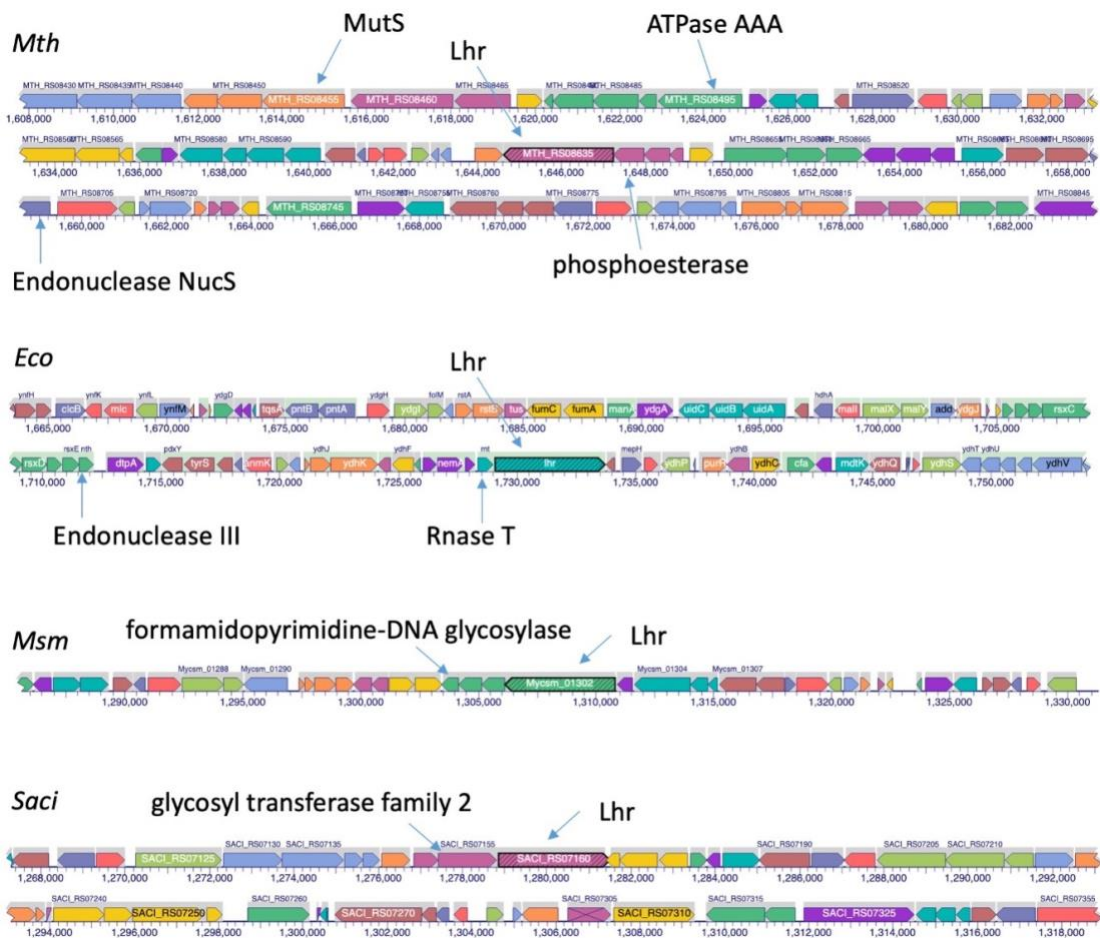


Figure 3.5 Lhr genomic context in multiple organisms.

Lhr and adjacent interesting proteins as annotated. Genome maps generated using ‘biocyc.org’¹⁸⁶. *Mth* – *Methanothermobacter thermautotrophicus*, *Eco* – *Escherichia coli*, *Msm* – *Mycobacterium smegmatis*, *Saci* – *Sulfolobus acidocaldarius*.

Lhr is commonly found as part of an operon. Gene clusters which form the operon are transcribed together producing a single mRNA which is later translated to produce individual proteins^{187,188}. Genetic arrangements such as this allow co-regulation of multiple genes for rapid responses to changes in the environment¹⁸⁸. Identification of conserved operons between multiple diverse organisms can allow predictions of genes of unknown function¹⁸⁹. Additionally, high levels of conservation may give an indication of importance to the respective genes. Lhr is found adjacent to multiple DNA modifying enzymes suggesting a potential regulated repair pathway.

Data presented in Figure 3.5 is in close agreement with a published observation presented in Ejaz et al. (2018)¹³⁶ however, a more recent phylogenetic study delved deeper, looking at instances of *lhr* in both bacteria and archaea. They suggest the genomic context is not conserved across all instances of *lhr* in archaea but some conservation was seen when looking at specific Lhr subtypes. They identified two genes which were in close proximity to *lhr* but were not conserved in relative positioning (upstream or downstream). One protein of importance belongs to the metallophosphatase (MPP) superfamily. These nucleases are able to hydrolyse phosphomono-, phosphodi- or phosphotri-esters using a metal co-factor allowing the targeted breakdown of nucleic acids^{190–192}. The presence of an MPP was also observed in bacteria further supporting previous published results¹³⁶. The second gene of interest was a glycosyltransferase thought to be involved in cell biogenesis but no speculation was made on its relevance¹²¹. Characterisation and identification of Lhr protein subfamilies may explain the significance of Lhr's variable genomic context. Once ascertained, this may allow more informed investigation of uncharacterised Lhr proteins.

As mentioned previously and displayed in Figure 3.5 and Figure 3.6, *E. coli* Lhr is present within an operon and is situated downstream of Rnt. RnT (RNaseT) is a DEDD superfamily, 3' to 5' hydrolytic exoribonuclease which has functions in multiple RNA metabolic processes such as a tRNA maturation and rRNA processing and is also required for normal cell growth^{193,194}. RnT is able to act redundantly across a wide range of functions in cells lacking other RNases and is able to reduce sensitivity to UV irradiation in *recJ*, *exo I*, *exo VII* deficient cells, suggesting a wider role in DNA metabolism^{195,196}. RnT's ability to participate in UV damage repair may be due to its ability to trim 3' ended structured DNA substrates¹⁹⁷. A functional link between the two proteins activities is yet to be reported.

3.2.3 Analysis of *E. coli* Lhr gene regulation and codon usage

The 'rnt-lhr' operon is controlled via two potential promoters. The 'rntp' promoter is located 27 base pairs upstream of *rnt*'s start codon and is bound by RpoD. This is a σ^{70} family sigma factor and is often utilised for proteins involved during exponential growth¹⁹⁸ and has the strongest interaction with the RNA polymerase core complex¹⁹⁹.

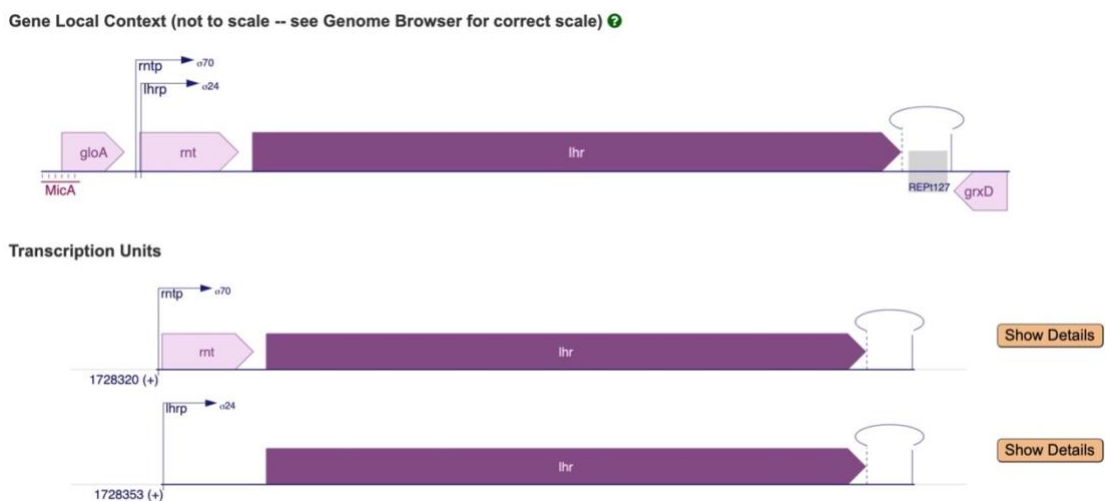


Figure 3.6 Lhr expression may be controlled by a temperature sensitive promoter.

Promoter information gathered from 'biocyc.org'¹⁸⁶.

The second promoter is located within the *rnt* gene starting after the second codon, 734 base pairs upstream from *lhr*'s start codon. This promoter is bound by RpoE, a σ^{24} sigma factor which specialises in responding to stresses such as heat shock^{200,201}, membrane damage²⁰² and accumulation of unassembled or misfolded proteins²⁰³. RpoE also controls expression of proteins which induce the mutagenic repair of DNA breaks²⁰⁴ as well as proteins associated with oxidative stress^{205,206}. Lhr's potential expression control by RpoE will be discussed further in section 6.5 of the discussion.

Regulation under RpoE gives strong evidence that Lhr may play a hand in a wide range of cellular pathways as well as in DNA repair. Due to the size of Lhr, regulation under this promoter may limit expression to only when additional chaperone proteins are present to ensure proper folding.

Lhr Rare Codon Analysis		
Number of bases in Sequence = 4617		
Number of Codons in Sequence = 1539		
Amino Acid	Rare Codon	Frequency of Occurrence
Arginine	CGA	15
	CGG	27
	AGG	4
	AGA	6
Glycine	GGA	15
	GGG	19
Isoleucine	AUA	8
Leucine	CUA	9
Proline	CCC	19
Threonine	ACG	25
Total %		9.55

Figure 3.7 Analysis of *E. coli* *lhr* codon usage.

Table generated from²⁰⁷ and redrawn.

Initial pilot overexpressions of Lhr yielded limited full length product but abundant overexpression of a protein of ~35 kDa in size. The first hypothesis was that this

protein was due to degradation of full length Lhr into small fragments, although we postulated that for this to be true, there should also be an abundance of larger sized degradation products. Overexpression optimisation continued by changing growth temperature, cell strains and concentration of transcription activators (i.e. IPTG and L-arabinose).

Successful overexpression occurred using Rosetta 2 *E. coli* cells which harbour the pRARE2 plasmid encoding 7 rare codon tRNAs. Improved expression in the presence of pRARE2 seemed counterintuitive so we investigated *lhr*'s codon usage. We analysed other *E. coli* proteins alongside (data not shown) and *lhr* contained an unusually high presence of rare codons by more than 3%. Analysis of rare codon position highlighted a cluster of codons which may cause extended replicative stalling, polymerase slippage and transcription of a pre-mature STOP codon (Figure 3.8). Little study has been completed into why genes would possess rare codons but in the case for Lhr we may infer that it is for protein regulation. This can be justified when considering the two promoters present within the 'rnt-lhr' operon. Having regulation by premature termination of the *lhr* gene without disrupting *rnt* expression may pose less strain on the cell. Regulated expression of this operon may also be more complex with both proteins being produced in response to multiple different stimuli. As *rnt* and *lhr* will share the same mRNA, the 35 kDa portion of Lhr may also serve a functional purpose to RNaseT.

ATG GCA GAT AAT CCA GAC CCT TCA TCG CTC CTG CCG GAC GTG TTT TCA CCG
 GCG ACC CGC GAC TGG TTT CTT CGC GCC TTT AAA CAG CCG ACC GCT GTC CAG
 CCG CAA ACC TGG CAT GTG GCG GCG CGA AGC GAA CAT GCG CTG GTG ATT GCA
 CCG ACC GGC TCC GGG AAA ACG CTG GCA GCA TTT CTC TAC GCC CTC GAT CGG
 CTC TTC CGC GAA GGC GGC GAA GAT ACC CGC GAG GCG CAT AAG CGT AAA ACC
 TCA CGC ATC CTC TAT ATT TCA CCG ATA AAA GCC CTG GGC ACC GAC GTT CAG
 CGC AAC TTG CAG ATC CCG TTG AAG GGT ATT GCC GAT GAA CGG CGG CGG CGC
 GGC GAA ACG GAA GTC AAT CTT CGC GTA GGG ATC CGT ACT GGC GAT ACG CCT
 GCA CAG GAA CGC AGC AAA CTC ACC CGT AAT CCG CCG GAT ATT CTG ATC ACC
 ACA CCC GAA TCA CTC TAT CTG ATG CTG ACC TCC CGC GCG CGC GAA ACG CTA
 CGC GGC GTC GAA ACG GTA ATT ATT GAT GAA GTC CAC GCG GTA GCG GGC AGT
 AAA CGT GGT GCG CAT CTG GCG TTA AGT CTG GAG CGG CTC GAT GCG CTG CTC
 CAC ACC TCA GCA CAG CGA ATT GGC CTT TCT GCC ACT GTG CGC TCA GCC AGC
 GAT GTG GCA GCA TTT CTT GGT GGC GAT CGC CCG GTT ACG GTA GTC AAC CCG
 CCC GCA ATG CGC CAT CCG CAG ATA CGA ATT GTC GTA CCG GTC GCC AAT ATG
 GAT GAT GTC TCA TCG GTC GCC AGC GGC ACC GGC GAA GAC AGC CAT GCC GGC
 CGG GAA GGC TCC ATC TGG CCA TAT ATT GAA ACG GGT ATC CTT GAT GAA GTG
 TTG CGC CAT CGC TCG ACC ATT GTC TTT ACT AAT TCG CGG GGG CTG GCG GAA
 AAA CTG ACG GCA CGA TTA AAT GAG CTT TAC GCC GCA CGC TTA CAG CGT TCC
 CCG TCT ATC GCC GTT GAT GCG GCC CAT TTC GAG TCG ACC TCC GGC GCA ACC
 TCT AAC CGT GTA CAA AGT AGC GAC GTT TTT ATT GCC CGC TCA CAC CAC GGC
 TCC GTC TCT AAA GAA CAA CGA GCA ATC ACC GAA CAG GCG CTG AAA TCG GGT

Figure 3.8 Identification of *lhr* rare codons within the first 313 codons.

Analysis of rare codon usage in *E. coli lhr* reveals high abundance and potential site of transcription stalling and premature termination. Codons CGG (4.1 bases per 100), GGG (8.6), ACG (11.5) and CGA (4.3) occur in close proximity to a TTA AAT codon pair which may be transcribed as TAA (cyan, *E. coli* most common STOP codon)²⁰⁸. Analysis results redrawn from²⁰⁷.

3.2.4 Identification of Lhr abundance in bacteria

Recent studies have highlighted the vast abundance of Lhr proteins across all domains of life^{110,121}. Hajj et al¹²¹ extensively investigated the lineages of 'Lhr-core' sequences, comprising of an SF2 helicase core (2x RecA-like domains), a winged-helix motif and an Lhr-specific domain 4. In Buckley et al¹¹⁰, Lhr was identified across archaea and a human structural homolog was also identified.

Although Hajj et. al.'s investigation was extensive, they omitted any C-terminal sequence owing to the variability of this domain when present. We thought to investigate ourselves into the bacterial abundance of Lhr at the phyla level, whilst playing close attention to the Blastp results to determine distribution of 'Lhr-core' and 'Lhr-extended' hits.

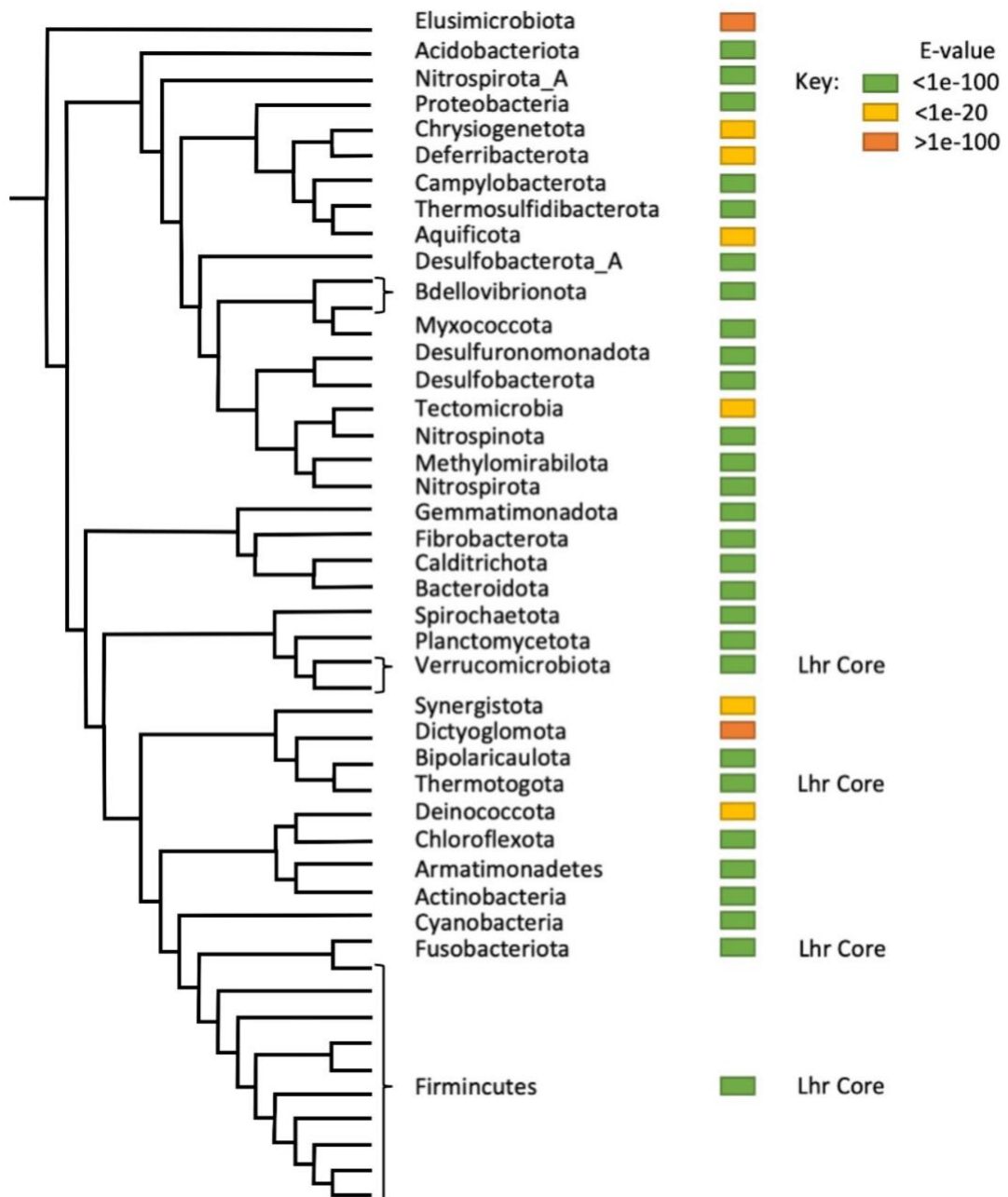


Figure 3.9 Lhr presence in bacterial phyla.

Green boxes indicate presence of Lhr, yellow represents potential presence and red indicates Lhr protein absence. Taken and adapted from²⁰⁹.

Lhr is found abundantly across bacteria and is almost entirely present in its extended form, although examples of ‘Lhr-core’ are present. Blastp²¹⁰ queries were performed using *E. coli* Lhr as the reference. Hits were graded on level of conservation using the ‘Expected (E)-values’ as displayed in the ‘description section’ of the results screen. E-

values become more statistically significant as they tend towards 0 and factor in length of query/match into the value given (small sequences have a high probability of occurring and matching than longer sequences). Presence of Lhr-extended or Lhr-core was determined when looking at the 'alignment' section of the results screen, only the top hit was considered. Results are intended as a snapshot and Lhr variation may be present between species within the same phyla.

For ease of viewing, results are colour coded with **green** representing very good matches, **yellow** depicting potential presence of Lhr and **red** showing Lhr absence. Figure 3.9 was taken and adapted from Gavriilidou et al²⁰⁹ showing a simple relationship between bacteria phyla. Figure 3.9 is slightly misleading due to the absence of some bacterial phyla where Lhr is not present. Full results and a phylogenetic tree of Lhr-extended proteins can be seen in section 8.2 of the appendix.

Using data gathered here and data presented in Hajj *et al.*, it is clear to see that Lhr is abundantly conserved across all domains of life¹²¹. Lhr is a relatively poorly studied protein considering its level of conservation so extensive characterisation is needed to determine its place in the DNA repair architecture. Additional study may highlight Lhr protein specialisation between species and whether defined roles are dictated upon the presence of a C-terminal extension.

3.3 Summary of Key findings

3.3.1 Phylogenetic analysis of RecA/Rad51 family proteins

As identified through MUSCLE sequence alignment, RecA/Rad51 family proteins display a high level of conservation within the Walker A and Walker B active sites (Figure 3.1). Conservation is situated with the P-loop, which contacts the γ -phosphate of ATP presented here as GX₄GK(S/T), and an aspartic acid within the Walker B hydrophobic patch. Level of conservation between these proteins may give an indication of functional properties. This is evident with *E. coli* RadA (Sms) which displays limited ATP hydrolysis as a monomer which may be due to a glycine residue being present instead of the Walker B catalytic aspartate.

Further conservation was identified throughout the proteins structure which are theorised to aid in formation of the ATP binding pocket and ATP coordination as displayed in Figure 3.2. A residue of great importance to ATP hydrolysis also showed conservation. Displayed in *H. sapiens* Rad51 in Figure 3.2 C, R299 represents the 'arginine finger' residue which extends into an adjacent Rad51 molecules to catalyse ATP hydrolysis when as part of a oligomer. Residues which aid in R299 positioning were also highly conserved.

Investigation into *T. brucei* RecA/Rad51 proteins as a model for a Rad51 paralogue system through phylogenetics suggested limited suitability. With reference to Figure 3.3, *T. brucei* Rad51 paralogues showed great phylogenetic similarity to *H. sapiens* Rad51 and DMC1 recombinase proteins, *E. coli* RecA and human Rad51 C/XRCC3 Rad51 paralogue proteins. A high phylogenetic match to a bacterial protein seemed

counterintuitive but this may be justified when considering the complex cell cycle of *T. brucei* displaying both eukaryotic and prokaryotic characteristics.

Subsequent phylogenetic analysis highlighted grouping of proteins into three potential subfamilies. RecA/Rad51 family protein separation displayed as 'recombinogenic-like', 'XRCC3-like' and 'Rad51 C-like' (Figure 3.4). Classifying uncharacterised proteins into these potential subfamilies may allow guided study across all domains of life and allow direct comparison of functionally homologous proteins.

3.3.2 Investigation into Lhr family protein genomic context and bacterial distribution

Lhr family proteins show conservation of genomic context in the organisms displayed in Figure 3.5. Numerous adjacent proteins are involved in DNA repair or DNA modification. Considering work produced by Hajj *et al.*, genomic context is not conserved between all instances of Lhr in bacteria and archaea. Here we suggest that genomic context may allow the linking of functionally similar Lhr family proteins allowing informed characterisation.

Analysis of *E. coli* Lhr gene regulation identified possible σ^{70} and σ^{24} promoter regulatory elements. These suggest expression during exponential growth or in situations of heat shock, respectively. Expression control in circumstances of heat shock, where additional chaperones are also expressed, may suggest the need for aid in proper protein folding. Codon usage investigation displayed an usually high % of

rare codons for an *E. coli* protein. This may allow further expression regulation through pre-mature transcription termination.

Lhr's distribution between bacterial clades showed high abundance in bLhr-HTH forms with 'Lhr-core' proteins also present when an extended Lhr was not present. Relative abundance in archaea and bacteria suggest Lhr family proteins are of high importance but biological function is yet to be fully appreciated.

Chapter 4 : ‘Mechanistic insights into Lhr helicase function in DNA repair’

4.1 Introduction to Lhr-core

Lhr is Superfamily 2 helicase which is able to translocate and unwind multiple nucleic acid substrates in an ATP dependent manner^{119,120,134,136}. It is highly conserved throughout archaea and is present in the extremely reduced Nanoarchaeota and all classes of the Asgardarchaeota^{110,116,121}. Lhr is present in two forms with ‘Lhr-core’, comprising of just the helicase domains and ‘Lhr-extended’, which includes a large C-terminal domain of unknown function. Genetic phenotypes for Lhr-extended and Lhr-core knockouts have remained relatively elusive, with previous studies in *Haloferax volcanii* showing little effect in cell survival in response to UV or γ irradiation²¹¹, and only a modest UV sensitivity in *Sulfolobus acidocaldarius*¹³². In response to genetic insult by mitomycin C, *lhr* from *M. tuberculosis* showed a 4-fold up regulation in transcription¹³⁰. Genetic analysis in *E. coli* cells treated when AZT showed a synergistic phenotype when *lhr* was knocked out alongside RadA (Sms), a Rad51/RecA family protein¹.

Lhr-core comprises only of the helicase domains, fielding similar domain structure and orientation to the archaeal Hel308, a Ski2-like DNA helicase^{110,133,212}. Thus far, biochemical analysis has been limited to just five organisms. Lhr has been shown to unwind a plethora of RNA and DNA substrates by translocation along single-stranded nucleic acid with a 3' to 5' directionality^{132,135}. Lhr from *M. thermotrophicus*

belongs to the aLhr-2 subfamily of Lhr-core proteins and thus far remains uncharacterised.

4.2 Genetic analysis of Lhr-core

4.2.1 Archaeal Lhr localises at stalled replication forks

M. thermautotrophicus Lhr-core's recruitment to stalled replication forks was identified using an *E. coli* reporter strain which carries a *dnaE486* mutation on the replicative DNA polymerase III and are deficient in RecQ, a protein which aids in stalled replication fork recovery and elicits a wider DNA repair SOS signal. The *dnaE* gene encodes for the α -subunit of DNA pol III and is responsible for the polymerising activity. The *dnaE486* mutation causes cells to become temperature sensitive. Cells grown at 30°C are permissive and are fully viable. Cells grown at 37°C are semi-permissive, slow growing and highly filamentous, a phenotype which is partially suppressed when *recQ* is knocked out. Cells grown at 42°C are inviable due to the loss of DNA Pol III activity⁹³.

At the semi-permissive growth temperature, DNA replication is perturbed and prone to stalling, mimicking inhibition by DNA damage. Nominally, replicative stress is recovered by activation of replication-coupled DNA repair pathways however, interference may occur when expressing heterologous proteins which localise to stalled replication forks. This causes a reduction in cell viability due to the impedance of additional non-native repair proteins.

Growing these cells and expressing heterologous suspect repair proteins at the three temperatures allows identification of potential replication-coupled repair proteins. Unaffected growth at 30°C confirms suspect protein is non-toxic to normal replication and no growth at 42°C ensures no suppressor mutations have arisen to give false positive results at 37°C.

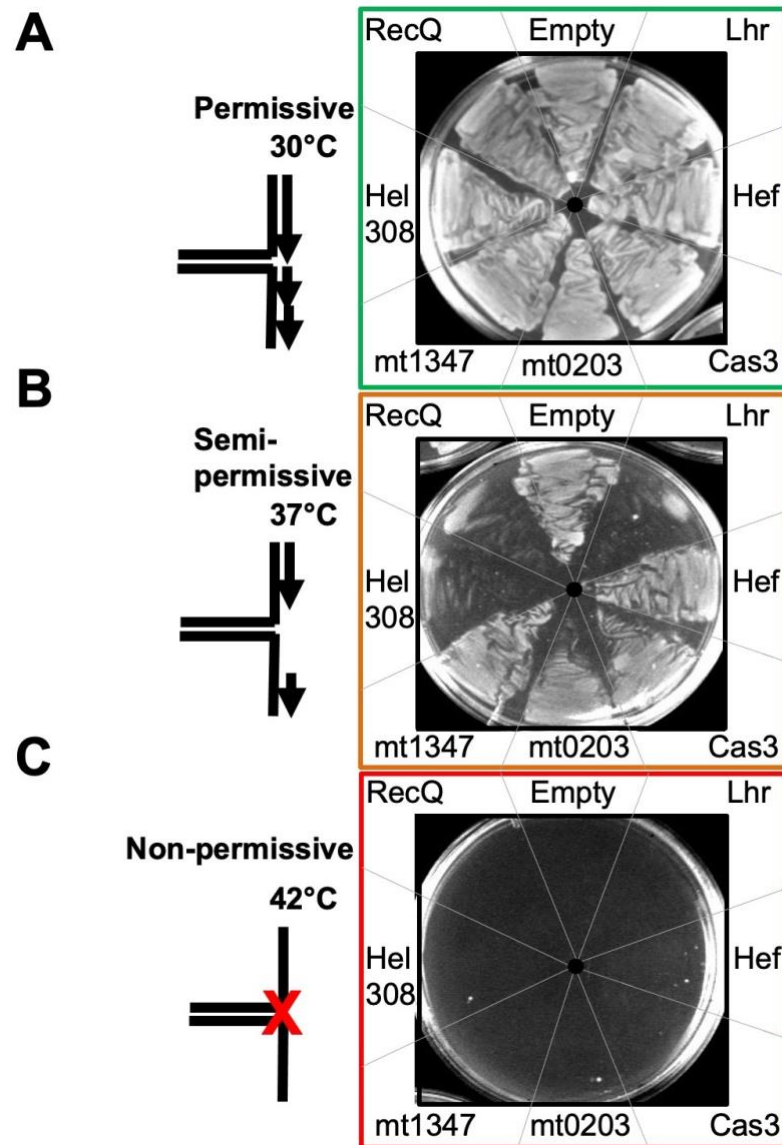


Figure 4.1 *M. thermautotrophicus* Lhr interacts with stalled replication forks in *E. coli* *dnaE486 ΔrecQ* cells.

Transformed cells containing suspected DNA repair helicases were grown at 30°C to OD₆₀₀ 0.5. 100 µl of cells were spread onto agar plates containing appropriate antibiotic and grown overnight at, (A) 30°C for permissive growth. Replication occurs unhindered and cells are fully viable. (B) 37°C for semi-permissive growth. Replication is destabilised by *dnaE486* allele and prone to stalling. (C) 42°C for inviable growth. Replication cannot fully complete due to the *dnaE486* allele, colonies present here represent survival through suppressor mutations. ‘Empty’ represents pT7-7 empty vector control, ‘*EcoRecQ*’ and ‘*MthHel308*’ give inviability phenotypes suggesting

interaction with stalled replication forks as did 'Lhr'. Taken from Buckley et al. (2020)¹¹⁰.

Results presented in Figure 4.1 echo previously published phenotypes of *E. coli dnaE* Δ *recQ* cells transformed with *EcoRecQ* and *MthHel308* repair helicases presenting as a dominant negative effect on growth at 37°C. These proteins have been attributed to localise at stalled replication forks and aid in fork recovery^{93,107}.

M. thermautotrophicus Lhr-core expression at 37°C also results in a negative growth phenotype suggesting a similar localisation/recruitment to stalled replication forks (Figure 4.1 **B**). Expressing Lhr-core at the permissive temperature (30°C, Figure 4.1 **A**) had no effect on cell viability. This confirms Lhr-core is non-toxic to *E. coli* and cells are replicating normally. No cell growth occurred at 42°C (Figure 4.1 **C**).

Lhr-core's interaction with forked DNA substrates was investigated further to support or debunk this observed phenotype.

4.3 Biochemical analysis of Lhr-core

DNA unwinding assays were used to determine Lhr-core's polarity and substrate preference. Reactions were loaded onto 10% TBE acrylamide gels allowing separation of DNA species by molecular weight. DNA products were visualised using a Storm™ scanner (Amersham) for phosphorimaging screens, after drying the gels under a vacuum on a flatbed gel dryer. Quantification was achieved from TIF files using the GelEval software. Boiled samples were used to show full dissociation of the DNA substrate. *M. thermautotrophicus* Lhr-core protein was previously purified during my MRes, purification protocols can be viewed in Buckley et al (2020) and Buckley MRes thesis (2017)^{110,156}.

4.3.1 Identification of Lhr-core polarity and optimal ATP:Mg²⁺ ratio

With reference to Figure 4.2, archaeal Lhr-core preferentially translocates in a 3' to 5' direction in an ATP dependent manner and is able to displace a complementary DNA strand as it does. This is shown by the accumulation of the 32 nucleotide strand (Figure 4.2 A, lane 2). Lhr-core unwinds partial duplex DNA with the highest efficiency when using ATP and magnesium in a 2:1 ratio (Figure 4.2 B).

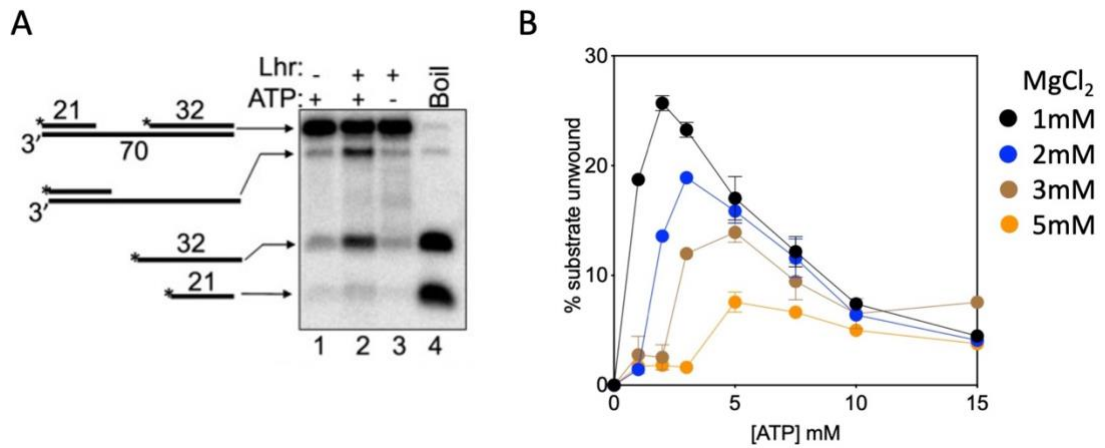


Figure 4.2 *M. thermotrophicus* Lhr-core preferentially translocates and unwinds branched DNA substrates with a 3' to 5' polarity.

DNA products were separated and visualised using 10% TBE acrylamide gels. Substrates were 5'-³²P end labelled as indicated (*) and used at 10nM. (A) Lhr-core (100 nM) displaced the 32 nucleotide strand from a gapped duplex (1 nM) in an ATP-dependent manner indicating 3' to 5' directionality. (B) Lhr-core processivity in buffer containing an increasing concentration of ATP and variable concentrations of magnesium on the gapped duplex DNA substrate used in A. Optimal ratio 2:1 ATP to MgCl₂. Reactions performed in triplicate and standard error shown. Taken and adapted from Buckley et al. (2020)¹¹⁰.

A helicases directional polarity gives clues to protein functionality through identification of preferred DNA substrates. The inability for Lhr to unwind the 21 nucleotide strand suggests the requirement for ssDNA for helicase unwinding as opposed to loading and separation from the duplex end. Determining polarity also allows determination of protein tracking and separated strands.

4.3.2 Lhr-core unwinds branched DNA substrates more readily

Lhr-core's substrate specificity was investigated further using duplex DNA (Figure 4.3 A, lanes 1-3), 5'-parital duplex (PD) and 3'-PD (Figure 4.3 A, lanes 4-6 and lanes 7-9) and a Holliday junction (HJ, Figure 4.3 A, lanes 10-12).

Lhr-core is able to minimally unwind both 5'-PD and 3'-PD duplex DNA in an ATP independent manner. This may be due to Lhr-core's ability to distort DNA base-pairing causing low levels of DNA unwinding (see Figure 4.6). A higher processivity of a 3'-PD is observed upon the addition of ATP. Comparing Figure 4.3 A lanes 5 and 8, Lhr-core is able to more readily unwind the 3'-PD substrate when in the presence of ATP. This data supports a 3' to 5' polarity as identified in Figure 4.2 A. Lhr-core is able to fully unwind a Holliday junction (HJ) branched substrate to ssDNA in the presence of ATP (Figure 4.3 A, lane 11). No minimal unwinding is observed in the absence of ATP as seen with 5'-PD and 3'-PD. This may be due to the higher stability afforded by a HJ substrate due to the increased number of DNA base pairings. Lhr-core's ability to unwind all three substrates was quantified further with respect to protein concentration (Figure 4.3 B) and as a function of time (Figure 4.3 C). In both cases, Lhr was able to produce 40% more unwound product with a HJ substrate as compared to a 3'-PD, further supporting a preference to branched DNA substrates.

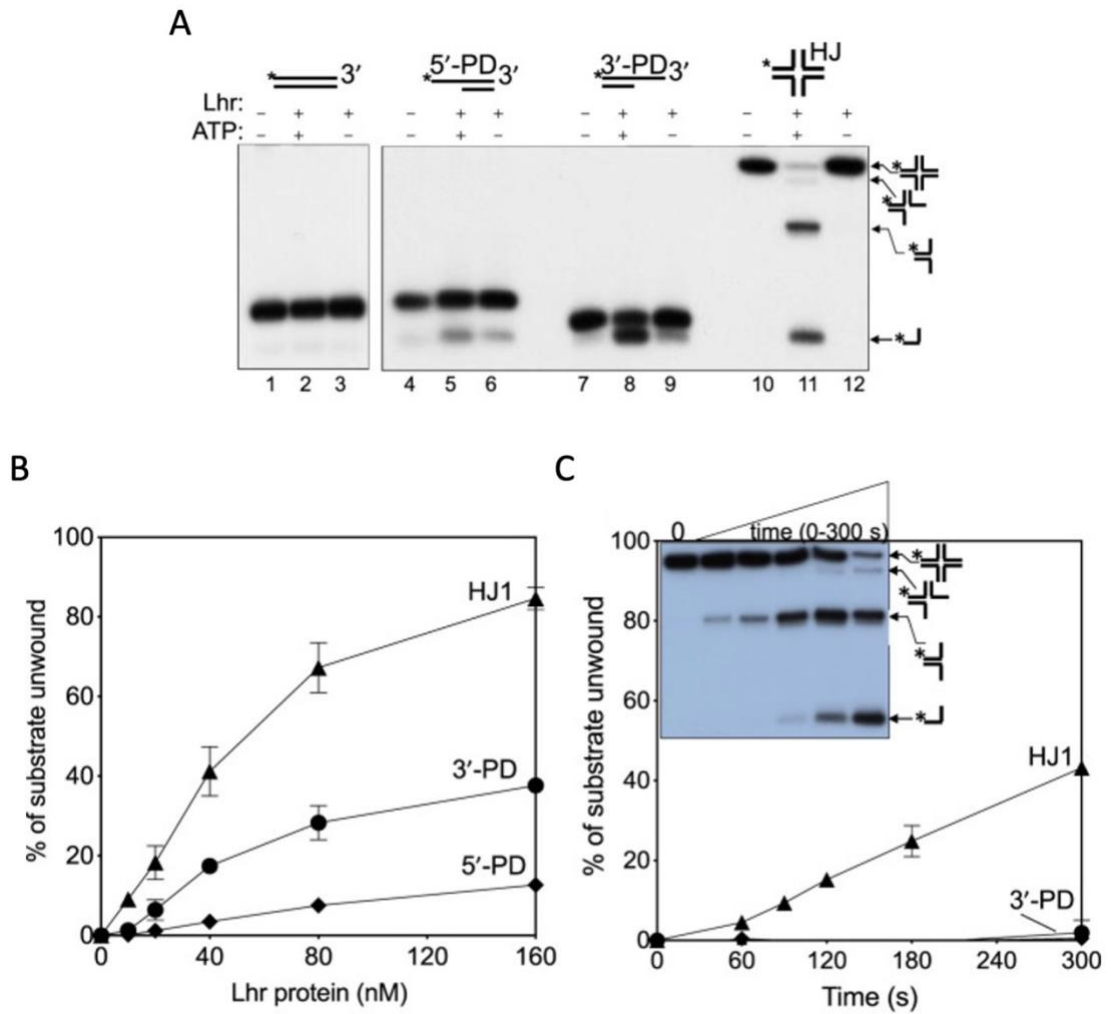


Figure 4.3 Lhr-core readily unwinds branched DNA substrates.

5'-³²P end labelled (*) products were separated and visualised using 10% TBE acrylamide gels. (A) Lhr-core (100 nM) activity on 1 nM of duplex DNA (lanes 1-3), 5'-partial duplex (PD) and 3'-PD (lanes 4-6 and 7-9 respectively), and Holliday junction (HJ) (lanes 10-12) DNA substrates. Data here is presented as with varying Lhr-core protein concentration (10-160 nM) in B and as a function of time in C, with 40 nM protein and 10 nM DNA. Reactions performed in triplicate and standard error shown. Taken and adapted from Buckley et al. (2020)¹¹⁰.

4.3.3 Lhr-core preferentially unwinds fully base-paired forked DNA substrates

Unwinding assays were performed to investigate Lhr-core's function on HJ substrates. Comparative analysis of HJ unwinding intermediates against the bona fide HJ resolvase enzyme RuvAB highlight a clear difference in processivity (Figure 4.4 A). Lhr-core causes complete unwinding of the HJ substrate suggesting non-specific targeting of forked DNA substrates. During HR, RuvAB migrates the HJ junction seeking preferred sites for incision by RuvC and resolution. HJ branch migration by RuvAB during this assay results in the accumulation of a flayed duplex. This is due to the movement of the branch point through one end of the HJ substrate causing subsequent separation. Lhr's added activity causing full unwinding suggests a more generalised localisation to branched DNA substrates.

To investigate substrate preference further, Lhr-core's unwinding capacity was compared between multiple forked and HJ substrates against a function of time (Figure 4.4 B). Lhr-core was able to unwind a flayed duplex (forked DNA lacking leading and lagging strands) to the same extent as both a mobile (HJ1) and immobile (HJ2) HJ substrate. A higher level of unwinding was seen with both fully base-paired forked DNA substrates. These substrates (Fork-1 and Fork-2) are fully base paired flayed duplex equivalents, with Fork 1 containing a breathable junction and Fork 2 a 'static' equivalent. This investigates Lhr's ability to cause local unwinding of the static fork for loading and helicase activity.

Figure 4.4 B supports genetic data shown in Figure 4.1 and suggests Lhr-core is localising to stalled replication forks as opposed to interfering with repair intermediates such as that generated in HJ resolution.

Previous work on Lhr-core showed an unwinding preference to substrates containing RNA^{120,134}. Figure 4.4 C shows *M. thermautotrophicus* Lhr-core preferentially unwinds a forked DNA:DNA as opposed to an equivalent DNA:RNA substrate.

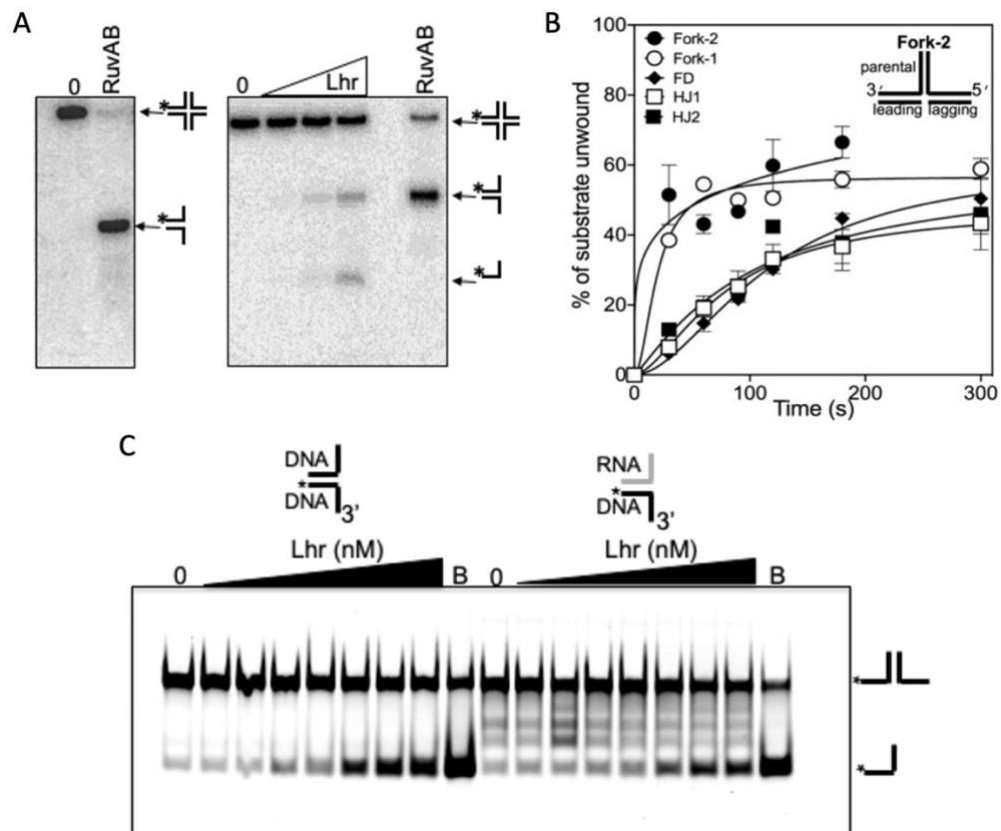


Figure 4.4 *M. thermautotrophicus* Lhr-core targets branched nucleic acids with a preference to fully base-paired DNA:DNA substrates.

5'-³²P end labelled (*) products were separated and visualised using 10% TBE acrylamide gels. (A) Comparison of Holliday junction (HJ) dissolution products of Lhr-core (5-50 nM) and RuvAB (40 nM) on 1 nM of HJ substrate in the presence of 2 mM ATP and 1 mM MgCl₂. Lhr-core unwound HJ DNA fully showing non-canonical HJ migration. (B) Lhr-core (40 nM) most effectively unwound fully based paired fork substrates when compared to a flayed duplex and HJ substrates. Reactions performed in triplicate with standard error from mean displayed. (C) Lhr-core unwinding activity on DNA:DNA and DNA:RNA flayed duplex substrates (20 nM). Protein concentrations ranged from 5 to 320 nM, 'B' indicates boiled samples.

4.3.4 Lhr-core preferentially unwinds through the parental fork DNA strands

Data presented in sections 4.3.4 and 4.3.5 were obtained in collaboration with Dr. Kevin Kramm and Dr. Dina Grohmann from the University of Regensburg.

Further DNA binding and unwinding assays were used to determine Lhr-core's action on forked DNA using substrates containing donor-acceptor dye-pairs. Dual substrate labelling of the lagging strand (ATTO 674N – **red**) and leading strand (ATTO 532 – **green**) allows differentiation of reaction products. Reactions were not subject to de-proteination to allow detection of both binding and unwinding reaction intermediates. Reactions were loaded onto 5% TBE gels to allow separation of binding and unwinding species of different molecular weight by electrophoresis.

With reference to Figure 4.5 lanes **2** and **5**, Lhr-core is able to bind a fully base paired fork and a fork substrate containing only a leading strand. Lhr-core binding to a fully base paired fork causes slight unwinding in the absence of ATP. Upon addition of ATP, Lhr-core is able to readily unwind both forked substrates (Figure 4.5 lanes **3** and **6**), preferentially unwinding through the parental fork strands in a 3' to 5' directionality. Confirmation of polarity is due to the presence of an intermediate green Lhr-core nucleic acid protein complex (as seen in lanes **3** and **6**, highlighted by large arrow) suggesting interaction and unwinding through this parental strand.

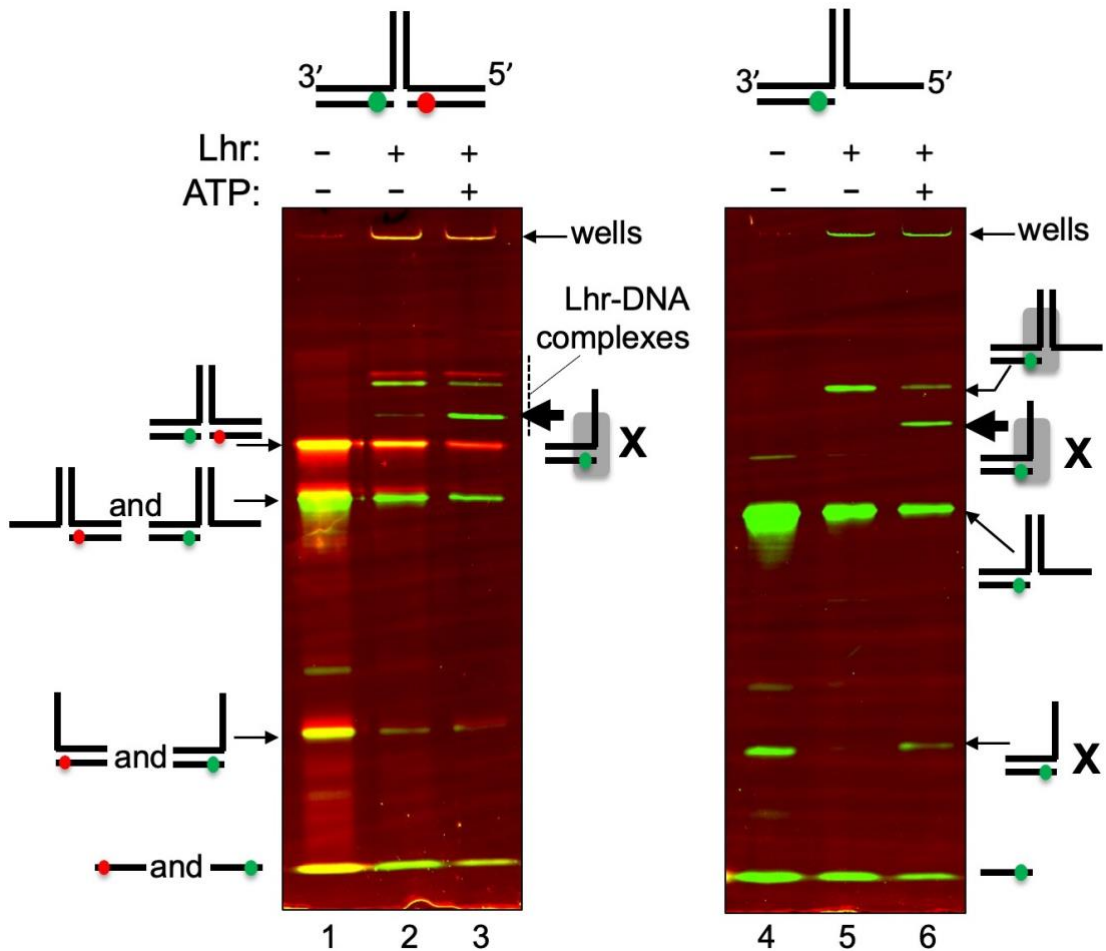


Figure 4.5 *M. thermautotrophicus* Lhr-core binding and unwinding of DNA fork substrates giving clues on unwinding directionality.

Reactions were carried out at room temperature for 20 minutes in buffer containing 1xHB, 10 nM of DNA, 2 mM ATP, 1 mM MgCl₂ and 25 mM DTT, initiated with the addition of protein to 100 nM. Fork substrate was labelled with ATTO 532 (green) and ATTO 647N (red) fluorophores as shown. Gel images represent one gel, split to allow room for annotation. DNA substrates consist of a fully base paired fork (**left**) and a partial fork lacking a lagging strand (**right**). Lanes **1** and **4** show bands representing free DNA substrates. Lanes **2** and **5** show Lhr-core binding in the absence of ATP. Lanes **3** and **6** show resulting products upon addition of ATP. Gels imaged using a ChemiDoc MP imaging system.

This data suggests a role for Lhr to unwind lesions present with the parental duplex. This may facilitate replication fork reversal, access to damaged site by additional repair proteins or may suggest an 'accessory helicase' role.

4.3.5 Lhr-core remodels fork DNA substrates in the absence of ATP and MgCl₂, prior to translocation and unwinding

Figure 4.3 shows that Lhr-core is unable to unwind duplex DNA and so requires ssDNA to be present for efficient unwinding. Using fluorescence resonance energy transfer (FRET) techniques at the single-molecule level we investigated how Lhr-core locally manipulates a fully base paired fork to facilitate loading, and subsequent unwinding. FRET assays have been used extensively in studying DNA-protein interactions by allowing observation of induced conformational changes and assembly dynamics. Such studies utilise the relationship between two dye pairs whereby an excited fluorophore is able to pass energy to the acceptor. The energy transfer efficiency (E) can be quantified and allow inference of relative distances, or changes in distance between dye pairs²¹³. A higher E value indicates fluorophores in close proximity. For this assay, both leading and lagging strand dyes are adjacent to the forked branch point. As E values increase, strands become stretched or move further away due to local unwinding.

Figure 4.6 **A** serves as a control FRET efficiency with an E value of 0.72. This represents the DNA substrate in a relaxed state as a single DNA population, in the absence of protein. Lagging and leading strand angle is at $\approx 130^\circ$. Addition of Lhr-core at room temperature (Figure 4.6 **B**, **F ii**) causes an increase in E value representing a constriction between the dye-pair. This suggests fork compaction or DNA rotation induced by Lhr-core. Figure 4.6 **C** and **F iii** shows resulting additional FRET populations upon shifting reactions to 45°C, these represent DNA stretching (E=0.50) and further compaction (E=0.92). Further incubation at 45°C (Figure 4.6 **D**) resulted in a decline in the higher E value populations (E=0.92 and E=0.78) and the emergence of lower FRET

efficiencies ($E=0.50$ and $E=0.12$) indicative of highly stretched or partly unwound DNA conformations. Addition of ATP and Mg^{2+} showed further disappearance of higher and intermediate E values and emergence of a FRET population with $E \approx 0$, representing full separation of the dye-pair due to fork unwinding (Figure 4.6 **E**, **F v**).

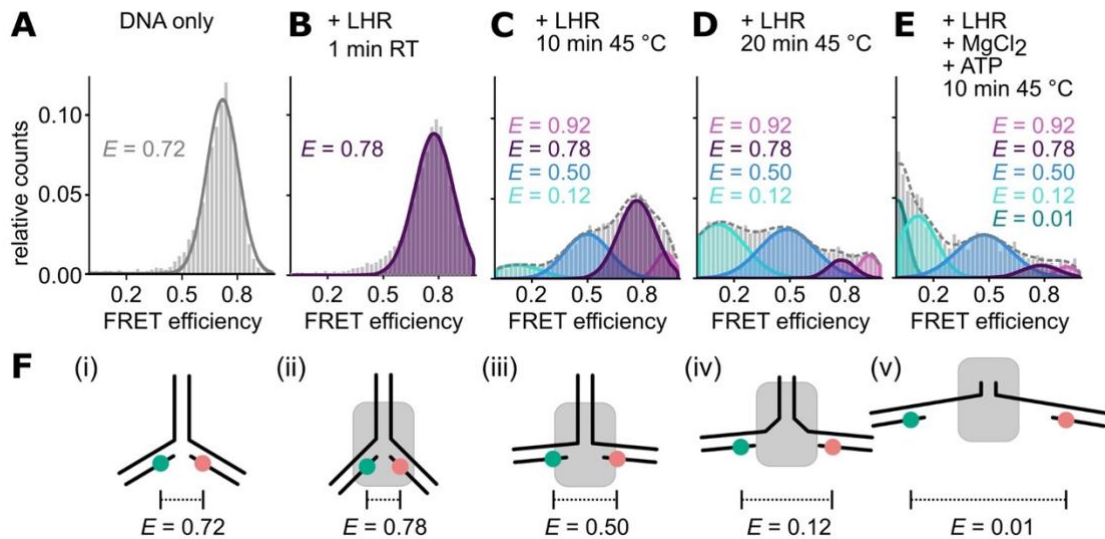


Figure 4.6 Single-molecule FRET analysis of conformational changes induced by *M. thermautotrophicus* Lhr-core on fully base paired forked DNA.

Measurements were performed on freely diffusing DNA/protein complexes using a dual labelled DNA substrate (1 nM) allowing FRET efficiency monitoring between ATTO532 (green-donor) and ATTO647N (red-acceptor) for inter-fluorophore distance calculations. FRET populations were fitted with multiple Gaussian distributions. Mean FRET efficiency is shown. **(A)** FRET efficiency in absence of protein. **(B)** Effect of addition of 1 μ M Lhr at room temperature (RT). Reactions were moved to 45°C and monitored after incubation for 10 minutes **(C)** and 20 minutes **(D)**. **(E)** Reactions were incubated at 45°C for 10 minutes with the addition of 1 mM MgCl₂ and 2 mM ATP. **(F)** Putative model for the mechanism of Lhr-dependent fork DNA unwinding. **(i)** Relaxed DNA in the absence of protein with leading (**green**, donor) and lagging (**red**, acceptor) strands labelled. **(ii)** Lhr binding (**grey**) causing compaction of fork. **(iii)** Fork stretching upon heat activation of Lhr. **(iv)** Lhr partially melts fork upon addition of ATP and Mg²⁺. **(v)** Mostly unwound fork with Lhr bound and unwinding through parental DNA strands. Reactions performed in triplicate. Taken from Buckley et al. (2020)¹¹⁰.

4.3.6 Lhr novel C-terminal region matches to a glycosylase repair protein

Data presented in section 4.3.6 was obtained in collaboration with Dr. Christopher Cooper from the University of Huddersfield.

M. thermautotrophicus 1802 (Lhr-core used in this work) was superimposed onto *Mycobacterium smegmatis* Lhr-core crystal structure (PDB: 5V9X) using PyMOL. *Mth1802* matches well with the mycobacterium Lhr-core crystal structure with a root-mean-square deviation (RMSD) of 0.8 Å. Domain structures as shown in Figure 4.7 A, are highlighted as follows, RecA domain 1 (**green**), RecA domain 2 (**blue**), winged helix (**yellow**), domain of unknown function (**pink**), with lighter shades representing *Mth1802*. Additional structure was identified through PHYRE² *ab initio* modelling and PSIPRED searches, including a previously unresolved 30-residue α -helical structure (**red**) located adjacent to the RecA-like DNA binding domains and the translocating DNA strand. This helical structure may facilitate additional contacts with forked DNA structures but further experimentation is needed.

Mycobacterium Lhr's previously unidentified extended C-terminus was modelled against PHYRE² and DALI servers. The most proximal region (residues 1139 to 1507) strongly matched protein folds present in the DNA glycosylase enzyme AlkZ with an RMSD of 1.6 Å⁵⁶. An intermediate region between Lhr-core and the C-terminus (residues 938 to 1077) showed strong similarity with the tandem winged helix domains present in the elongation factor SelB (RMSD 6.9 Å). A cartoon summary of this can be seen in Figure 4.7 B and structural predictions displayed in C.

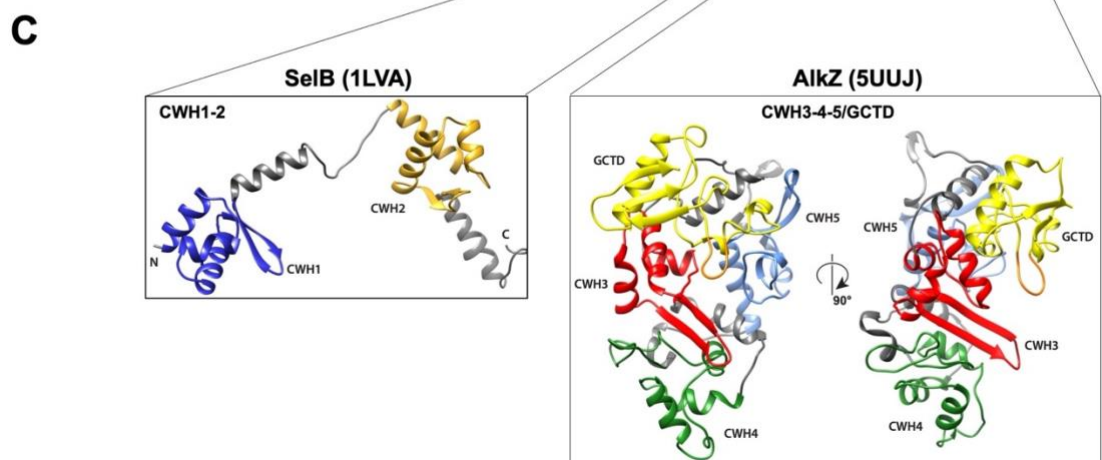
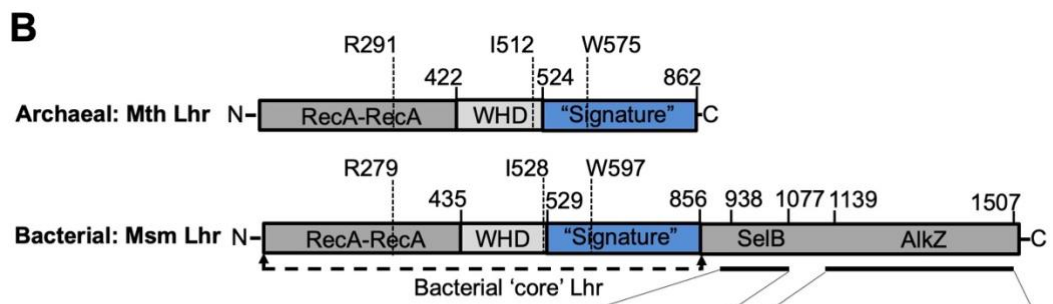
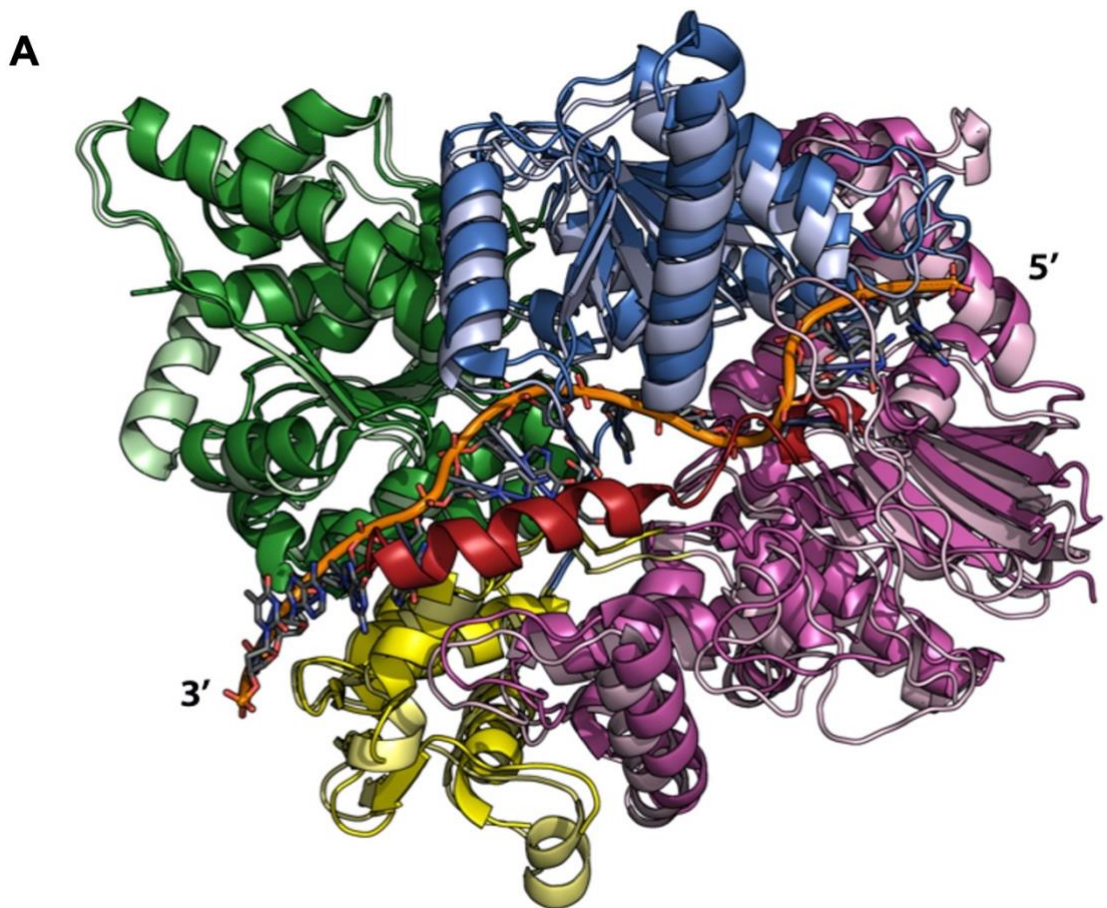


Figure 4.7 Bioinformatic analysis of Lhr-core and extended C-terminus. (A) *Mth1802* Lhr was superimposed onto *M. smegmatis* Lhr-core crystal structure (PDB: 5V9X). Protein orientation and domain colours consistent with¹³⁴. Domains labelled as follows, RecA domain 1 (**green**), RecA domain 2 (**blue**), winged helix (**yellow**), domain of unknown function (**pink**), with lighter shades representing *Mth1802*. ssDNA is shaded in **orange** and *ab initio* modelled *Mth1802* C-terminal 30 residues is shaded **red**. (B) Labelled cartoon summary of the domain organization of Lhr proteins from archaea (*Mth*) and bacteria (*Msm*). Lhr-core containing two RecA-like domains, winged helix domains (WHD) and 'signature' domain of unknown function are highlighted. Amino acid positions are indicated, including invariant amino acids that are required for helicase activity of the bacterial Lhr¹³⁴. C-terminal region of bacterial Lhr is also highlighted. (C) Summarizes two parts of the bacterial C-terminal Lhr region that match with structural folds of AlkZ and SelB proteins: **CWH**, C-terminal winged helix-turn-helix motif, **GCTD**, glycosylase C-terminal domain.

This data suggests bacterial Lhr may be a fusion of two proteins giving it dual functionality. Characterisation of Lhr-CTD will build a greater picture of where these proteins fit into the existing DNA repair architecture or may allow identification of a novel uncharacterised pathway.

4.4 Summary of Key findings

4.4.1 Initial identification of replication-coupled DNA repair protein

M. thermautotrophicus Lhr is able to localise to stalled replication forks in an *E. coli* *dnaE486* reporter strain, similar to a phenotype identified by Hel308 and RecQ DNA helicases. As displayed in Figure 4.1, expression of Lhr causes a loss in cellular viability due to the recombinant protein being unable to perform proper function in its *E. coli* host. Recombinant protein interference of host repair pathways adds additional strain onto the cell causing cellular death.

4.4.2 Assessment of *MthLhr* unwinding characteristics and substrate preference

Instigation into Lhr's polarity using a gapped duplex DNA substrate displayed a 3' to 5' preference in line with other characterised Lhr family proteins. This is shown through separation of the 32 nucleotide labelled strand (Figure 4.2, **A**). Optimal ATP to Mg²⁺ ratio was investigated, as displayed in Figure 4.2 **B**, showing a higher unwinding potential when present in a 2:1 ratio.

MthLhr's substrate preference was investigated further using duplex, 3'- and 5'-partial duplex and Holliday junction DNA substrates. Figure 4.3 confirmed Lhr's 3' to 5' polarity through little relative ability to unwind a 5'-PD. Lhr showed a preference to a HJ junction DNA substrate displaying full substrate unwinding and at least a twofold increased activity as compared to a 3'-PD.

This unwinding ability was developed with comparison to *E. coli* RuvAB, displaying differences in HJ processing. Data in Figure 4.4 suggests Lhr targeting of branch DNA

substrates as opposed to a HJ branch migration functionality. Lhr's activity on forked DNA substrates as compared to a HJ was investigated. This showed a strong preference to fully base paired 'replication fork' DNA, displaying a twofold increased rate of product formation against a function of time. Flayed duplex DNA:RNA unwinding ability showed a preference to an equivalent DNA only substrate which is in contrast to other reported Lhr family proteins suggesting a diversity in function.

4.4.3 *Mth*Lhr remodels replication fork DNA and unwinds through parental DNA strands

Lhr was shown to be able to bind and unwind a variety of replication fork intermediates shown by analysis of Lhr action on dual labelled DNA substrates. Initial Lhr action causes binding of forked substrates and unwinding through the parental duplex. This is displayed by the ATTO 532 green 'EMSA' band in lane **3** of Figure 4.5. Lhr shows limited separation of the green labelled strand further supporting a 3' to 5' directionality.

Using smFRET, Lhr was shown to be able to remodel fully base paired forked DNA substrates in the absence of Mg^{2+} and ATP (Figure 4.6, **C** and **D**). This reveals ssDNA regions for protein loading and unwinding again through the duplex upon the addition of Mg^{2+} and ATP.

4.4.4 Computational identification of an α -helical bundle and identification of AlkZ-like CTD

MthLhr showed a high level of conserved structural topology when modelled against *MsmLhr* crystal structure bound to ssDNA. An additional α -helical bundle which escaped previous structural determination was identified (Figure 4.7, red), theorised to be involved with contacts to duplex regions of forked DNA substrates.

MsmLhr's extended C-terminus was modelled through Phyre² and DALI servers displaying structural hits to SelB and AlkZ suggesting protein dual functionality.

Chapter 5 ‘The *E. coli* DNA helicase Lhr is also a DNA-uracil glycosylase’

5.1 Introduction

Lhr is a Superfamily 2 ATP-dependent 3' to 5' DNA translocase found extensively in both bacteria and archaea^{110,119,121,191}. Recent phylogenetic analysis in prokaryotes has identified the breadth of Lhr and Lhr-like proteins, showing great diversity within its own distinct clade^{121,134}. Lhr family proteins show high levels of conservation (≈30% identity) within the N-terminal 800-900 amino acids. This is known as ‘Lhr-core’ and comprises of, two RecA-like domains, a winged helix motif reminiscent of Hel308 proteins and a signature domain 4 of unknown function¹³⁴. Lhr-core is responsible for helicase activity and has been characterised extensively in both bacteria and archaea^{110,121,131,135,136}.

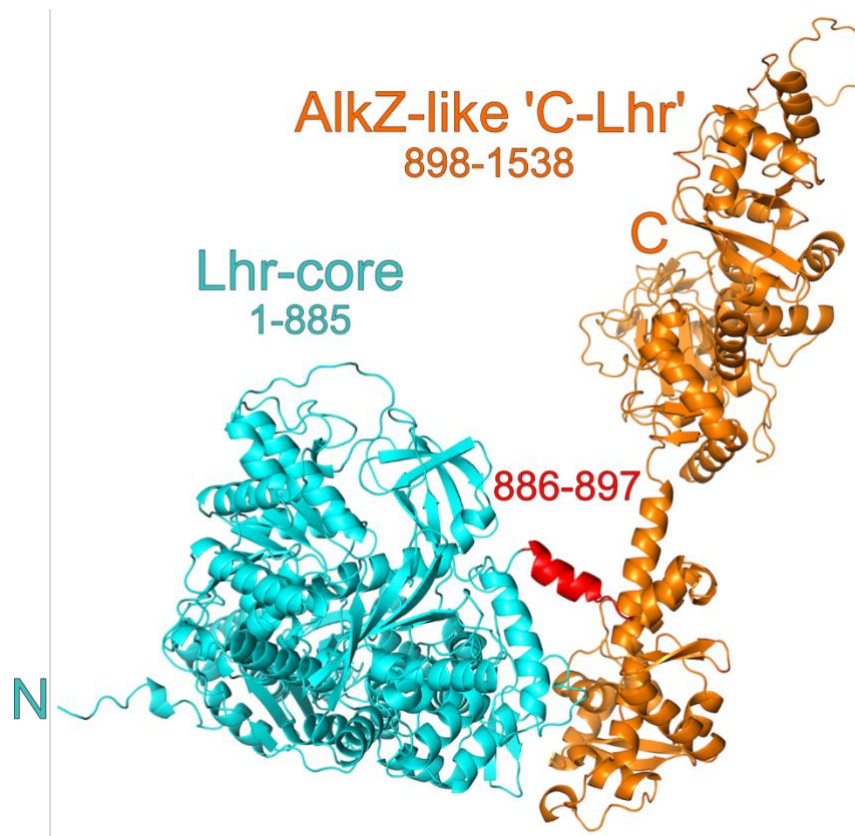


Figure 5.1 AlphaFold predicted structure of *E. coli* Lhr.

Displayed are Lhr helicase-core (**cyan**), extended AlkZ-like C-terminal domain (**orange**) termed here as 'C-Lhr' or 'CTD' and Lhr-core to 'C-Lhr' linker helix (**red**).

aLhr2 and bLhr-HTH proteins contain an extended C-terminal domain (CTD) of unknown function¹²¹. The structure of this additional 500-700 amino acids has been modelled and solved through cryo-EM displaying a domain containing extended multiple winged helix motifs and a β -barrel module reminiscent of *Streptomyces sahachiroi* AlkZ, a HTH_42 DNA glycosylase^{56,57,110,125}.

Genetic analysis of Lhr family proteins has shown variable phenotypes in response to various DNA damaging agents reminiscent of its specific preference of DNA substrates between characterised examples. The most prominent repair phenotypes display Lhr's involvement in relieving replicative stress and in the repair of alkylated DNA damage through roles in homologous recombination^{1,127,129,131,132}.

5.2 Identification of DNA repair phenotypes

Escherichia coli knockout strains were produced to allow identification of genetic repair phenotypes. Comparison of growth between wild type and knockout cell strains grown in the presence of DNA damaging agents allows inference of an associated repair pathway through damage specificities associated with each mutagenic source. Growth differences may be shown as a change in cellular viability or as a disruption in growth rate. Repair pathway impedance may also be displayed through an abnormal accumulation of random mutations in the presence of a selective pressure. Genetic analysis in this way allows guidance to *in vitro* study by giving clues to potential preferred DNA substrates.

5.2.1 Production of knockout strains for genetic analysis

E. coli MG1655 Δlhr , $\Delta lhr\Delta radA$ and $\Delta radA$ cells were generated via P1 transduction of knockout-resistant markers from corresponding 'Keio collection' donor strains. The Keio collection is a resource which comprises of *E. coli* K12 cell strains which have single-gene deletions of all non-essential genes to allow systematic analysis of genes of unknown function. A kanamycin resistance cassette replaces the targeted gene, achieved through λ Red recombineering¹³⁹, allowing selection of successful knockout cells. The kanamycin resistance marker is flanked by FLP recognition target (FRT) sites to allow cassette removal through temperature sensitive expression of FLP. The resistance cassette also contains 'gene homology arms' at each end, necessary for initial gene targeting during λ Red recombineering. Removal of the resistance cassette leaves behind a scar sequence of ≈ 100 nucleotides²¹⁴.

Figure 5.2 shows PCR verification of successful P1 transduction (**A** for Δlhr and **C**, lanes 3 and 5 for $\Delta lhr\Delta radA$ and $\Delta radA$ respectively) and removal of resistance cassette (**B**, lane 3 for Δlhr and **C**, lanes 4 and 6 for $\Delta lhr\Delta radA$, and $\Delta radA$ respectively).

Generation of Δlhr Kan^R (RB001a) displayed multiple higher bands of various sizes, with a band of ≈ 1 kb being the majority product. This band is thought to be the kanamycin resistance cassette as highlighted. The additional bands raised concern but were not thought of as an issue due to their relative disappearance after resistance cassette removal, and absence of a band of similar size to the *lhr* gene. The strain isolate represented in Figure 5.2 **A** lane 4 was selected for treatment with pCP20 and resistance cassette removal. pCP20 contains a gene encoding for *Saccharomyces cerevisiae* FLP and has a temperature sensitive origin of replication. This allows tuned

removal of the resistance cassette and curing of the plasmid in one growth step^{138-140,214}. For this work this was achieved with an overnight growth temperature of 45°C.

Generation of $\Delta lhr\Delta radA$ (RB006a) and $\Delta radA$ (RB007a) cell strains was performed in a similar fashion. With reference to Figure 5.2 C, colony PCR analysis again shows multiple bands present after P1 transduction. For these cell strains the resistance cassette product is ≈ 1.2 kb. This is due to the strain verification primers binding 100 base pairs either side of the *radA* gene adding an additional 200 base pairs to the resultant band. Additional bands are removed following pCP20 treatment.

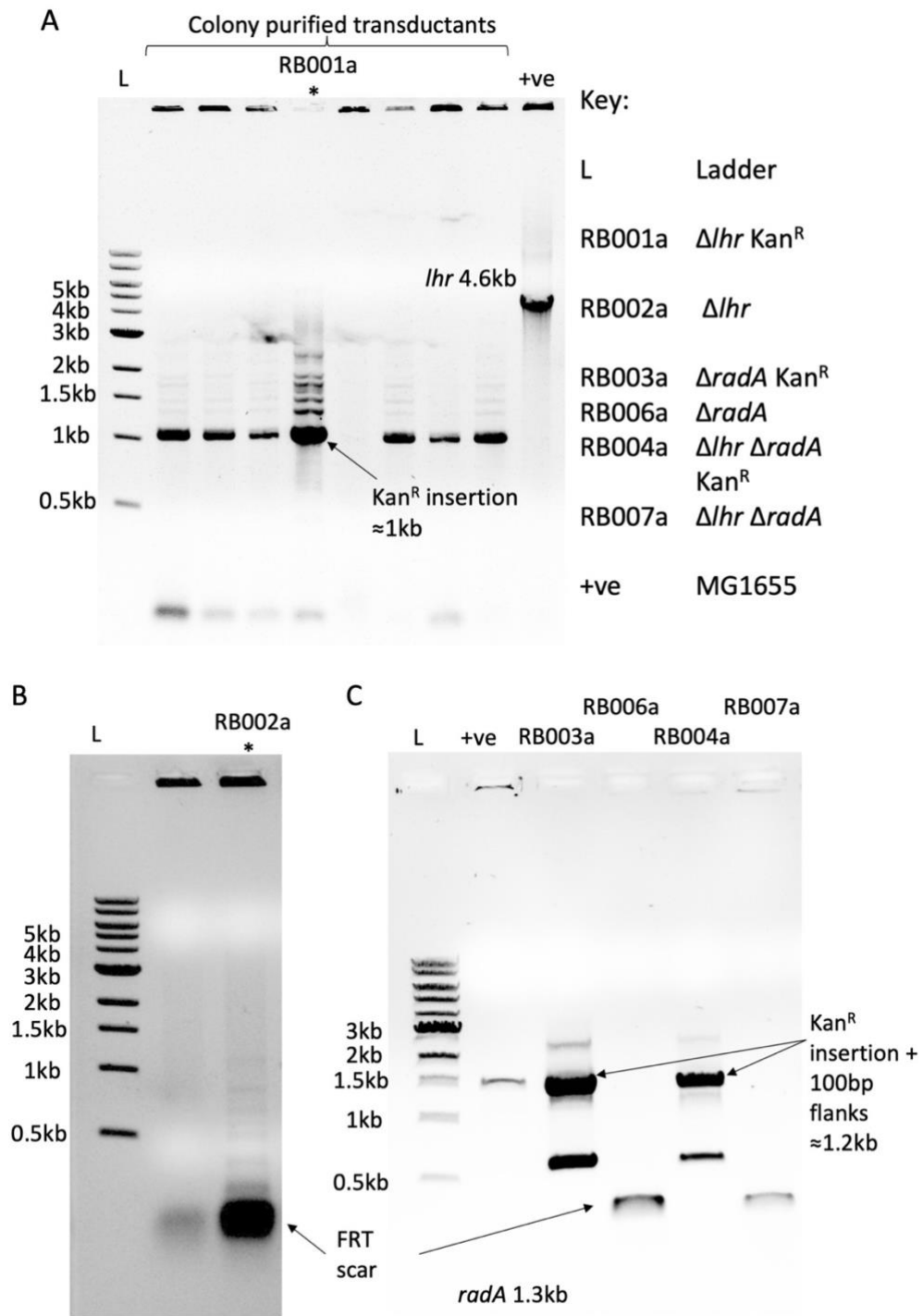


Figure 5.2 Generation of knockout strains for use in genetic analysis.

Colony PCR samples were loaded onto 0.5% agarose TBE gels stained with 0.2 μ g/ml ethidium bromide and run at 100 V until sufficient migration. KO cell generation achieved through P1 transduction from 'Kieo collection' donor and pCP20 treatment for resistance marker removal. (A) and (B) display Δlhr generation and (C), for $\Delta lhr \Delta radA$ and $\Delta radA$ cells.

A colony PCR of the donor Keio collection cell strains may allow inference of the origin of the additional bands shown in Figure 5.2 **A**, lane 4 and Figure 5.2 **C**, lanes 3 and 5. This would show if these bands are present in the original strain samples or have occurred through the P1 transduction process, serving as an added control along with data presented below in section 5.2.2.

5.2.2 Lhr may be involved in replication-associated DNA break repair

Previous study of *Ecolhr* identified a growth sensitivity to azidothymidine (AZT) when *lhr* was knocked-out alongside *radA*¹. To confirm our cell strains behave in a similar fashion, we sought to repeat this phenotype through viability spot tests.

AZT is an antiretroviral drug used for the treatment of human immunodeficiency virus-1 (HIV-1). AZT is an analog of thymidine and elicits inhibitory function by blocking subsequent nucleoside addition after AZT incorporation. For HIV-1, this acts to block reverse transcription of the viral RNA upon cellular infection^{215,216}. In *E. coli*, AZT causes replication arrest through chain termination, leading to the occurrence of single-stranded and double-stranded DNA gaps. For repair, a plethora of DNA repair proteins are deployed such as RecF (ssDNA gaps), RecBCD complex (dsDNA breaks) and RuvAB/RuvC (in late stage recombination)²¹⁵. Replicative stressed caused by AZT is also alleviated through template switching, which often leads to an accumulation of mutations at sites of short palindromic repeats¹²⁶.

With reference to Figure 5.3, all *E. coli* knockout cells used within this experiment show an increased sensitivity to AZT, as compared to wild type MG1655. Δlhr and $\Delta radA$ single mutants show a slight increase in survival when exposed to AZT at 2.5 ng/ μ l. This observation is also seen in *lexA3-Ind*⁻ cells¹ although in both experiments, increased survival may not have any significant biological relevance. Δlhr cells also showed increased survival at the very highest AZT concentration (Figure 5.3 A, furthest right panel), this may be due to suppressor mutations which allow increased survival. Due to tight budgeting, glycerol stocks of these cells were stored to allow future analysis to identify common mutations between isolates. This would be achieved

through whole genome sequencing. The $\Delta lhr\Delta radA$ cell strain (red triangle) showed the most sensitivity to growth in the presence of AZT, data of which is in close agreement to work presented by Cooper *et al.*¹.

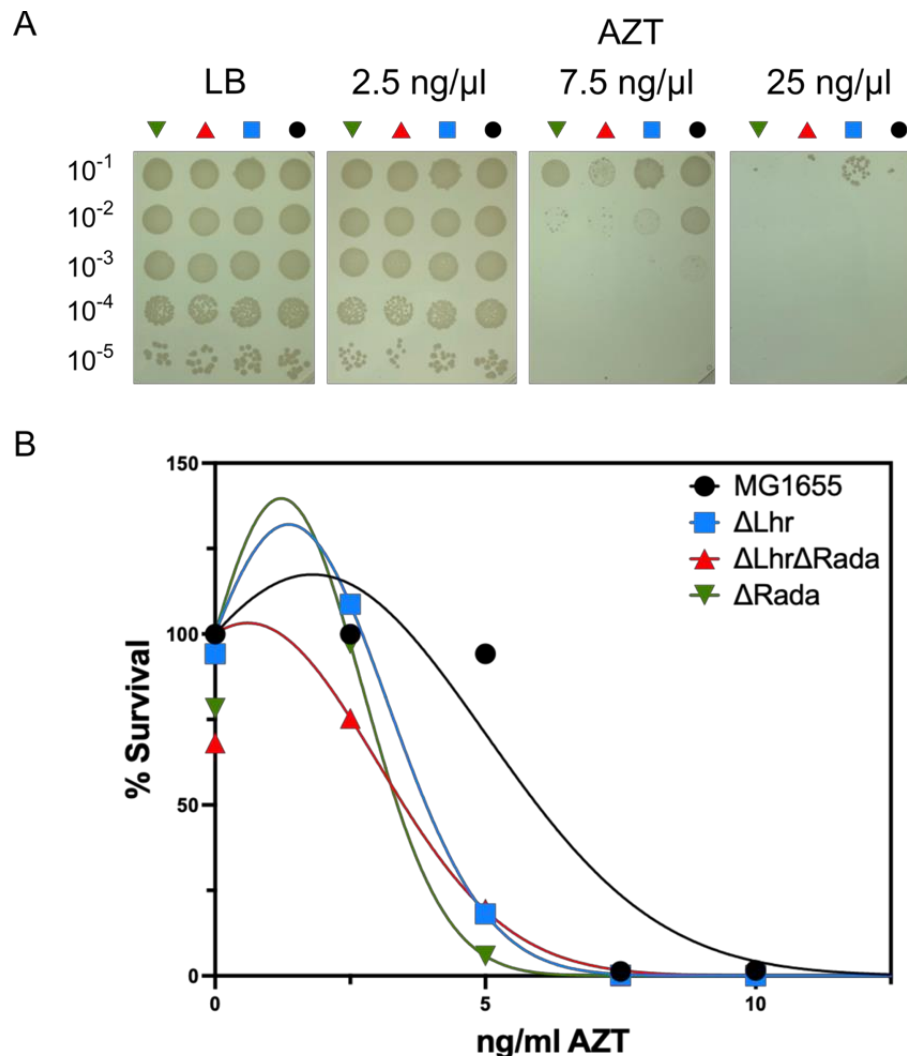


Figure 5.3 Confirmation of *E. coli* $\Delta lhr\Delta radA$ sensitivity to azidothymidine exposure highlighting potential roles in resolving DNA break associated replicative-stress.

Cellular viability of *E. coli* Δlhr , $\Delta lhr\Delta radA$ and $\Delta radA$ cells is significantly reduced when plated onto agar plates containing AZT and grown overnight. Δlhr showed an initial increase in viability in the presence of 2.5 ng/μl AZT as compared to wild type MG1655. (A) Shows a visual example of spot viability agar plates highlighting $\Delta lhr\Delta radA$ increased sensitivity. (B) Shows extended AZT concentration results (2.5 ng/μl, 5 ng/μl, 7.5 ng/μl, 10 ng/μl and 25 ng/μl), plotting linear quadratic survival. Graphs show data plotted from at least two repeats.

EcoLhr localisation at replication forks supports data presented in section 4.2. Sensitivity to AZT may occur through disruption of multiple DNA repair pathways so further analysis is needed to pinpoint *Lhr*'s involvement. The synergistic phenotype observed here is in close agreement with Cooper *et al.*¹, reaffirming *Lhr* and *RadA* (*Sms*)'s cooperativity in a potentially as yet uncharacterised DNA repair pathway.

5.2.3 Lhr is involved in a repair response following exposure to oxidative agents

After confirming our knockout strains behaved as previously published, we decided to investigate further with a diverse set of mutagenic sources.

Here, we investigate changes in cellular viability (Figure 5.4 **A**) and effects on growth rate (Figure 5.4 **B** and **C**) when cells were grown in the presence of hydrogen peroxide.

H₂O₂ is a toxic oxidant which often forms intrinsically within the cells of aerobic organisms as a by-product during normal metabolic reactions. H₂O₂ is a key member of the reactive oxygen species (ROS) class of cellular damaging agents but also has important physiological roles in 'redox-signalling' and in cellular differentiation and proliferation in higher order organisms. Excessive levels of H₂O₂ can lead to oxidative stress through the production of HO•, a highly toxic and reactive free radical, causing DNA damage and the progression of diseases such as cancer^{217,218}. ROS cause damage by reacting with a wide range of cellular components such as proteins, lipids and nucleic acids. This can have disastrous consequences reducing protein functionality, disrupting cellular membranes and in the accumulation of mutations, as well as leading directly to cell death²¹⁸. Cells have therefore developed mechanisms to protect themselves against oxidative damage. In *E. coli*, various peroxidases and catalases sequester H₂O₂ and rapidly degrade it to limit unwanted damage¹⁵¹. However, these protein systems may still become overwhelmed.

The most relevant forms of oxidative damage to this piece of work are those which manifest as DNA lesions. The most prominent lesion is 8-oxoguanine which is a major source of GC→AT transitions. In *E. coli* this damage is repaired by the DNA glycosylase

MutM as a form of BER as presented in section 1.5.2, although other proteins can act redundantly²¹⁹.

E. coli Δ *lhr* cells show increased sensitivity when grown in the presence of H₂O₂ as compared to wild type MG1655 cells. This is displayed in both cellular viability (Figure 5.4 A) and in growth rate (Figure 5.4 B). Unlike data presented in Figure 5.3 and Figure 5.5, viability spot tests were performed by growing H₂O₂ exposed cells in liquid culture for a further generation after reaching OD before spotting. This is in contrast to spotting directly onto LB Broth agar containing damaging agents of various concentrations but was necessary due to the reactivity of H₂O₂ within agar which generated bubbles, rendering the plates unsuitable. *Lhr* knockout cells were shown to take four to five times longer to reach exponential phase (OD₆₀₀ 0.2) in the presence of H₂O₂ but show a growth rate recovery after 330 minutes and reach a similar maximum OD as MG1655 cells. Growth recovery may be due to successful sequestering of ROS or the emergence suppressor mutations.

With reference to Figure 5.4 C, Δ *lhr* Δ *radA* and Δ *radA* cell strains also show increased sensitivity to damage caused by H₂O₂ as compared to MG1655. Δ *radA* cells show the greatest variation of survival represented by large error bars. The *lhr radA* dual knockout cells show improved survival as compared to Δ *lhr* suggesting the activation of an alternative repair pathway when both proteins are absent.

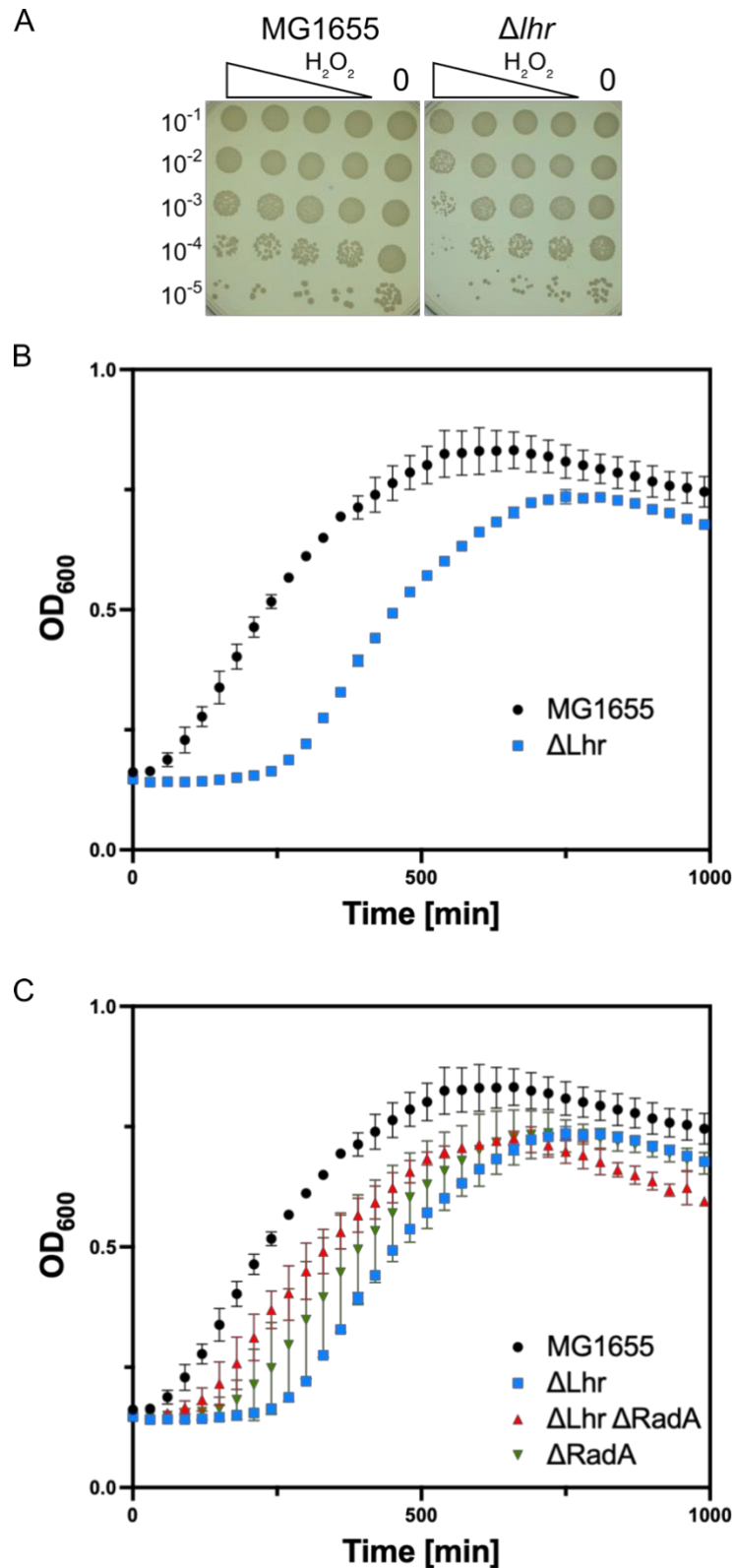


Figure 5.4 *E. coli* Lhr is involved in oxidative damage repair.

E. coli Δlhr cellular viability (A) and growth rate (B) are affected following exposure to 12.5 nM H_2O_2 . H_2O_2 sensitivity is reduced in $\Delta lhr \Delta radA$ cells (C). Graphs show data plotted from at least two repeats with error bars depicting standard error from mean.

Great difficulty was seen in obtaining repeats of data displayed in Figure 5.4. This may be due to the cells ability to recover from H₂O₂ insult as shown in Figure 5.4 **B** and **C**, causing repair phenotypes to be masked when scoring through viability spot assays. Another plausible reason is that Lhr is indirectly involved in the repair of 8-oxoguanine lesions or, is involved in the repair of rare forms of oxidative damage. This point is discussed further below in section 5.6 when investigating Lhr's preferred DNA substrate. This indirect involvement in oxidative damage repair and ability to recover to stationary phase may explain why Reuven *et al.* were unable to describe a similar genetic phenotype as displayed here. However, methodology in how they tested sensitivity of Δ *lhr* cell to H₂O₂ is not described so it is difficult to compare results¹¹⁹.

The involvement of RadA (Sms) in oxidative damage repair is not investigated further and is something which may be explored as a continuation of this work.

5.2.4 Lhr is not directly related to the repair of ICL or double strand DNA breaks

E. coli Lhr and RadA (Sms)'s involvement in interstrand crosslink (ICL) repair was also investigated. RadA (Sms) has been shown to promote branch migration of recombination intermediates and in promoting RecA-mediated strand invasion^{165,167}. Due to ICL damage requiring HR directed repair for resolution, it was logical to test RadA (Sms)'s involvement alone and in the context of Lhr, something which hasn't been documented before. To achieve this we performed viability spot assays in the presence of mitomycin C (MMC).

MMC is a potent DNA intercalator used as a chemotherapy agent for numerous cancers. Its toxic action is elicited by formation of a covalent cross-links between guanines of opposing strands, acting as a bulky adduct causing inhibition to many DNA metabolism proteins^{220,221}. MMC can also act as a source of oxidative damage²²². Repair of MMC adduct repair relies on NER, HR and TLS repair pathways which are presented in detail in section 1.5.

Cell strains lacking Lhr appear largely unaffected when grown in the presence of MMC (Figure 5.5, **blue box**). A mild reduction in cell viability is seen at the highest MMC concentration (1 µg/µl), but this difference is largely insignificant when compared to wild type MG1655 cells. When comparing viability spots in Figure 5.5 **A**, $\Delta lhr\Delta radA$ and $\Delta radA$ show a 10-fold reduction in cellular viability as compared to Δlhr and MG1655. This is reciprocated by the tightness of the **red** and **green** lines shown in Figure 5.5 **B**, suggesting involvement in ICL repair. Presentation of this mild phenotype may be due to redundancy of repair pathways in *E. coli*. Additional gene knockouts in combination with $\Delta radA$ may illuminate its importance in this type of repair.

The close alignment of $\Delta lhr\Delta radA$ and $\Delta radA$ suggest that for ICL damage repair, Lhr and RadA (Sms) are not involved in the same repair pathway. The negligible sensitivity of Δlhr cells further supports its absence in ICL repair.

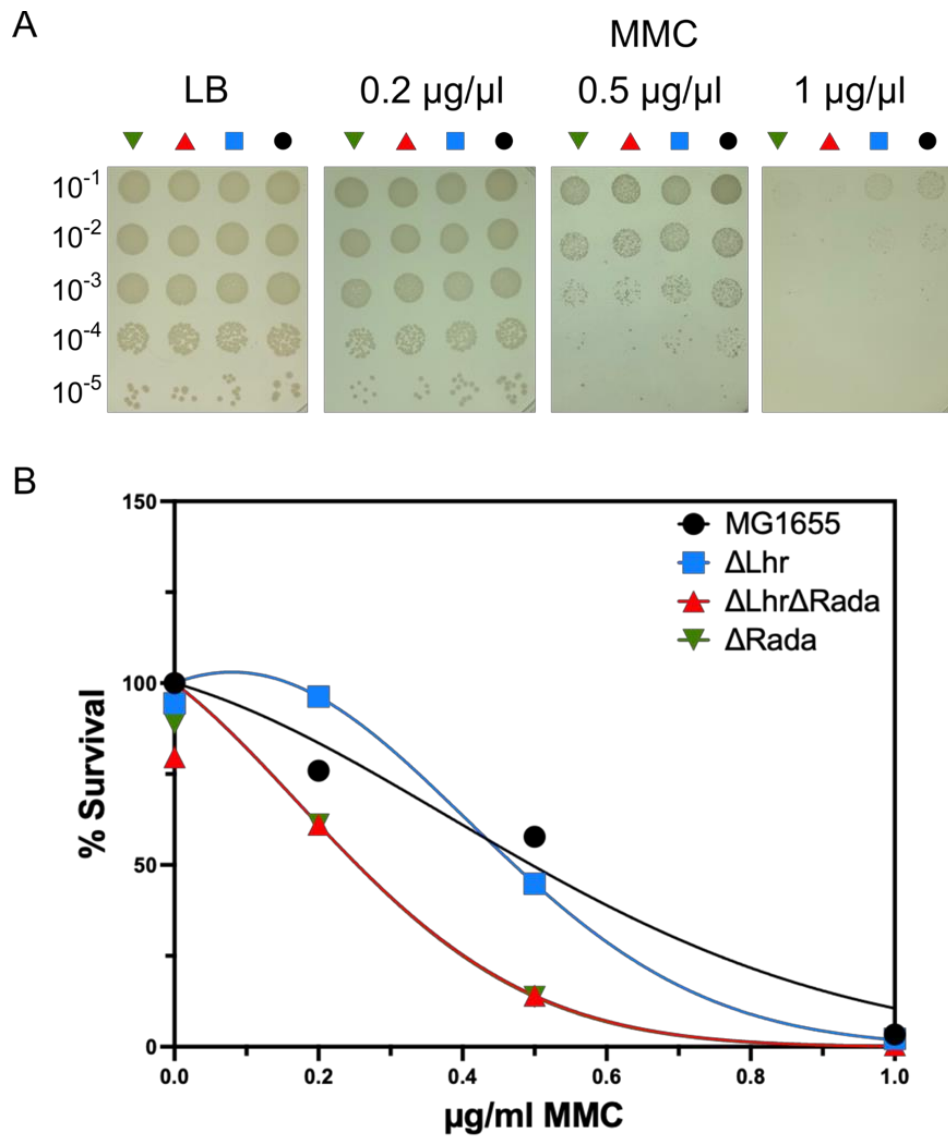


Figure 5.5 *E. coli* RadA (Sms) is involved in ICL repair but does not function alongside Lhr.

E. coli Δlhr knockout cells are largely unaffected when grown in the presence of mitomycin C (MMC). $\Delta lhr\Delta radA$ and $\Delta radA$ cells show identical loss of viability suggesting RadA (Sms) involvement is distinct from Lhr. Cells are still viable with *radA* deletion suggesting repair pathway redundancy. (A) Displays viability spot agar plate examples. (B) Is data of at least two repeats with quadratic survival curve plotted.

5.2.5 Lhr and RadA (Sms) may be part of a mutation inducing repair pathway

E. coli Lhr and RadA (Sms) knockout strains were subject to a 'rifampicin mutation repair phenotype' assay to determine the extent of random mutations in the absence of protein(s). This assay is scored through relative acquired resistance to rifampicin through accumulation of mutations in the *E. coli* RNA polymerase β subunit gene *rpoB*^{223,224}. Rifampicin is a broad spectrum antibiotic which binds RpoB with a high affinity causing inhibition through steric blockage of the elongating RNA strand²²⁴. This results in RNA polymerase pausing causing the complex to act as a roadblock which in turn inhibits other cellular functions such as DNA replication. Resistance develops through specific mutations of *rpoB* as detailed by Garibyan *et al.*, often occurring through single base substitutions²²⁵.

For this assay, cells are grown in LB Broth and are plated onto agar plates containing multiple rifampicin concentrations at various growth stages. This allows us to see if mutations arise at a specific point i.e. in stationary phase, or if they occur throughout the growth cycle. Cells within stationary phase often see an increase in mutation rates due to exhaustion of resources and a population crash due to cell lysis. This can change the properties of the surrounding media, such as its pH, causing selection of advantageous genetic traits²²⁶. Plating of cell samples onto LB Broth agar plates containing various concentrations of rifampicin gives an indication of mutation rate between cell strains. We hypothesised that our knockout cell strains would show a greater amount of colony formation and therefore acquired resistance due to the absence of Lhr and RadA (Sms) repair proteins and disruption of associated repair pathways.

With reference to Figure 5.6, wild type MG1655 cells showed the highest rate of acquired resistance as compared to the knockout cell strains. Both Δ/hr and $\Delta/radA$ single knockouts showed a decrease in acquired resistance as compared to MG1655. This difference maybe up to a 2x less, although colonies were not counted for this experiment. The most surprising result is the distinct lack of colonies on the $\Delta/hr\Delta/radA$ dual knockout. This suggests that this cell strain shows a reduced rate of mutation, directly contradicting our initial hypothesis. This result is repeated and shown in the '1 day' column as highlighted in **green** in Figure 5.6, which originally was a control as part of a similar experiment involving 'indole' as the mutagenic source (data not shown).

Observations are consistent across all rifampicin concentrations and 'cell growth' samples excluding those taken at OD₆₀₀ 0.4 (early exponential phase) where acquired resistance is identical.

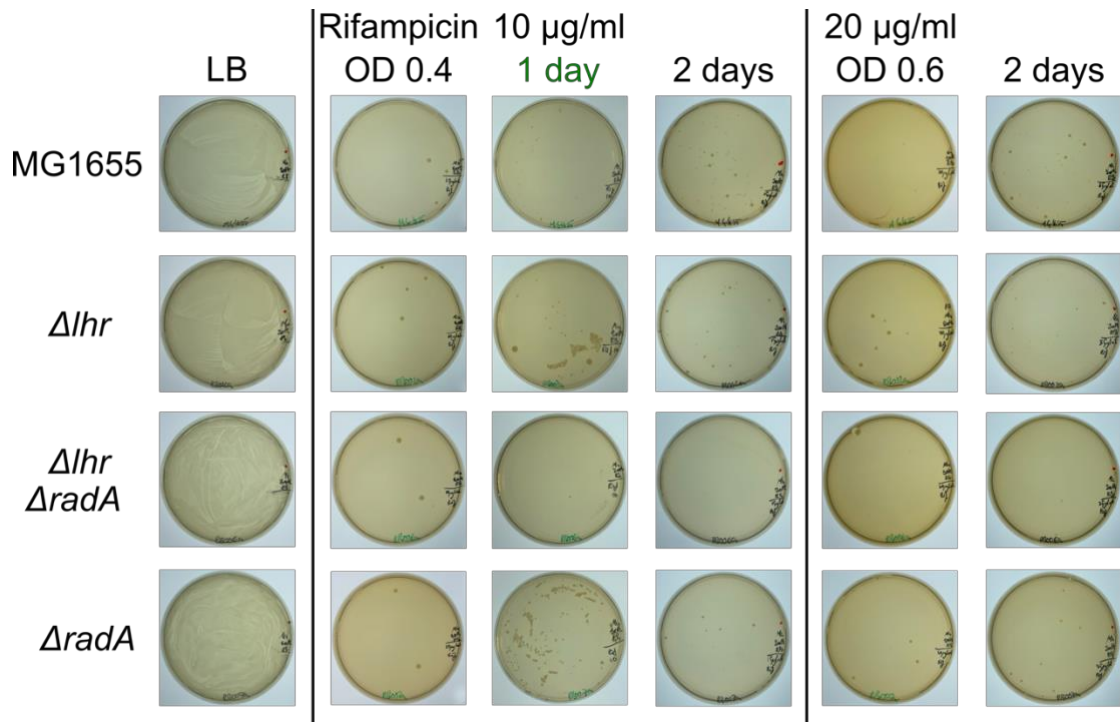


Figure 5.6 *E. coli* Lhr and RadA (Sms) knockout cells show reduced acquired resistance to rifampicin as compared to wild type cells.

E. coli MG1655, Δlhr , $\Delta lhr \Delta radA$, and $\Delta radA$ cells were plated onto rifampicin containing agar plates (10 µg/ml and 20 µg/ml as labelled) at various cell densities/growth stages. Resistance to rifampicin occurs through accumulation of specific point mutations in the *ropB* gene. Here, the $\Delta lhr \Delta radA$ cell strain shows a dramatically reduced level of acquired resistance as compared to wild type MG1655. The green '1 day' column shows an experimental repeat performed independently from the rest. Plates within the '2 day' column showed increased resistance of 'stationary phase' cells as compared to OD₆₀₀ 0.4 and OD₆₀₀ 0.6 'log phase' cells. Plates on the furthest left column, 'LB', are positive growth controls of each cell strain after 2 days of growth.

A 'hypomutation' phenotype is exceedingly rare with only three instances being reported. In these cases phenotypes present through controlled (often increased) expression of MMR proteins MutL^{227,228} or MutS²²⁹ leading to a state of adaptive mutation suppression. How these observations relate to Lhr and RadA (Sms) may only be justified through speculation. It must be noted that the $\Delta lhr\Delta radA$ deficient cells were much slower in reaching exponential phase as compared to the other cell strains. This may afford justification for a reduced mutation rate due to a greater chance of DNA repair per growth cycle. Further characterisation into when these proteins are expressed and in what repair pathways as well as a deeper investigation on their influence in cellular growth rate may shed light on to the importance of this result.

5.3 Purification of *E. coli* Lhr-extended

Protein overexpression and purification strategies were developed to investigate *E. coli* Lhr's potential role in DNA damage repair. Obtaining purified protein allows analysis by biochemical assays to support genetic data presented in section 5.2. Plasmid expression of a large, native *E. coli* protein using *E. coli* expression systems proved challenging.

E. coli expression systems are commonly used due to their relative cheapness as compared to eukaryotic systems, and the rapid growth seen when using rich broth media. Protein production is achieved using an expression plasmid cloned with recombinant protein. Expression plasmids are vast in features to allow tuned expression and often include affinity tags to aid protein purification. Expression with *E. coli* systems most commonly use B line cell strains such as BL21 and BL21 (DE3) which are deficient in Lon and OmpT proteases to reduce recombinant protein degradation. B line cells often contain a chromosomal T7 RNA polymerase under an inducible promoter to allow further control of protein expression²³⁰. Different combinations of cell line and plasmid allow optimisation of expression, dependent on individual protein characteristics.

5.3.1 Optimised expression of non-tagged *E. coli* Lhr

E. coli Lhr expression was initially attempted using a pACYC Duet based vector, cloned by Dr. Edward Bolt (pEB692). For this attempt, shown in Figure 5.7 A, *E. coli* BL21 AI cells were used. BL21 AI cells contain a chromosomal T7 RNA polymerase under an araBad promoter for induction upon addition of L-arabinose. Further expression control is achieved through the use of the pACYC based plasmid which inhibits protein production in the absence of IPTG. When added, IPTG causes the removal of the *lac* repressor from the T7 promoter allowing access by the T7 polymerase for protein production. Two stage protein expression control is often used for tight regulation to limit the toxic effects of recombinant proteins.

For pilot overexpression using this method, cultures were grown in baffled flasks at 37°C to OD₆₀₀ 1.2. Cells were chilled on ice before 0.8% IPTG and 0.2% L-arabinose was added for induction. Flasks were moved to 18°C for overnight growth after which 1ml samples were taken, spun down and prepared for SDS PAGE analysis. BL21 AI cells transformed with pACYC Duet empty vector was grown alongside as a negative control.

Successful protein expression is confirmed by comparing negative control and overexpression samples. Full length Lhr is expected at 169 kDa. With reference to Figure 5.7 A, limited full length protein expression can be seen (lane 3, labelled o/e). A highly abundant band of ≈35 kDa is highlighted. Presence of this band due to degradation would likely show further bands of varying length totaling 169 kDa. Their stark absence suggests this protein species to be a truncated N-terminal Lhr product as opposed to protein degradation. Further investigation into why truncated expression would occur lead to the discovery of *lhr*'s rare codon composition (data

presented in section 3.2.3). From this, we discovered a series of rare codons located upstream of a potential pre-mature termination site. We theorize that inclusion of rare codons cause RNAP stuttering (lag due to absence of correct tRNA) and slippage resulting in pre-mature termination (Figure 3.8). This would result in production of a 79 kDa protein which subsequently degrades into two 35 kDa products and is observed as seen below.

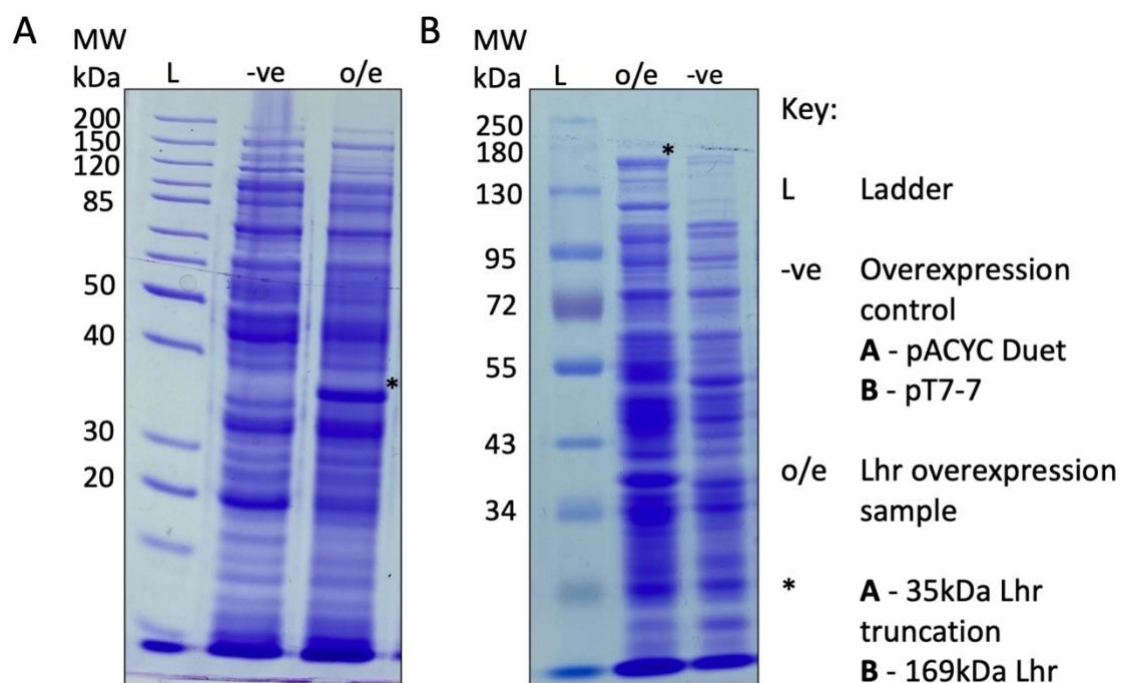


Figure 5.7 Overexpression optimisation of non-tagged full length *E.coli* Lhr.

Coomassie stained 8% acrylamide SDS PAGE analysis of Lhr pilot overexpression samples. **(A)** Overexpression samples of BL21 AI cells transformed with a plasmid containing Lhr in pACYC Duet (o/e) and pACYC Duet empty vector control (-ve). **(B)** Overexpression samples using Rosetta 2 (DE3) cells transformed with a plasmid containing Lhr in pT7-7 (o/e) and pT7-7 empty vector as control (-ve). Bands of interest are highlighted with an ‘*’ as indicated.

To navigate overexpression problems due to *lhr*'s rare codon composition, the *lhr* gene from pEB692 was PCR amplified to insert a C-terminal HindIII restriction site for cloning into pT7-7 (pRJB15, ampicillin). This plasmid is often used for genetic and overexpression studies. This allowed pilot overexpressions to be performed using Rosetta 2 (DE3) *E. coli* cells. Rosetta 2 (DE3) carry a chromosomal copy of T7 RNA polymerase under control of the *lacUV5* promoter, and a chloramphenicol resistant plasmid which supplies the cell with 7 rare codon tRNAs. pT7-7 is known as a 'leaky' plasmid due to the absence of a plasmid repression system however, due to the regulation control afforded by the Rosetta 2 (DE3) cells, expression is tightly controlled to only occur upon induction by IPTG.

For pilot overexpression, cultures were grown in baffled flasks at 37°C to OD₆₀₀ 1.2. Cells were chilled on ice before 0.8% IPTG was added for induction. Flasks were moved to 18°C for overnight growth after which 1 ml samples were taken, spun down and prepared for SDS PAGE analysis. With reference to Figure 5.7 B, non-tagged full length *E. coli* Lhr was successfully overexpressed using pRJB15 transformed into Rosetta 2 (DE3) cells. Identification of full length Lhr (marked with an '*') is achieved upon comparison between lanes 2 and 3. The truncated ≈35 kDa protein species is still seen within the sample but with full length protein at a much higher abundance than in A. An enriched protein band can be seen in the 'o/e' lane of Figure 5.7 B situated below the 130 kDa marker. This band is absent from the adjacent '-ve' control lane and is thought to be due to Lhr protein degradation. Addition of this band and the ≈35 kDa protein species does not quite equal the total length of *E. coli* Lhr suggesting the origin of these two protein species may occur independently. This method was upscaled x30 to generate biomass for purification attempts.

5.3.2 Purification of non-tagged *E. coli* Lhr-extended

Protein purification steps were optimised by trial and error using multiple FPLC chromatography columns. Initial Lhr purification attempts followed similar methodology to that stated in Buckley *et al.*¹¹⁰ however, *E. coli* Lhr displayed limited stable binding to a heparin column. Instead, a butyl sepharose column was introduced at the first step. Butyl sepharose columns are a form of 'reverse-phase chromatography' whereby protein mixtures (mobile polar phase) are separated dependent on the level of hydrophobic interactions with the butyl side groups (organic fixed phase) of the column. Soluble proteins are loaded in buffer containing 1.5 M ammonium sulphate. This decreases the availability of water molecules within the solution through water solvation of the salt ions. This exposes hydrophobic patches on the proteins surface to allow interactions with the chromatography column. Protein elution is achieved through a reduction of salt concentration²³¹. Resuspending harvested cells in this buffer proved vital in increasing full length Lhr abundance within the soluble fraction post sonication. This increased the relative abundance of available Lhr protein to a sufficient level for purification attempts.

To purify, soluble cell lysate was loaded onto a butyl sepharose column and eluted over multiple fractions against a gradient of decreasing ammonium sulphate. Lowering the amount of ammonium sulphate within the buffer weakens hydrophobic interactions allowing protein elution. Suspected Lhr fractions, shown in Figure 5.8, were pooled for overnight dialysis into suitable buffer for loading onto a heparin column.

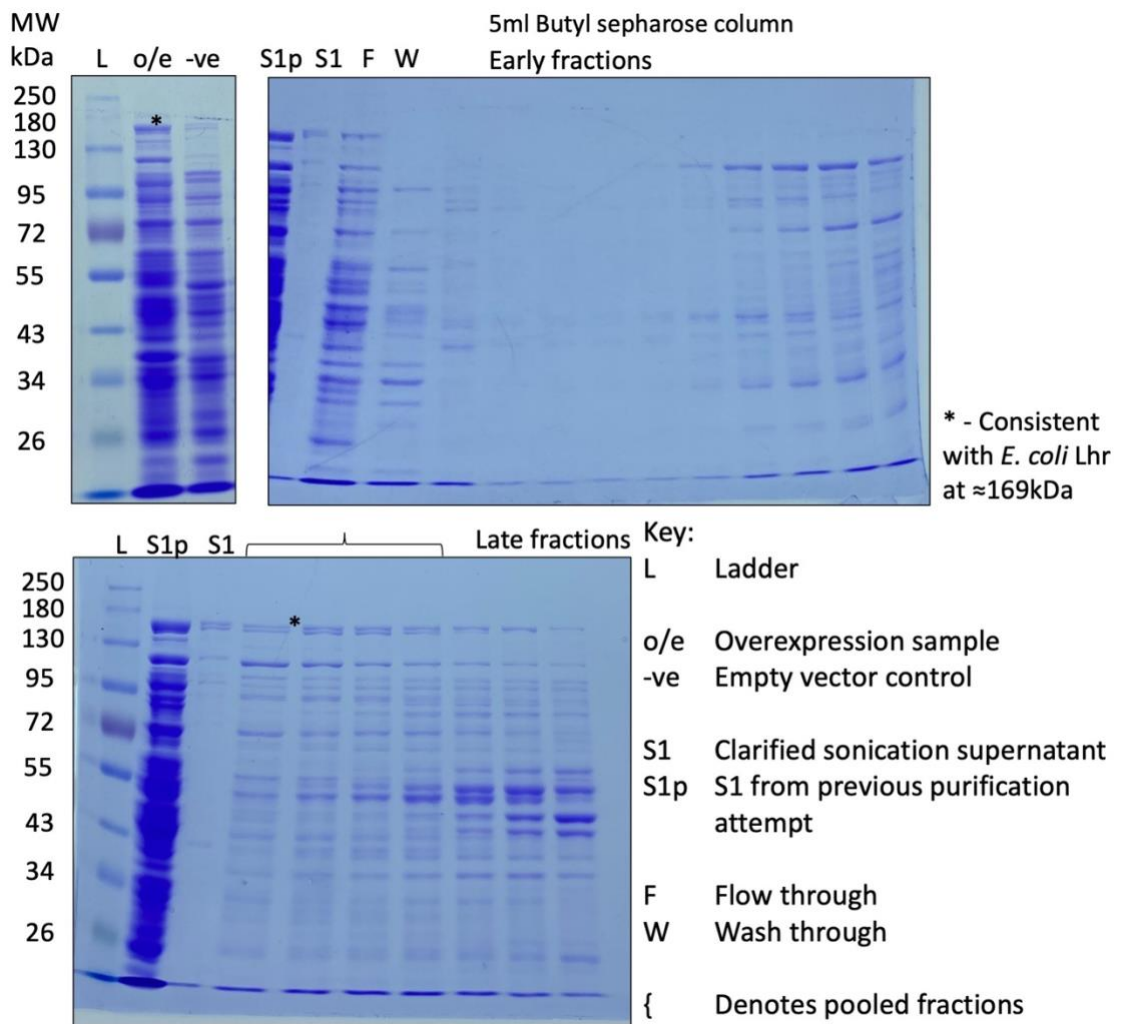


Figure 5.8 *E.coli* Lhr elutes from a butyl sepharose column at a range of salt concentrations.

Coomassie stained 8% acrylamide SDS PAGE analysis of initial non-tagged Lhr purification steps. Pilot overexpression gel is included for reference. Bands of interest are highlighted with an ‘*’ as indicated.

A heparin column contains sulphated polysaccharides which mimic the DNA phosphodiester backbone. This property is often exploited when purifying suspected DNA binding proteins. Here, the majority of pooled *E. coli* Lhr displays weak association and elutes within the first fraction of a gradient of increasing ionic strength. Some Lhr elutes in the later fractions but in too low a concentration to be used in subsequent steps. Fractions were pooled as denoted in Figure 5.9. Pooled

fractions are then loaded directly onto a Q sepharose column without needing further dialysis due to the low salt concentration afforded from early elution from the heparin.

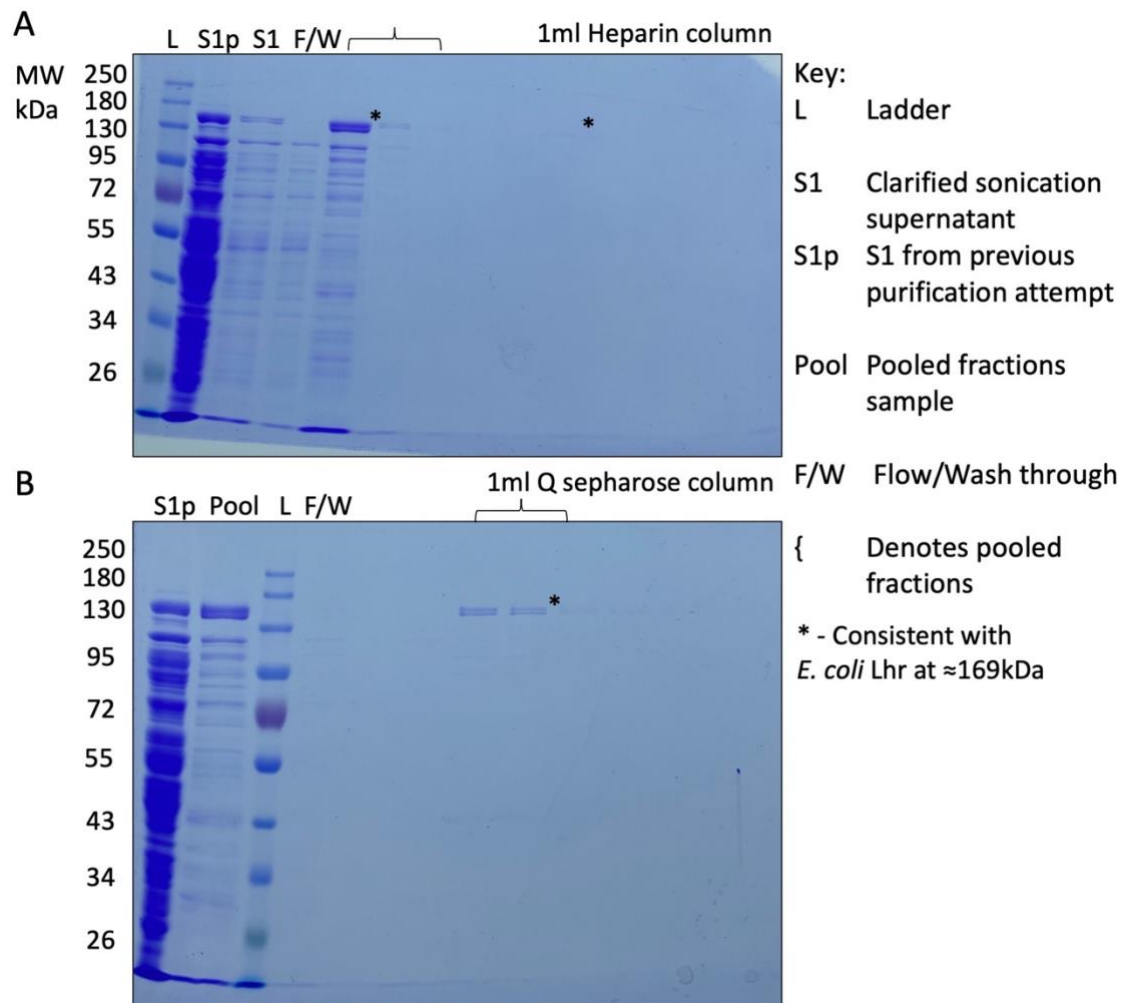


Figure 5.9 *E. coli* Lhr is purified using ion exchange chromatography.

Coomassie stained 8% acrylamide SDS PAGE analysis of final non-tagged Lhr purification steps. **(A)** Further isolation of Lhr proteins using Heparin affinity chromatography. **(B)** Purification of Lhr by Q sepharose affinity chromatography. Bands of interest are highlighted with an ‘*’ as indicated.

E. coli Lhr was shown to bind well to a Q sepharose column due to possessing a predicted charge of -12.8 in pH 8 (pI of 6.75), allowing strong binding to the positive beads of the column. Pooled fractions presented as a purified doublet with a small

abundance of suspected degradation product of ≈ 40 kDa. Purified protein sample was used in initial biochemical analysis to ascertain ability to bind DNA.

Approximate protein concentration was determined using the DeNovix spectrophotometer absorption reading at 280 nm and Lhrs' extinction coefficient value (178105), which were applied to the Beer-Lambert law. To determine protein concentration more accurately multiple assays may be employed such as Bradfords or BCA, however in this case this was not needed

5.3.3 Purification of his-tagged *E. coli* Lhr-extended

Multiple purifications were attempted using the 'non-his tagged' method as described in section 2.10.2 and presented in section 5.3.2. Repetition of this method proved difficult and unreliable for success. To remedy this, we sought to clone in a histidine tag to aid in future purification attempts. Affinity tags are a powerful tool widely used in protein purification due to their highly specific properties. Histidine is able to form strong coordination bonds with the immobilised ions which 'charge' the columns matrix, this interaction is facilitated through the electron donor groups of histidine's imidazole ring. Bound proteins are then displaced through competitive binding upon addition of free imidazole contained within the elution buffer²³².

For this work, we inserted a 6x His-tag between the first and second codon of *lhr* from pRJB15 (resulting plasmid named pRJB28). This is a relatively unstructured region as shown on PyMOL, so we predicted it would have little effect on protein function or assembly.

To purify, soluble cell lysate was loaded onto a butyl sepharose column and subsequently washed with 60% 'No Salt buffer B' to remove weakly bound contaminating proteins. This wash step was determined using Figure 5.8 as a reference during the purification of non-tagged Lhr. A pre-equilibrated Ni-NTA column was then attached in tandem and both columns were washed with 100% 'No Salt buffer B'. This caused *E. coli* Lhr to elute from the butyl sepharose column and bind to the Ni-NTA through interactions with the histidine tag. The butyl sepharose column was detached and proteins were eluted from the Ni-NTA column over multiple fractions against a gradient of increasing imidazole. Increasing the amount of

imidazole competitively removes bound proteins allowing protein elution. Two pools were selected for overnight dialysis into suitable buffer for loading onto a Q sepharose column. Pool 1 appeared to contain a doublet of similar size to that purified in Figure 5.8 but was markedly more dirty after Q sepharose elution. Pool 2 was much cleaner but did not contain the *E. coli* Lhr doublet previously see in Figure 5.9 **B**.

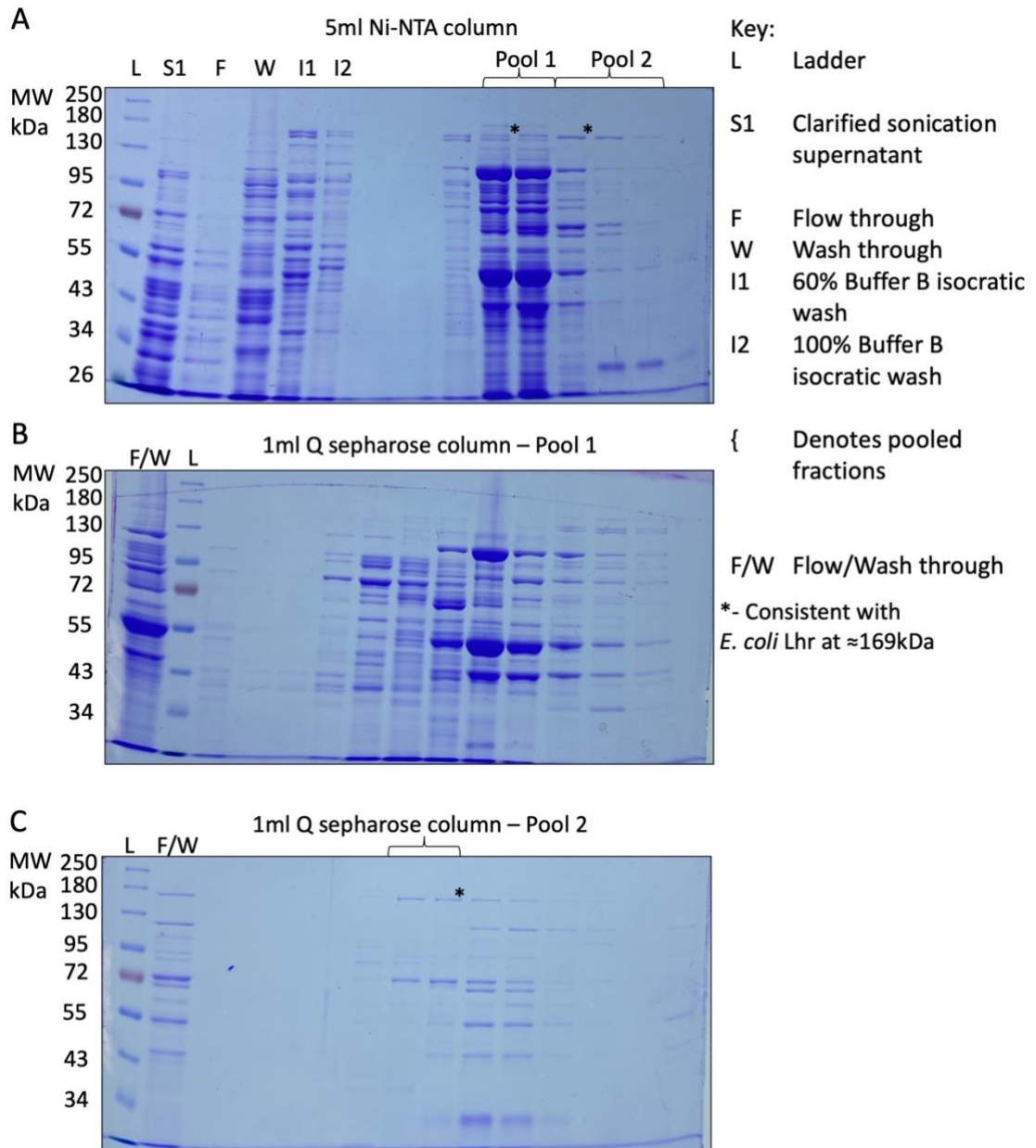


Figure 5.10 Purification of *E. coli* Lhr.

Coomassie stained 8% acrylamide SDS PAGE analysis of final his-tagged Lhr purification steps. (A) Butyl sepharose wash and Ni²⁺-NTA chromatography. Two pools containing suspected Lhr were dialysed into suitable buffer for Q sepharose column loading. (B) Q sepharose chromatography of 'pool 1'. Suspected Lhr remains dirty and was not taken further. (C) Q sepharose chromatography of 'pool 2'. Full length Lhr was purified as indicated with suspected degradation products. Bands of interest are highlighted with an '*' as indicated.

E. coli Lhr again bound well to a Q sepharose column eluting over a gradient of increasing ionic strength. Fractions eluting from pool 1 remained highly contaminated so were not purified further or dialysed for storage. Fractions containing *E. coli* Lhr eluting from pool 2, as indicated in Figure 5.10 C, were dialysed for storage. Protein was used in glycosylase activity assays and EMSA binding studies. Extra bands seen in Figure 5.10 C are attributed to protein degradation products and do not appear to have uncharacteristic inhibitory or additive protein function as determined through biochemical analysis (as presented below).

Approximate protein concentration was determined using the DeNovix spectrophotometer absorption reading at 280 nm and Lhr's extinction coefficient value (178105), which were applied to the Beer-Lambert law as before.

5.4 Purification of *E. coli* C-Lhr

5.4.1 Cloning of *E. coli* C-Lhr

E. coli lhr residues 876-1538 were amplified using polymerase chain reaction (PCR) and cloned into pNH-TrxT using ligation independent cloning. This was performed by Nadia Ahmed in Dr. Christopher Cooper's lab (University of Huddersfield). This plasmid was denoted as pRJB23 for this work. pRJB23 was transformed into competent *E. coli* Rosetta 2 (DE3) cells and an overexpression was performed as described in section 2.10.3.

5.4.2 Purification of *E. coli* C-Lhr

To purify, soluble cell lysate was loaded onto a Ni-NTA column to allow purification from contaminating proteins through selection of the histidine tag. Fractions were collected over an increasing gradient of imidazole for protein elution. For this purification an old chromatography column was used. This had an unusual effect on protein binding, causing a small amount of C-terminal Lhr to elute pure in the later fractions. Nominal his-tagged protein elution is achieved well before the imidazole concentration reaches $\approx 300\text{mM}$. Here, protein is eluting well passed this. Unfortunately, some time after this purification the column was broken and its unusual properties were lost forever. This led me to develop this purification further to purify mutant C-Lhr proteins as described below.

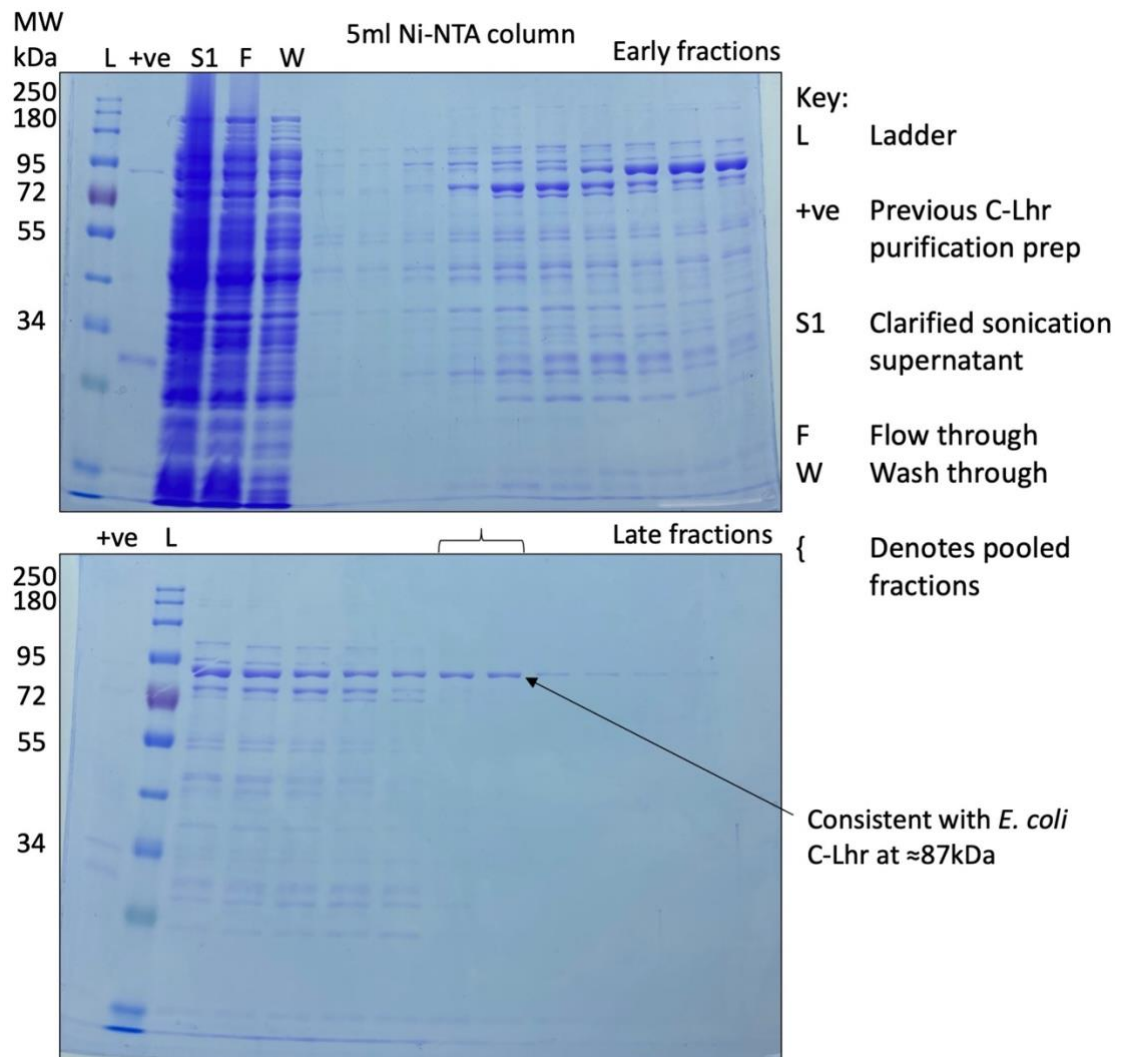


Figure 5.11 Purification of *E. coli* Lhr C-terminus.

Coomassie stained 10% acrylamide SDS PAGE analysis of C-terminal Lhr purification steps. Purification achieved through Ni²⁺-NTA chromatography.

Fractions containing purified C-Lhr were pooled and dialysed for storage. Protein was used in glycosylase activity assays and EMSA binding studies.

Approximate protein concentration was determined using the DeNovix spectrophotometer absorption reading at 280 nm and Lhr's extinction coefficient value (178105), which were applied to the Beer-Lambert law as before.

5.5 Investigation into *E. coli* Lhr DNA binding capacity

Electrophoretic mobility shift assays (EMSAs) were used to assess the binding abilities of full length *E. coli* Lhr (FL-Lhr) and C-terminal Lhr (C-Lhr) as described in section 2.9.3. EMSAs rely on the relative migration capacity of free DNA (fast migration) compared to that which is bound by protein (slower migration). These assays allow simplistic evaluation of DNA binding capacity relatively quickly²³³. Quantitative assessment may be achieved with additional methods such as anisotropy, but for this piece of work this was deemed not necessary.

Protein concentration titrations were performed using 5' Cy5-labeled DNA substrates as indicated in respective figures. Protein-nucleic acid complexes were visualised using a Typhoon phosphor-imager after migration on a 5% native polyacrylamide gel.

5.5.1 *E. coli* Lhr requires single stranded DNA for binding

Lhr's binding capacity was investigated using single stranded DNA (Figure 5.12 A), flayed duplex DNA (Figure 5.12 B, left) and duplex DNA (Figure 5.12 B, right) substrates.

Lhr is able to bind ssDNA and flayed duplex DNA at concentrations as low as 12.5 nM (Figure 5.12 A and B, first Lhr concentration) and achieves full binding of 12.5 nM of Cy5-labelled DNA at a concentration of 200 nM. At higher Lhr concentrations, protein-nucleic acid aggregation occurs, whereby bands appear 'stuck' with the gels wells. This is most evident in Figure 5.12 B as highlighted. Lhr appears unable to stably bind a duplex DNA substrate (Figure 5.12 B, right titre), although some binding may be present when considering 'band streaking'. Inability for Lhr to bind duplex DNA for function is supported below in section 5.6.4, through lack of observable glycosylase activity. This data is further supported in section 4.3.2 through *MthLhr*'s inability to unwind duplex DNA¹¹⁰. DNA binding displayed for flayed duplex DNA (Figure 5.12 B, left titre) shows both stable complex formation and protein-nucleic acid aggregation, this may be due to the substrate having both duplex and single stranded regions so shows both interactions. *E. coli* Lhr's preference for regions of ssDNA to elicit function is supported by data presented by Warren *et al.* and confirms our Lhr is behaving normally¹²⁵.

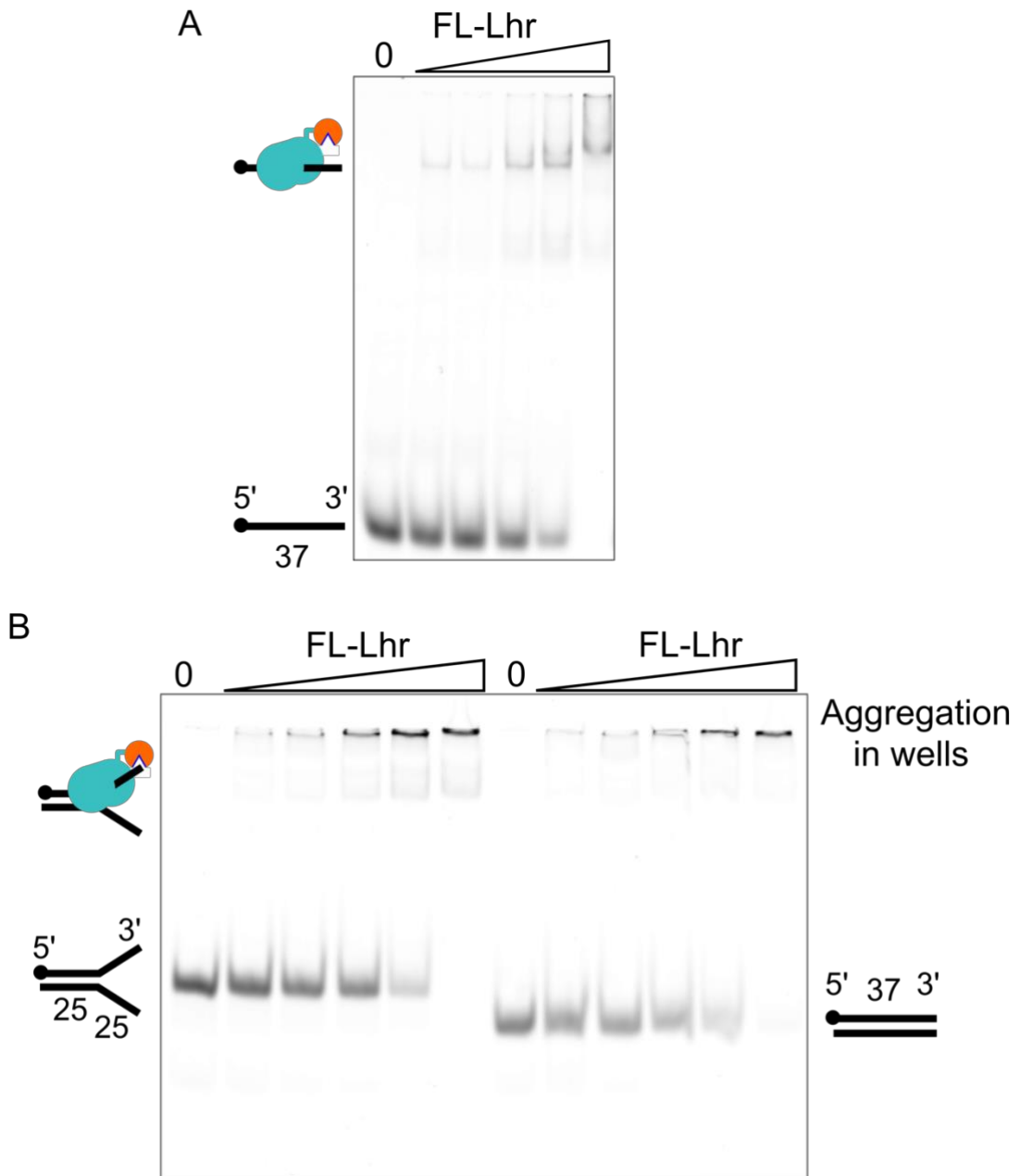


Figure 5.12 *E. coli* Lhr requires regions of ssDNA for stable binding.

5% Native acrylamide EMSA gels showing binding abilities of full length Lhr on 12.5 nM of various 5' Cy5 labelled DNA substrates. Protein concentrations increase from 12.5 to 200 nM linearly. **(A)** Lhr is able to bind ssDNA to a high affinity achieving full binding at 200 nM of protein. **(B)** Lhr shows binding and protein-nucleic acid aggregation on substrates containing regions of fully base paired DNA. Lhr is able to stably bind a flayed duplex but not a blunt ended DNA substrate. Full DNA immobilisation is again achieved at 200 nM of protein.

5.5.2 *E. coli* Lhr stably binds 'damaged' DNA substrates

E. coli Lhr's extended C-terminus shows great similarity to the *Streptomyces sahachiori* glycosylase AlkZ, as discussed in section 4.3.6 and presented by Werner *et al.*^{110,125}. To investigate Lhr glycosylase activity, we constructed DNA substrates containing modifications which mimicked chemical characteristics of DNA damage. These substrates contained d-uracil and 8-oxo-d-guanine nucleotides. The biological context and occurrence of these two nucleotides is presented extensively in sections 1.4 - 1.5 and will be discussed further below in relation to Lhr.

The protein concentration titre, as presented in Figure 5.13, shows full length Lhr is able to bind to d-uracil containing ssDNA to a similar affinity to 'undamaged' ssDNA. Full binding of 12.5 nM DNA substrate is achieved at a protein concentration of 200 nM for 'undamaged' DNA. Full substrate binding appears to occur at a lower protein concentration for the d-uracil DNA substrate however, this substrate appears to be of lower overall intensity as compared to 'undamaged' DNA so falsifying this observation. Lhr binding to the d-uracil substrate appears to lack a strongly defined protein-nucleic acid band. This may be due to limited glycosylase action elicited by Lhr, even in the presence of EDTA, which alters its binding profile (see Figure 5.19).

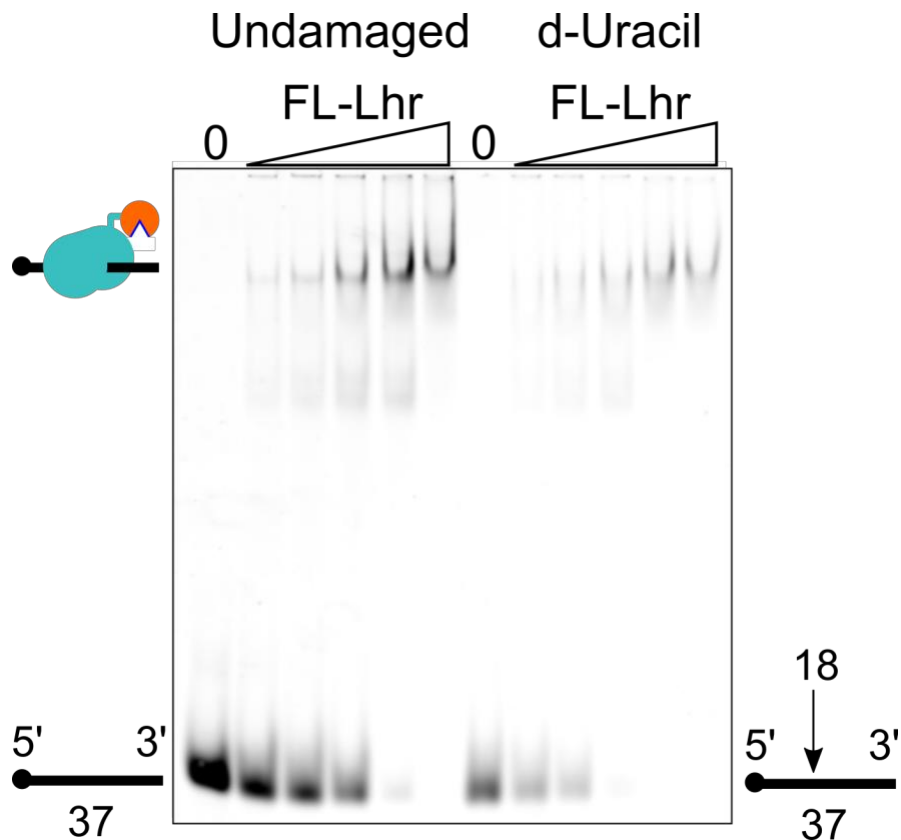


Figure 5.13 EMSA of *E. coli* Lhr's ability to bind a d-uracil containing DNA substrate. 5% Native acrylamide EMSA gel showing binding abilities of full length Lhr on 12.5 nM of 'undamaged' (left) and d-uracil 'damaged' (right) 5' Cy5 labelled DNA substrates. Protein concentrations increase from 12.5 to 200 nM linearly. DNA binding species are slightly different between substrates potentially due to limited Lhr glycosylase activity. Full DNA immobilisation is achieved at 200 nM of protein.

In both Figure 5.12 A and Figure 5.13 intermediate binding species can be seen. The presence of lower molecular weight bands, which seemingly shift to a higher molecular weight as protein concentration is increased, suggest multiple Lhr-DNA binding capacities. The ability for Lhr to bind cooperatively on ssDNA and not just as a monomeric protein may be investigated further to fully understand the presence of these bands. This may be achieved through size exclusion chromatography. This method may also illuminate differences in binding capacity between DNA substrates.

5.5.3 Lhr-CTD is unable to stably bind DNA in an EMSA

DNA binding ability was similarly investigated with purified C-terminal Lhr protein. With reference to Figure 5.14, C-Lhr is unable to stably bind DNA within this assay. This data may suggest C-Lhr DNA binding is either enhanced when part of the full protein, C-Lhr requires the 'helicase core' for DNA interaction, or this method of detecting DNA binding is inappropriate for C-Lhr. Justification for weak or transient DNA binding may be found when considering UvrAB and the Mut repair pathways where glycosylase action is promoted through interactions with additional proteins such as the DNA helicase. Here, the glycosylase and helicase functions would be tethered within a single protein.

Further interrogation into C-Lhr's ability to bind DNA may be achieved through more sensitive methods such as anisotropy as suggested above.

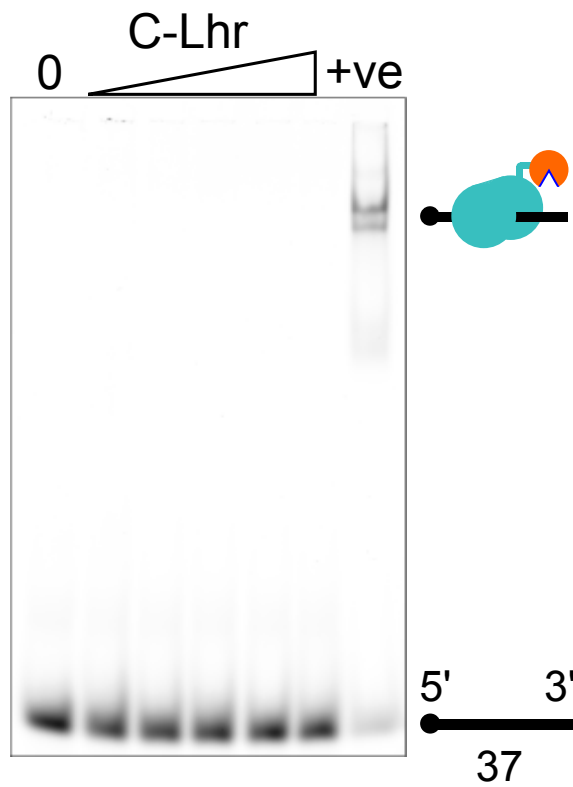


Figure 5.14 *E. coli* Lhr-CTD is unable to stably bind a ssDNA substrate.

5% Native acrylamide EMSA gel showing binding ability of C-Lhr on 12.5 nM of 5' Cy5 labelled ssDNA substrate. Protein concentration increases from 12.5 to 200 nM linearly. C-Lhr is unable to bind DNA stably as compared to a full length Lhr control (40 nM).

5.6 Investigation into *E. coli* Lhr DNA glycosylase activity

Initial EMSA testing of non-tagged full length Lhr displayed a mysterious band migrating faster than the 37 nucleotide no protein control. When a repeat assay was analysed using a denaturing acrylamide gel, it revealed multiple small DNA products of various sizes suggesting the presence of DNA cutting activity (data not shown). Due to limited non-tagged full length Lhr protein and the relative ease of purifying C-terminal Lhr, we decided to investigate and optimise nuclease/glycosylase assays using only C-Lhr.

Denaturing urea acrylamide gels allow separation and visualisation of RNA or DNA species with a resolution able to distinguish between single nucleotide differences. Preparation of the reaction sample before loading and subsequent running of the gel denatures nucleic acid secondary structure for true molecular weight migration²³⁴.

Protein concentration titrations and time course reactions were performed using 5' Cy5-labeled DNA substrates as indicated in respective figures, and as detailed in section 2.9.5. Protein-nucleic acid complexes were visualised using a Typhoon phosphor-imager after migration on 15% or 18% denaturing polyacrylamide gels as stated. A 'ladder' lane containing 5' Cy5-labeled DNA of 50, 37, 20, 18 and 16 nucleotides was often run alongside glycosylase reactions to allow determination of DNA product sizes.

5.6.1 *E. coli* Lhr-CTD acts as a d-uracil stimulated DNA glycosylase

C-terminal Lhr glycosylase activity was investigated on DNA substrates containing d-uracil and 8-oxo-d-guanine. Addition of NaOH proved key to resolving the glycosylated product for visualisation and migration of its true nucleotide length (comparison of Figure 5.15, lower gel). NaOH acts upon the abasic (AP) site, generated by Lhr glycosylase activity, causing backbone cleavage by β/δ elimination, similar to the action of an AP lyase²³⁵. Mild nuclease chewing can be seen on the 'undamaged' DNA substrate (Figure 5.15, top gel). This is in contrast to the targeted formation of a small DNA product of below 20 nucleotides as seen in the d-uracil containing reaction (lower gel). C-Lhr showed limited activity on an equivalent DNA substrate containing 8-oxo-guanine. This gel has been omitted due to the inclusion of Figure 5.20, where full length Lhr also shows no glycosylase activity in the presence of 8-oxoguanine.

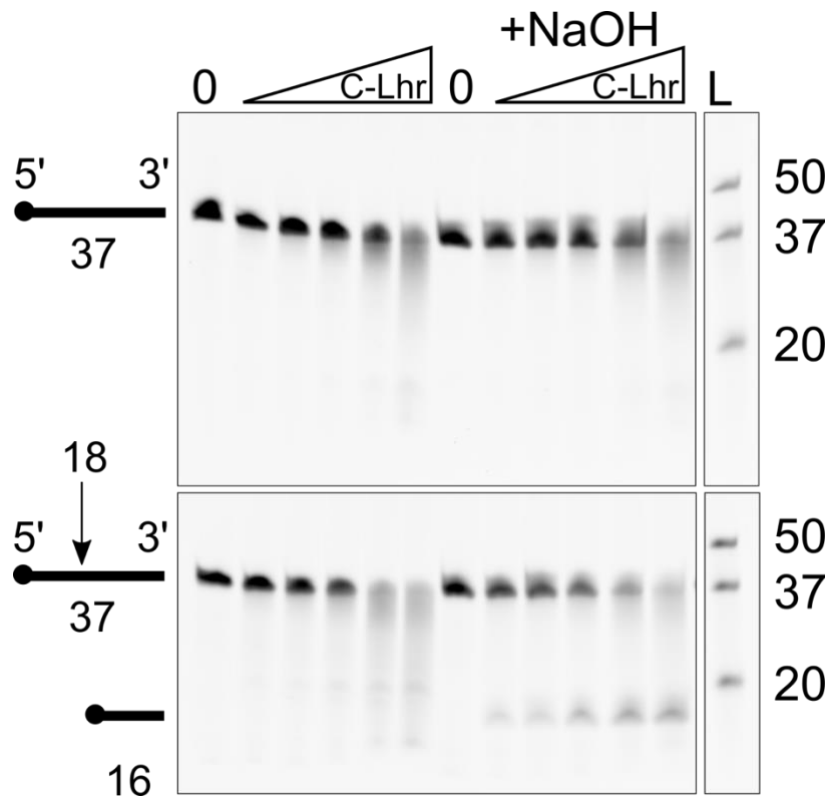


Figure 5.15 *E. coli* Lhr-CTD glycosylase activity on ‘undamaged’ and d-uracil ‘damaged’ DNA substrates.

15% Denaturing acrylamide gel showing a concentration titration of C-Lhr acting on 12.5 nM of 5' Cy5-ssDNA (top) and 5' Cy5-ssDNA of equivalent sequence containing a d-Uracil base 18 nucleotides from the fluorescent moiety as indicated (bottom).

Addition of NaOH causes β/δ elimination of the DNA species formed by C-Lhr activity resulting in DNA backbone cleavage confirming C-Lhr mediated glycosylase activity.

Protein concentration increases from 50 to 800 nM linearly.

5.6.2 *E. coli* full length Lhr shows focused glycosylase activity in the presence of Mn²⁺

The effect of different metal ions on glycosylase activity was investigated using his-tagged full length Lhr. Previous assays showed in the presence of Mg²⁺ ions, FL-Lhr would produce 'glycosylase' products of variable sizes. We theorised that this may be due to 1) activation of glycosylase activity by identification of the d-uracil base and 2) 'helicase-core' movement in the presence of Mg²⁺ causing off-target effects. To uncouple the 'helicase-core' from Lhr's glycosylase domain, we substituted Mg²⁺ for Mn²⁺. As can be seen in Figure 5.16, this achieved the desired effect. This assay also allowed us to ascertain the nucleotide position to which Lhr was breaking the glycosidic bond, two nucleotides upstream from the d-uracil base. The biological relevance of using manganese instead of magnesium is not known but using it within this context allows us to visualise Lhr's glycosylase activity more effectively.

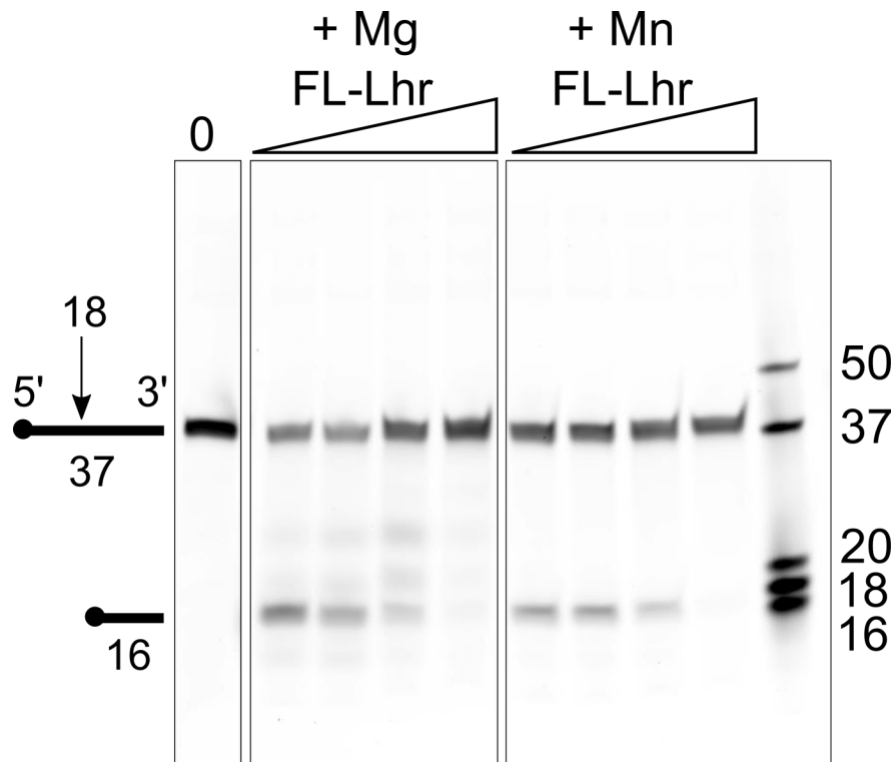


Figure 5.16 *E. coli* full length Lhr glycosylase activity in the presence of Mg^{2+} and Mn^{2+} .

15% denaturing acrylamide TBE gel showing FL-Lhr glycosylase activity on 12.5 nM ss-d-uracil containing DNA in the presence of $MgCl_2$ (left) and $MnCl_2$ (right). The d-uracil nucleotide is located 18 nucleotides from the 5' Cy5 moiety as shown. Protein concentrations increase from 25 to 200 nM linearly. Loss of activity is seen at higher concentrations thought to be due to protein-nucleic acid aggregation or protein auto-inhibition.

Increasing Lhr protein concentration resulted in loss of glycosylase function. This may be due to protein-nucleic acid aggregation (as seen in Figure 5.12) and/or auto-inhibition due to molecular crowding. The later would cause a 'loss of function' through steric inhibition.

5.6.3 Investigation of *E. coli* full length and Lhr-CTD glycosylase activities

Comparable analysis of FL-Lhr and C-Lhr glycosylase activities on d-uracil containing ssDNA substrates was performed. FL-Lhr is shown to have as much as 4x more glycosylase activity as compared to C-Lhr. This may be due to the full length proteins ability to stably bind DNA allowing improved targeting by its C-terminus. Comparison of FL-Lhr and C-Lhr activity in Figure 5.17 B show both Lhr species generating a DNA product of the same length.

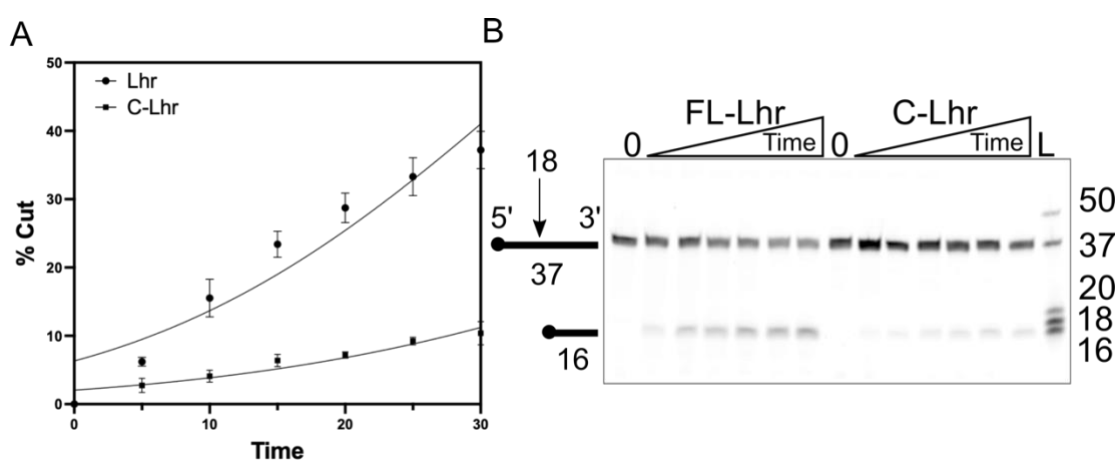


Figure 5.17 Comparative analysis of *E. coli* full length Lhr and Lhr-CTD glycosylase activities on a ssDNA d-uracil containing substrate.

Data generated using 50 nM of protein on 12.5 nM 5' Cy5 labelled DNA substrate. (A) Graphed data detailing the processivity differences between FL-Lhr and C-Lhr. Full length Lhr shows the formation of 4x more glycosylase product as compared to C-Lhr. Graphed data was repeated twice with standard error bars shown and Cumulative Gaussian fitted. (B) Example of 18% denaturing acrylamide TBE gel showing FL-Lhr and C-Lhr glycosylase activities on ss-d-uracil containing DNA. Time course reactions were completed over the course of 30 minutes.

5.6.4 *E. coli* full length Lhr glycosylase activity was investigated on multiple substrates

Full length Lhr was shown to preferentially target a d-uracil containing flayed duplex DNA substrate as compared to sequence equivalent ss- and duplex DNA. As shown in Figure 5.18 A, almost all flayed duplex DNA substrate is glycosylated after 30 minutes. This is almost twice as much as for ssDNA. Lhr is unable to glycosylate a duplex DNA substrate. This action is contrary to other uracil DNA glycosylases such as UDG²³⁶. Unwinding and protein action elicited on the parental strand of a DNA fork substrate is in support of data as presented in section 4.3.4¹¹⁰. This may suggest a potential role for Lhr to unwind and remove inhibitory substrates ahead of the replication fork.

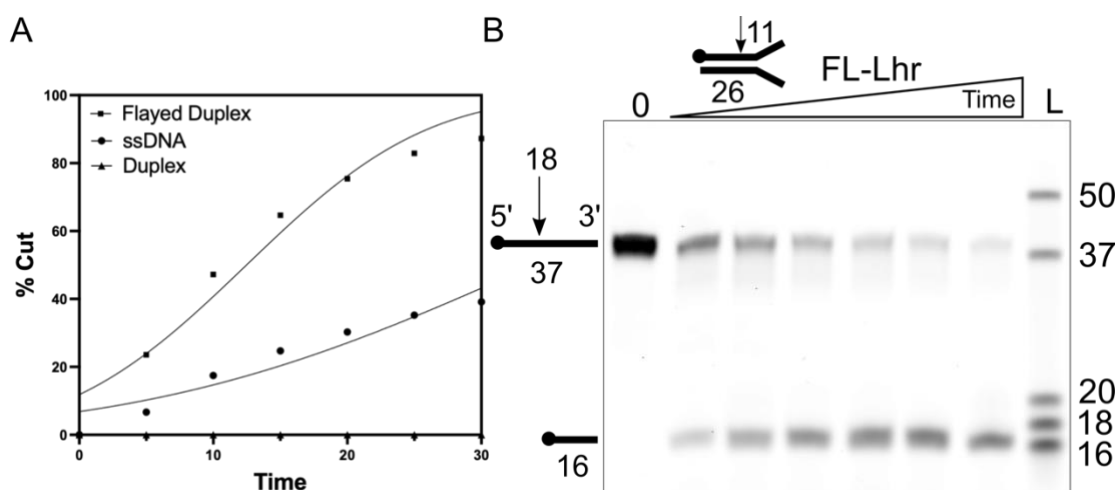


Figure 5.18 Comparative analysis of *E. coli* full length Lhr glycosylase activities on multiple d-uracil containing DNA substrates.

(A) Graphed data detailing the processivity differences between 80 nM of FL-Lhr on 12.5 nM of d-uracil containing ss-, ds- and flayed duplex DNA. FL-Lhr is more than twice as processive on flayed duplex DNA as compared to ssDNA and shows no activity on fully base paired duplex DNA. Graphed data was repeated twice and bars show standard error. (B) Example of 18% denaturing acrylamide TBE gel showing 80 nM of FL-Lhr glycosylase activity on flayed duplex d-uracil containing DNA. Time course reactions were completed over the course of 30 minutes.

C-terminal Lhr's ability to glycosylate a d-uracil containing flayed duplex was also investigated and is presented in Figure 8.10. In short, C-Lhr shows a 50% increase (15% total product formation) in glycosylase activity as compared to a ssDNA substrate (11% product formation). This increase is marginal when compared to FL-Lhr, further supporting enhanced activity as part of the full protein.

DNA substrates containing d-uracil within the duplex region contained a guanine nucleotide opposite the uracil 'deaminated' base to mimic any structural characteristics this may cause. Structural changes to the DNA because of damage and alternative base binding is often key for localisation and detection for proper repair protein function.

5.6.5 Determination of *E. coli* Lhr glycosylase activity in the presence of metal ions and ATP

Full length Lhr glycosylase activity in the presence of various metal ions and ATP was investigated. Previous work described the 'focusing' of Lhr's glycosylase activity upon the inclusion of manganese metal ions (Figure 5.16). Warren *et al.* described calcium as the most effective divalent cation cofactor for *E. coli* Lhr unwinding. This gives justification for CaCl₂ selection as opposed to other metal ions such as magnesium. EDTA is a strong metal ion chelator, here it is used as a 'no metal ion' control²³⁷. Repair pathways and helicase proteins often require energy released through ATP hydrolysis to perform function⁴⁸, here Lhr glycosylase activity in the absence of ATP is also investigated.

With reference to Figure 5.19 **A**, Lhr is able to perform glycosylase function in the absence of metal ions and ATP within the reaction buffer. The starkest difference in product formation is in the absence of ATP as shown on the right gel image. This may be due to the requirement of helicase activity to unwind the duplex region containing the d-uracil nucleotide allowing access.

Lhr is unable to perform glycosylase activity on a flayed duplex containing no d-uracil nucleotides. Mild nuclease activity is seen in the presence of Mn²⁺ ions (Figure 5.19 **B**, lanes **3** and **4**).

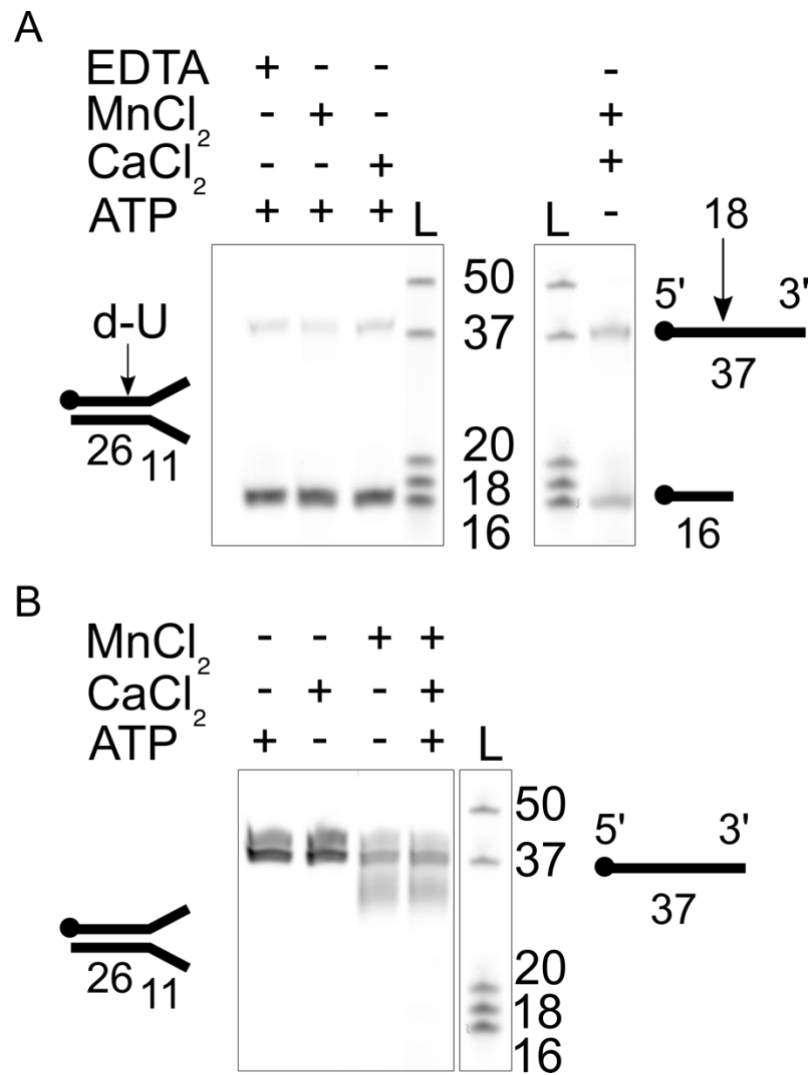


Figure 5.19 Analysis of *E. coli* full length Lhr glycosylase activity in the presence and absence of various metal coordination ions and ATP.

18% denaturing acrylamide TBE gels showing 80 nM FL-Lhr glycosylase activity on 12.5 nM of flayed duplex DNA substrates containing d-uracil (**A**) and 'undamaged' DNA (**B**). (**A**) FL-Lhr is able to elicit glycosylase function on d-Uracil flayed duplex DNA in all combinations. Activity is increased in the presence of ATP. (**B**) FL-Lhr shows no specific glycosylase activity on 'undamaged' DNA although some non-specific chewing is seen which is enhanced in the presence of Mn²⁺ ions.

5.6.6 Comparison of *E. coli* Lhr glycosylase activity to UDG and Fpg commercial DNA glycosylases

We sought to compare Lhr glycosylase activities with commercially available DNA glycosylases. This allows direct comparison of DNA products within the context of our assays. Both *E. coli* UDG (NEB M0280S) and Fpg (MutM, NEB M0240S) are quoted as DNA glycosylases with AP lyase activity, thus not requiring NaOH for backbone cleavage. To keep conditions consistent between reactions, NaOH was still added prior to loading to allow direct comparison of DNA species. Flayed duplex DNA substrates were used due to UDG and Fpg requiring 'damaged' nucleotides to be within the duplex region.

Lhr is shown to generate a DNA product of the same length on a d-uracil containing substrate as Fpg does with a substrate containing 8-oxoguanine. This is in contrast to the action of UDG which forms a DNA product of 20 nucleotides in length. Both FL and C-Lhr are unable to form a distinct glycosylase products on a 8-oxoguanine containing DNA substrate. This result supports the theory that Lhr is not directly involved in oxidative DNA damage or is involved in rare forms of oxidative damage which are not always present during genetic assays (see section 5.2.3).

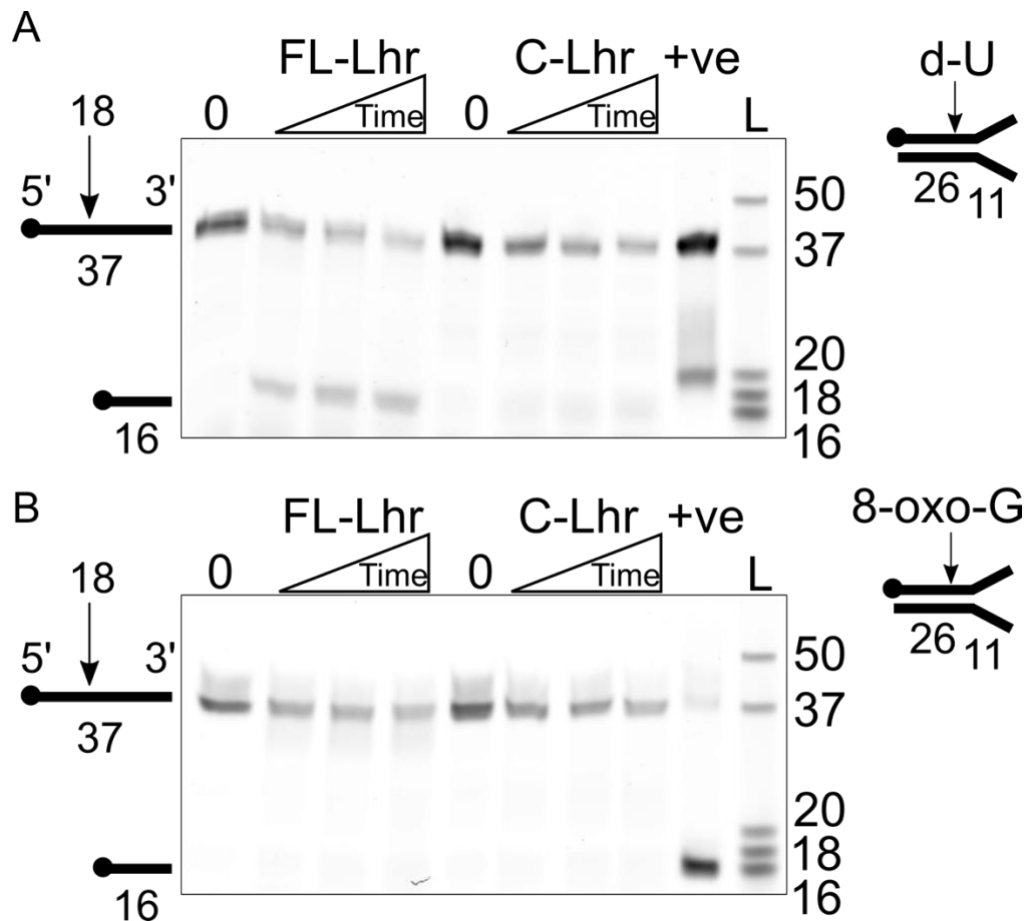


Figure 5.20 Analysis of *E. coli* full length Lhr glycosylase activity on d-uracil and 8-oxo-d-guanine containing DNA substrates as compared to commercial glycosylases. 18% denaturing acrylamide TBE gel showing 80 nM FL-Lhr glycosylase activity on flayed duplex DNA substrates containing d-uracil (A) and 8-oxo-d-Guanine (B). (A) Lhr proteins show different resulting products as compared to UDG (+ve), which excises the base 2 nucleotides upstream from the d-uracil base. (B) Lhr proteins are unable to elicit glycosylase activity on DNA substrates containing 8-oxo-d-guanine. Time points for Lhr proteins are 10, 20 and 30 minutes. Positive control lanes show resultant glycosylase activity after 30 minutes

Occurrence of d-uracil formation by cytosine deamination in the context of oxidative damage has been documented on a few occasions^{50–52,238}. This will be discussed in the context of Lhr in section 6.5 in the discussion.

5.7 Investigation of Lhr-CTD glycosylase active site residues

C-terminal Lhr was modelled using Phyre² analysis. This web-based service predicts a 3-D protein structure by aligning a submitted sequence with proteins which structures have already been solved. In addition to structure prediction, Phyre² details a list of closest structural matches allowing targeted inference of query protein function²³⁹.

Phyre² produces a '.pdb' file of the predicted structure which can be opened in molecular visualisation software such as PyMOL. This allows further in depth analysis where residue position can be observed and can also highlight areas of the protein where structural prediction was not possible.

5.7.1 Bioinformatic analysis of *E. coli* Lhr-CTD active site

We sought to identify essential catalytic residues responsible for observed glycosylase activity through structural prediction modelling. Phyre² analysis of C-terminal Lhr predicts a classic 'C-shape' topology reminiscent of *E. coli* AlkZ and *B. cereus* AlkD DNA glycosylases^{240,241}. Probable active site residues were highlighted through consideration of side chain extension into 'C-shape' pocket and side chain property as shown in Figure 5.21. Of the highlighted residues, a 'QxQ' motif was notably absent suggesting an alternative mechanism of glycosylase action to AlkZ, a protein which Lhr-CTD has been directly compared to^{110,125}. We hypothesised that amino acids of positive nature (arginine, R and lysine, L) would be responsible for DNA backbone contacts, hydrophobic residues (tryptophan, W) would serve to stabilise DNA base stacking and polar residues (glutamine, Q) would allow discrimination between cytosine and uracil⁵⁵. This allows attack of the glycosidic bond by the proposed catalytic aspartic acid residue (D 1536), facilitated by activation of a water molecule⁵³⁻⁵⁵. This mechanism would be similar to those described by other d-uracil glycosylases such as human UNG and SMUG1, as well as archaeal *MthMIG*^{48,242}.

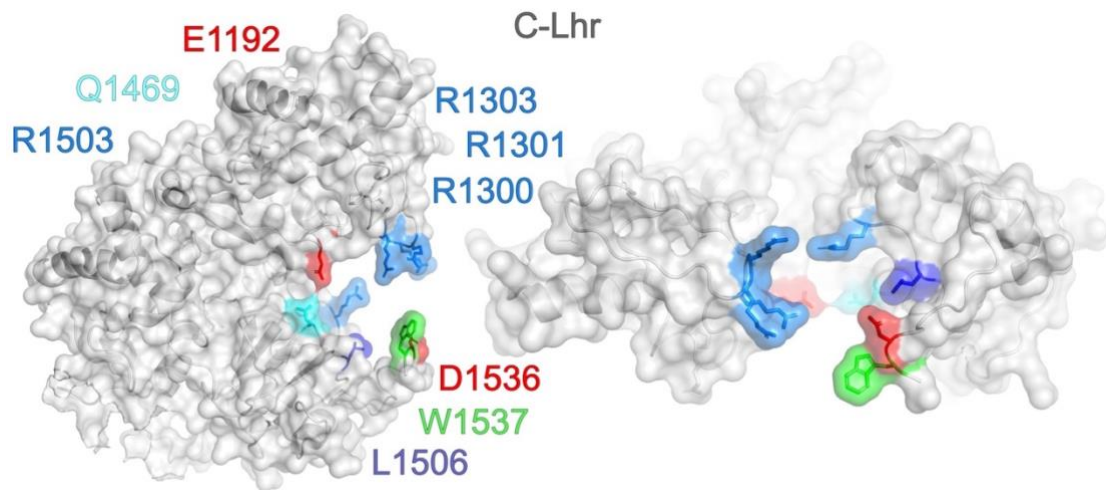


Figure 5.21 Phyre² structural model of *E. coli* Lhr-CTD with proposed active site residues as annotated.

C-terminal Lhr is shown in two profiles. Left depicts the characteristic ‘C-shape’ denoted by many other glycosylase enzymes, right shows the extension of highlighted residue side chains into the active site where they may form contacts with DNA.

5.7.2 Cloning and purification of *E. coli* Lhr CTD mutant D1536A

pRJB23 containing *E. coli* *lhr* residues 876-1538 was mutated using Gibson assembly which allows molecular cloning using sequence homology as opposed to using restriction endonuclease sites²⁴³. Great difficulty was found in attempting to mutate aspartic acid (D) residue 1536 to alanine (A) using a standard Q5 site directed mutagenesis protocol²⁴⁴. Here, Gibson assembly allowed efficient cloning by elevating trouble in optimising PCR conditions to amplify the 7.6 kb pRJB23 plasmid, which was thought to be the main source of difficulty. pRJB23 containing the D1536A mutation was named pRJB29 and was transformed into Rosetta 2 (DE3) cells for overexpression and purification. Overexpression was performed as stated with 'wild type' C-Lhr.

To purify, soluble cell lysate was loaded onto a Ni-NTA column to allow purification from contaminating proteins through selection of the histidine tag. Fractions were collected over an increasing gradient of imidazole for protein elution. C-Lhr D1536A containing fractions, as denoted in Figure 5.22, were dialysed into 'Low salt buffer A' and loaded onto a 1 ml Q sepharose column. *E. coli* C-Lhr was theorised to bind to a Q sepharose column due to its predicted charge of -16.3 in pH 8 (pI of 5.66), allowing strong binding to the positive beads of the column.

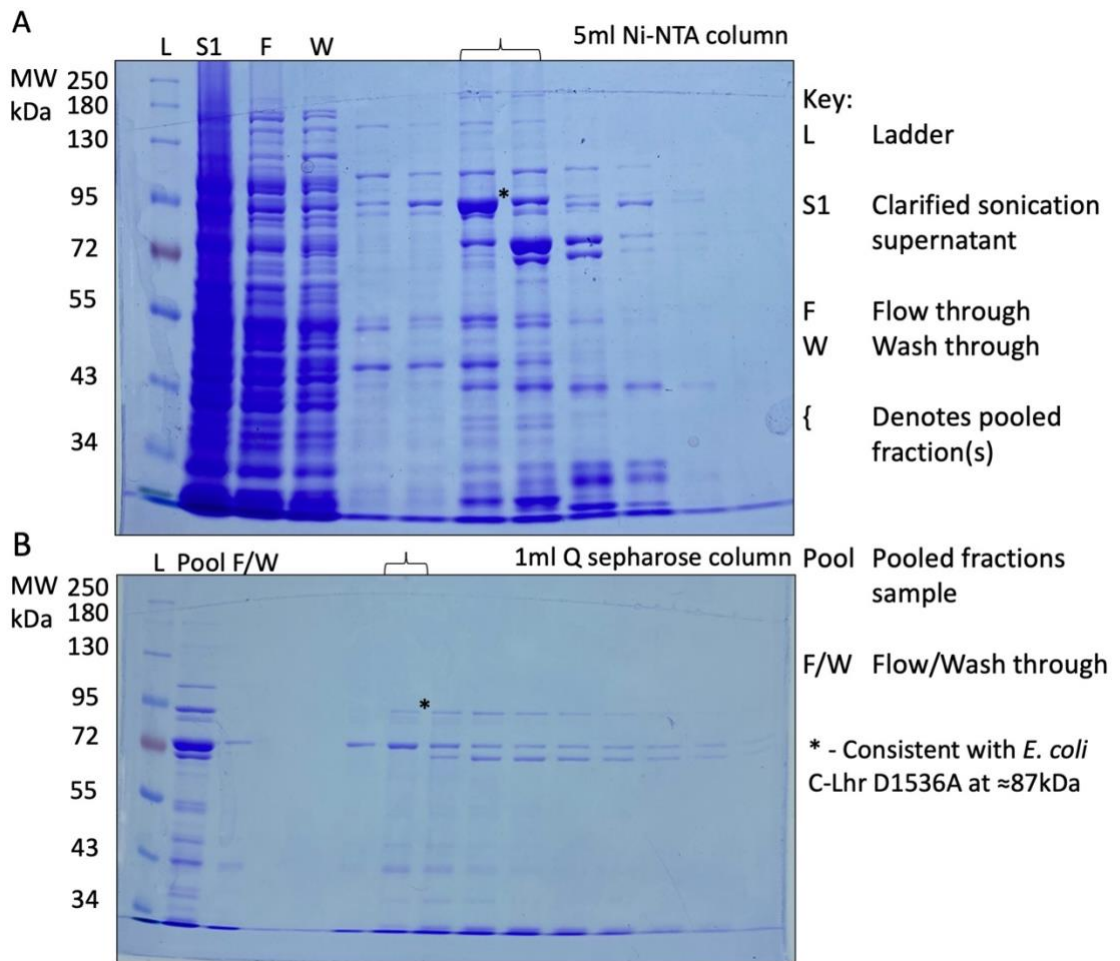


Figure 5.22 Purification of *E. coli* Lhr CTD D1536A mutant.

Coomassie stained 10% acrylamide SDS PAGE analysis of C-terminal Lhr D1536A mutant purification steps. **(A)** Ni²⁺-NTA chromatography. Pooled fractions were dialysed for loading onto a Q sepharose column. **(B)** Q sepharose chromatography for successful purification. Bands of interest are highlighted with an ‘*’ as indicated.

C-terminal Lhr D1536A mutant eluted early from the Q sepharose column. The highlighted protein fraction was dialysed for storage and biochemical analysis as indicated. Additional protein bands are thought to be degradation products conferring similar sizes to those seen during full length Lhr purification (Figure 5.10).

Approximate protein concentration was determined using the DeNovix spectrophotometer absorption reading at 280 nm and Lhr’s extinction coefficient value (178105), which were applied to the Beer-Lambert law as before.

5.7.3 Cloning and purification of *E. coli* full length Lhr mutant D1536A

pRJB28 containing His-tagged *E. coli lhr* was mutated using Gibson assembly. This was performed for similar reasons as stated in section 5.7.2. pRJB28 containing the D1536A mutation was named pRJB32 and was transformed into Rosetta 2 (DE3) cells for overexpression and purification. Overexpression was performed as with wild type full length Lhr.

Purification steps were performed as stated in section 5.3.3. Suspected degradation products are consistent between all purifications involving Lhr with Ni²⁺-NTA and Q sepharose chromatography steps.

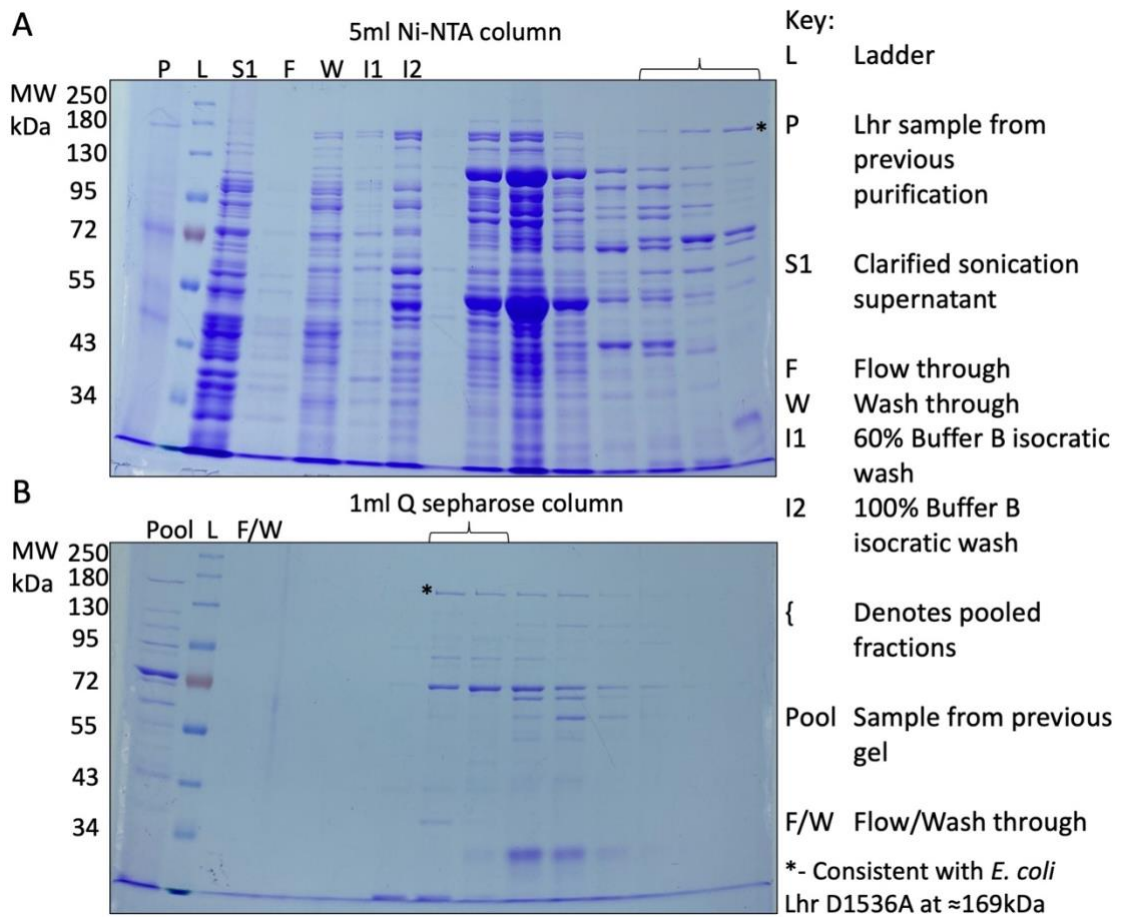


Figure 5.23 Purification of *E. coli* Lhr D1536A mutant.

Coomassie stained 8% acrylamide SDS PAGE analysis of Lhr D1536A mutant purification steps. **(A)** Ni²⁺-NTA chromatography. Pooled fractions were dialysed for loading onto a Q sepharose column. **(B)** Q sepharose chromatography for successful purification. Bands of interest are highlighted with an ‘*’ as indicated.

Approximate protein concentration was determined using the DeNovix spectrophotometer absorption reading at 280 nm and Lhr’s extinction coefficient value (178105), which were applied to the Beer-Lambert law as before.

5.8 Investigation into glycosylase active site residues

Purified D1536A mutant proteins were subject to biochemical interrogation with activities compared to 'wild type' Lhr proteins. Initial testing was performed through glycosylase assays which were run on denaturing polyacrylamide gels for visualisation. This was followed by DNA binding and unwinding assays to check for a false positive glycosylase inactive mutation.

5.8.1 D1536A mutation causes loss of glycosylase function in both full length and CTD Lhr

With reference to Figure 5.24, investigation into the effect of the D1536A mutation showed loss of glycosylase function in both FL-Lhr and C-Lhr protein species. This was confirmed as part of a concentration titration (A) and as a function of time (B).

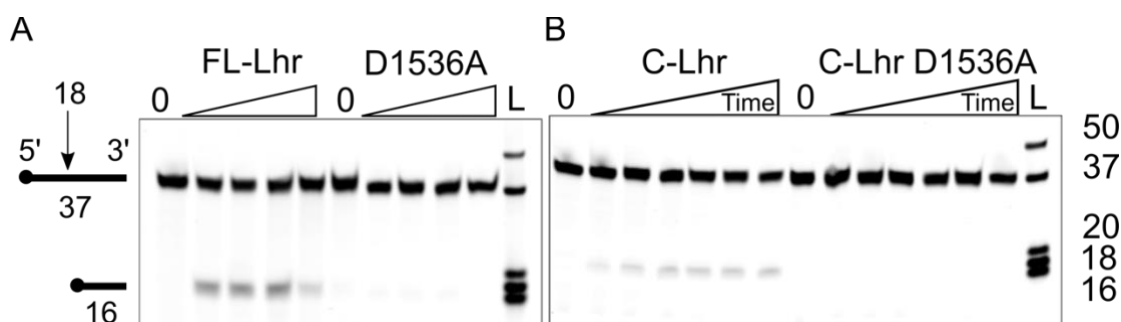


Figure 5.24 Effect of D1536A mutation on *E. coli* Lhr glycosylase activity.

Identification of catalytic aspartic acid residue through loss of function in both full length Lhr (A) and C-terminal Lhr (B). (A) Loss of glycosylase activity is shown in a concentration titration of full length Lhr (25, 50, 100 and 200 nM of protein) and as a function of time in C-terminal Lhr (200 nM) over the course of 30 minutes (B). 12.5 nM 5' Cy5 labelled ssDNA was used as substrate.

5.8.2 FL-Lhr D1536A binds and unwinds DNA

Upon confirming the D1536A mutation inactivates d-uracil glycosylase activity in both full length and C-terminal Lhr, we sought to investigate if this mutation had in fact caused wider structural effects leading to a reduction in DNA binding and unwinding abilities. This investigation would eliminate D1536A as a false positive glycosylase inactive mutant. To achieve this we performed EMSAs alongside full length wild type Lhr and compared unwinding abilities. Here, unwinding assays were performed using 5' Cy5-labelled flayed duplex DNA substrates. Reactions were run on 10% native polyacrylamide gels to allow separation of DNA species by molecular weight. Accumulation of single stranded DNA unwinding product was visualised using a Typhoon phosphor-imager.

Comparison of full length wild type Lhr and D1536A mutant shows negligible differences between DNA binding (Figure 5.25 **A**) or in ability to unwind DNA (Figure 5.25 **B**). Formation of a tight protein-nucleic acid complex band is absent in the D1536A mutant as compared to wild type protein.

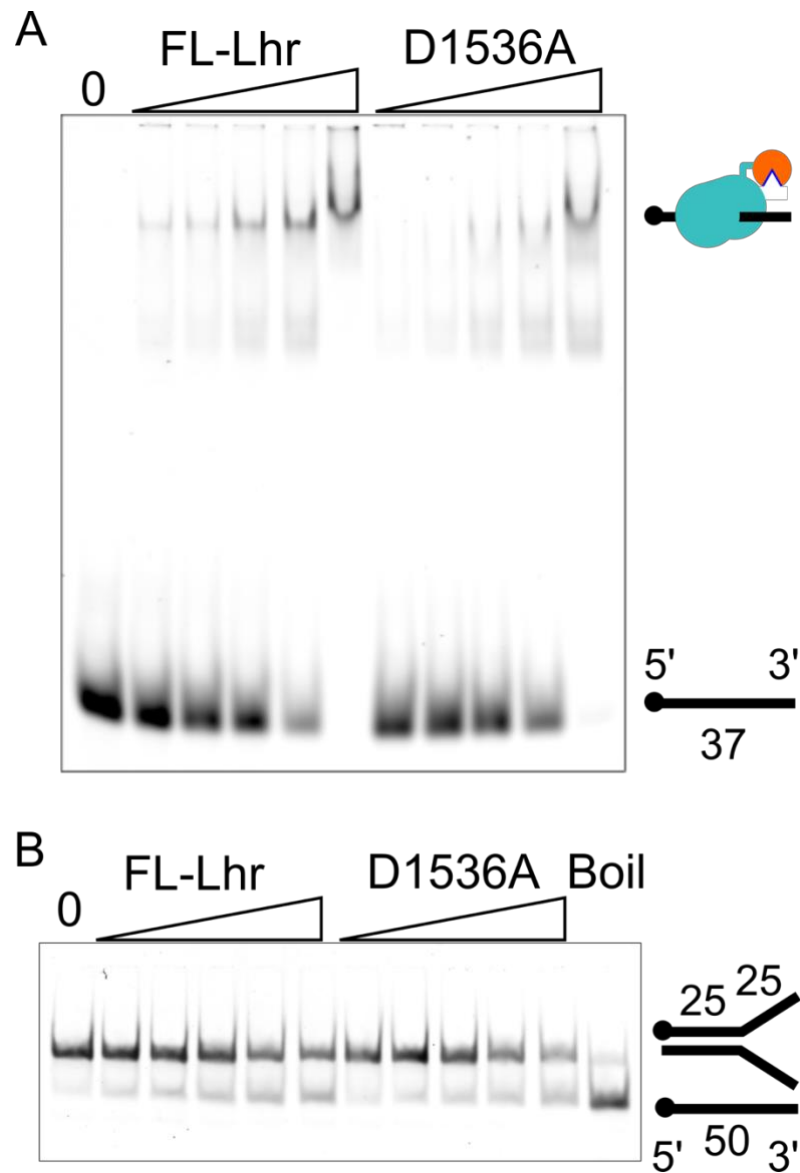


Figure 5.25 *E. coli* Lhr DNA binding and unwinding activity is unaffected by a D1536A mutation.

Investigation of full length Lhr and D1536A mutant DNA binding and unwinding capabilities. Protein concentration increases from 12.5 to 200 nM linearly. (A) 5% Native acrylamide EMSA gel showing DNA binding ability of Lhr proteins on 12.5 nM of 5' Cy5 labelled ssDNA substrate. (B) 10% Native acrylamide gel showing DNA unwinding ability of Lhr proteins on 12.5 nM of 5' Cy5 labelled flayed duplex DNA.

5.9 Summary of key findings

5.9.1 Lhr is involved in suppressing replicative stress and in oxidative damage repair

E. coli genetic knockout studies reconfirmed a synergistic repair phenotype as reported by Cooper *et al.* (2015). *Lhr* and *radA* deficient cells showed increased sensitivity when grown in the presence of AZT which causes accumulation of ssDNA gaps due to replication chain termination (Figure 5.3). An additional repair phenotype was displayed in the presence of H₂O₂ showing a reduction in cellular viability and a growth lag for entering the exponential growth phase (Figure 5.4). Here, *E. coli* *Lhr* knockout cells showed limited sensitivity to ICL damage repair when grown in the presence of mitomycin C. However, some sensitivity was shown in all $\Delta radA$ cell strains.

5.9.2 Lhr and RadA (Sms) may be part of a mutagenic repair pathway

Accumulation of random mutations in the absence of *Lhr* and *RadA* (*Sms*) proteins was investigated through use of a 'acquired rifampicin resistance' assay. Data displayed in Figure 5.6 show a reduction of resistant colony formation within all knockout cell strains used. Particular interest was seen within the *lhr radA* dual knockout cell strain which showed a drastic reduction in acquired resistance suggesting potential involvement in a mutagenic repair pathway. This phenomenon has only been reported with overexpression of MMR pathway proteins and is exceedingly rare. Normal growth without rifampicin was shown to be unaffected.

5.9.3 Lhr-core is required for Lhr binding and loading onto exposed ssDNA

Lhr is able to bind DNA substrates which have ssDNA regions which may facilitate protein loading (Figure 5.12). Protein-nucleic acid aggregation is seen in DNA substrates which contain duplex regions. Figure 5.13 shows negligible difference between Lhr binding to a 'damaged' ssDNA substrate containing a d-uracil nucleotide and an equivalent 'undamaged' DNA oligo. Investigation into purified Lhr-CTD showed an inability to stably bind DNA as detectable by an ESMA (Figure 5.14). This suggests a requirement of Lhr-core for proper protein function.

5.9.4 Lhr displays d-uracil DNA glycosylase activity

Initial suspected glycosylase activity was investigated using Lhr-CTD on DNA substrates containing d-uracil and 8-oxo-d-guanine nucleotides. Lhr-CTD or 'C-Lhr' showed glycosylase activity only in the presence of d-uracil. This was justified in Figure 5.15, lower gel, due to the proper DNA product migration only occurring in the presence of NaOH, which cleaves the AP site generated by Lhr-CTD through β/δ elimination. This also suggests Lhr does not display AP lyase activity.

Further characterisation using the extended full length *E. coli* Lhr protein showed a honing of glycosylase activity in the presence of Mn^{2+} ions (Figure 5.16) through potential uncoupling of the helicase and glycosylase domains.

With reference to Figure 5.17 and Figure 5.18, FL-Lhr is up to 4x more processive than Lhr-CTD alone and shows a strong preference to d-uracil containing forked DNA substrates. Both Lhr proteins were unable to produce glycosylase products on d-uracil containing duplex DNA.

Figure 5.19 showed FL-Lhr glycosylase activity is specific to d-uracil containing DNA and action occurs in the absence of metal coordination ions but activity is increased in the presence of ATP.

Lhr demonstrated an alternative mechanism of product formation as compared to *E. coli* uracil glycosylase UNG, displayed as DNA products of different length (Figure 5.20, **A**). Comparison of Lhr to *E. coli* Fpg (MutM), an 8-oxoguanine DNA glycosylase, showed similar product formation but in different 'damaged' substrates. Here, Lhr is shown to be unable to elicit glycosylase activity on 8-oxo-d-guanine containing DNA (Figure 5.20, **B**).

5.9.5 Identification of a key aspartic acid residue for d-uracil glycosylase activity

Lhr-CTD catalytic active site was determined using Phyre² structural modelling. Residue D1536 was targeted for site directed mutagenesis as the proposed catalytic amino acid. This was theorised through deduction analysis of residues which displayed side chains that extended into the presented active site as labelled in Figure 5.21.

Subsequent purification and glycosylase analysis of D1536A full length and CTD-Lhr proteins showed abolishment of glycosylase activity. This was presented as a function of protein concentration and as a function of time (Figure 5.24).

The D1536A mutation was also shown to have little to know effect on protein DNA binding and unwinding as compared to wild type full length protein (Figure 5.25). This data suggests targeted removal of glycosylase activity without affecting Lhr-cores functionality.

Chapter 6 :Discussion and future research

6.1 Assessment of project aims

6.1.1 RecA/Rad51 family proteins

The initial aim for this project was to identify *Trypanosoma brucei* Rad51 paralogue proteins and investigate them structurally using protein modelling. This would allow identification of human homologue partners through structural comparison, which would then be tested biochemically after protein purification. A great deal of time was spent in optimisation of protein overexpression and purification with limited success (data not shown). This was due to the insolubility and toxicity of expressing recombinant *T. brucei* proteins within *Escherichia coli* expression systems.

Due to the closure of research labs during the Covid19 lockdown in 2020, time was freed to allow in depth phylogenetic analysis to ascertain *T. brucei* Rad51 paralogue proteins' suitability as a model for the human homologues. This data, as presented in Chapter 1, highlighted the stark differences in Rad51 paralogue proteins between *Homo sapiens* and *T brucei*. Further investigation revealed an interesting relationship between RecA/Rad51 family proteins which may allow division into three distinct protein subfamilies.

Identification of *T. brucei's* limited suitability as a Rad51 paralogue protein model, the great difficulty seen in obtaining purified proteins and the loss of time due to the Covid19 pandemic led to the abandoning of this project and a change of study into Lhr family proteins. This would build upon work produced during my MRes and represented a more streamlined project with outcomes more easily obtainable within

the time remaining in my PhD. This project would still allow investigation into proteins thought to be involved within the early stages of homologous recombination and would build upon knowledge obtained pre-pandemic.

6.1.2 Lhr family proteins

The revised aims of this study were to (a) investigate Lhr protein abundance across archaea and bacteria with consideration to genomic context, (b) characterise *Methanothermobacter thermautotrophicus* 'Lhr-core' unwinding preferences building upon previous MRes work, (c) develop *E. coli* genetic analysis of Lhr and RadA (Sms) proteins and (d) investigate Lhr's extended C-terminus of unknown function.

Here we present data which highlights the vast abundance of Lhr family proteins in both bacteria and archaea (chapters 3 and 4). We shed light onto the quiriness of *E. coli lhr* gene regulation through two possible promoters and rare codon usage (chapter 3). Characterise an example of 'Lhr-core' describing substrate preference and ability to remodel branched DNA to allow protein loading, and subsequent unwinding through the parental duplex (chapter 4). Describe additional repair phenotypes of *lhr* knockout cells through sensitivities to oxidative damage and suggest a potential involvement in a mutagenic repair pathway along with RadA (Sms) (chapter 5). We also present the first characterisation of Lhr's extended C-terminal domain which displays d-uracil targeted DNA glycosylase activity and identify the key catalytic residue through targeted mutagenesis (chapter 5).

This data represents a significant breakthrough in the study of Lhr family proteins through characterisation of *E. coli* Lhr's C-terminal domain. This will allow further

study into Lhr family proteins to show conservation or diversification of glycosylase activity between protein examples. Additionally, data here gives clues into Lhr's involvement in maintaining genomic stability by relieving replicative stress and gives further biological context in oxidative stress repair and adaptive mutation with a protein partner. Here I will discuss these wider implications describing any experimental limitations and additional work needed.

6.2 Evaluation of RecA/Rad51 protein phylogenetics

RecA/Rad51 family proteins are found extensively in eukaryotes with variable numbers of non-recombinogenic Rad51 paralogue proteins being present between organisms. Rad51 paralogue abundance, or lack thereof, may be a direct indication of the most common DNA metabolism processes occurring. This may present as an organisms continued exposure to mutagenic sources, dependent on cellular environment, requiring utilisation of specific repair pathways or, the complexity of an organisms cell cycle requiring genetic variability through HR processes.

This became particularly evident when evaluating *T. brucei* Rad51 paralogue proteins as shown in chapter 3. In Figure 3.3, we identified a close evolutionary relationship between a *T. brucei* paralogue protein termed as 'p6' and bacterial *E. coli* RecA. Upon further investigation of the literature, this protein has been shown to be involved in anti-genic variation in *T. brucei*^{245,246}. This process involves the altering of expression of variant surface glycoproteins (VSG) by replacement of a previously silent VSG gene into the actively transcribed 'expression site', directed through HR. VSG switching is reminiscent to anti-genic variation strategies deployed by other pathogenic organisms for immune evasion^{247,248}. Phylogenetic analysis through conservation of active site residues as performed here may allow the identification of previously unknown Rad51 paralogue proteins within pathogenic organisms which use this process. These additional proteins may then be investigated to determine the evolutionary origin of anti-genic variation as either through a common ancestor or as convergent evolution during the 'arms race' between pathogen and host. These proteins may also represent

a target for inhibition to reduce HR associated pathways to help combat infection and limit disease posed by this pathogens.

Rad51 paralogue proteins are markedly absent within bacteria with only a few instances being reported. *E. coli* RadA (Sms) represents the most well studied bacterial Rad51 paralogue with recent study showing structural and functional conservation in *Arabidopsis thaliana* and presence in other land plants and also in algae²⁴⁹. Data presented in Figure 3.4, suggests an interesting relationship between RadA (Sms) and *H. sapiens* Rad51 C. To my knowledge, direct comparison between Rad51 paralogue proteins in bacteria and eukaryotes has not been reported. Rad51 C represents a key Rad51 paralogue protein in humans being part of both paralogue complexes and having additional roles within mitochondrial recombination²⁵⁰. The potential subdivision into recombinogenic, XRCC3-like and Rad51 C-like proteins as presented in Figure 3.4 may allow identification and targeted study of other Rad51 paralogue proteins using a new angle of approach. RadA (Sms)'s likeness in being a Rad51 C-like protein is further supported in Chevigny *et al.* where they report its localisation in *A. thaliana*'s mitochondria and its observed branch-migration functionality. This could represent a direct link to the Rad51 C/XRCC3 complex in humans but further study of both proteins would be required^{249,250}.

Phylogenetic analysis was performed with an eclectic sample of RecA/Rad51 family proteins and did not include a vast sample size. This may cause limitation in how reliable the evolutionary relationships are depicted. However, when further iterations were performed using a wider range of proteins, MUSCLE and Gblocks began to identify additional structural elements outside of the catalytic domains which may

have hid useful insight and produced bias in a different way. Here lies the delicate balance of sequence based techniques to help ascertain protein functionality.

6.3 Lhr substrate preference

Lhr has widely been reported as an ATP-dependent 3' to 5' DNA translocase which requires regions of ssDNA to facilitate loading and subsequent unwinding. Between studied examples, Lhr exhibits the ability to unwind both DNA:DNA and DNA:RNA substrates when the protein translocates along the DNA containing strand^{120,125,134}. Study into archaeal Lhr has described additional RNA:RNA substrate unwinding and strand annealing functionality^{121,131}. The diversity in substrate preference may be owed to the functional diversity seen between Lhr family proteins¹²¹.

Studies of Lhr from *Mycobacterium* performed by Stewart Shumans' lab primarily focused on characterisation of protein translocation/unwinding polarity, ATPase activity in the presence of single stranded nucleic acid, divalent cation specificities and unwinding ability on 3'-partial duplex DNA. Here we present an in depth analysis of Lhr substrate preference across a variety of partial duplex and branched DNA substrates (Chapter 4). *M. thermautotrophicus* Lhr-core is consistent in translocation polarity shown through displacement of a 3'-partial duplex DNA substrate (Figure 4.2). Investigation into *MthLhr*'s role in HR through action on Holliday junction DNA displayed differing branch migration capabilities to the *E. coli* HJ resolvase RuvAB. Data as seen in Figure 4.3 and Figure 4.4, suggest branched DNA targeting as opposed to *bone fide* branch migration functionality¹¹⁰. This is in relative agreement to data presented by Suzuki *et al.* who include a series of 'trap' DNA oligonucleotides to

capture unwound HJ junction intermediates produced by *SacaLhr*¹³¹. Here, *SacaLhr* displays reminiscent full unwinding of the HJ substrate with intermediates as with *MthLhr* (Figure 4.3) and prompts comparison to Hjm/Hel308 unwinding capabilities with suggested involvement on HR. Involvement, but not direct reliance on Lhr's HJ processing functionality is displayed through various non-lethal genetic phenotypes, such as the limited sensitivity to MMC reported here (Figure 5.5), so exact function within this context still remains elusive.

Further investigation into *MthLhr*'s action on branched DNA substrates demonstrated a strong unwinding preference to fully base paired 'replication fork' DNA substrates, as shown in Figure 4.4. This is in direct agreement to genetic analysis presented by Cooper and Rand *et al.* and in our initial assay, which highlighted *MthLhr*'s localisation and interference of *E. coli* stalled replication forks (Figure 4.1)^{1,110,130}. Further support for Lhr's involvement in replication-coupled repair is displayed by *S. acidocaldarius* Lhr, which more readily unwound flayed duplex DNA as opposed to HJ, partial-duplex and duplex DNA¹³².

MthLhr's action on forked DNA substrates was further characterised as shown in Figure 4.5, showing binding and unwinding through the parental duplex, giving further context to its action in replication-coupled repair. Analysis through smFRET showed Lhr's capability to locally distort and remodel fully base paired DNA to allow loading and unwinding in the presence of Mg²⁺ and ATP¹¹⁰. This subsequently bypasses Lhr's requirement for instances of ssDNA to facilitate loading, as is displayed with other characterised examples.

A strong preference to replication fork, flayed duplex DNA substrates was displayed in *EcoLhr*'s glycosylase activity. As shown in Figure 5.18, *EcoLhr* is able to produce over twice the amount of glycosylated product on flayed duplex DNA as compared to an equivalent ssDNA substrate. An increased processivity occurs with seemingly similar binding efficiencies to both substrates (Figure 5.12). *EcoLhr* in our hands also displayed reasonable unwinding of a DNA:DNA substrate which is in slight disagreement to data presented in Warren *et al.*, although full investigation using the same substrates would be needed to ascertain any functional differences between protein preparations¹²⁵.

6.4 Lhr as a dual function protein

Lhr family proteins are highly abundant proteins found extensively in bacteria, as shown here in Figure 3.9, and in archaea as reported by Hajj *et al.* Previous studies of Lhr identified an uncharacterised extended C-terminus of bacterial Lhr which, at the time, displayed no homology to any other proteins outside of the Lhr clade^{119,120}. In Chapter 4 Figure 4.7, we present an in depth computational analysis of *MsmLhr*-CTD revealing structural similarity to SelB (an mRNA elongation factor) and AlkZ (a DNA glycosylase)¹¹⁰. Similarity to AlkZ was later confirmed in Warren *et al.* who presented the solved cryo-EM structure, depicting extensive structural homology through the β -barrel and extended multiple winged-helix motifs¹²⁵. AlkZ is a representative of an emerging 'HTH_42' superfamily of DNA repair proteins thought to be involved in a wide range of DNA repair pathways⁵⁷.

Investigation into the proposed AlkZ-like *E. coli* Lhr-CTD, as shown in Chapter 5, displayed the activity of a d-uracil stimulated DNA glycosylase. This activity was consistent in both purified full length *EcoLhr* and the isolated C-terminal domain (Figure 5.17). Full length *EcoLhr* was able to produce 4x more glycosylated product than the CTD alone. This may be due to lack of observable stable DNA binding displayed by *EcoLhr*-CTD. This data suggests a tethering of the helicase/DNA binding and DNA glycosylase activities similar to the cooperativity of the UvrB helicase and UvrC excinuclease proteins in the *E. coli* NER pathway^{46,72}. Further investigation into *EcoLhr*-CTD's DNA binding ability may shed light into the extent of DNA association through this domain and may be achieved using more sensitive methods such as anisotropy.

Full length *EcoLhr* showed a strong substrate preference to a flayed duplex containing d-uracil substrate, as mentioned before (Figure 5.18), this further strengthens its link to replication-coupled repair. Glycosylase activity was further investigated on 8-oxo-d-guanine containing DNA but displayed no discernible activity. This is unexpected due to the observed H₂O₂ sensitivity phenotype as report here in Figure 5.4.

One notable difference between Lhr-CTD and AlkZ is the stark absence of a catalytic 'QxQ' motif as described here in section 5.7.1 and highlighted in Bradley *et al*⁵⁷. Through comparison to other uracil DNA glycosylases, we identified an aspartic acid as the potential catalytic residue and investigated through site directed mutagenesis. Data presented in Figure 5.24 and Figure 5.25 shows complete inhibition of glycosylase activity in both FL-Lhr and CTD D1536A mutant proteins without affecting DNA binding or unwinding activities. This suggests a catalytic mechanism through

nucleophilic attack by an activated water molecule by the aspartic acid residue similar to that of other uracil DNA glycosylases^{48,53–55,242}. Further investigation into the proposed active site as presented in Figure 5.21 would shed more light onto the intricacies of mechanistic action. Of particular interest to me is the glutamic acid residue 1192 which resides deep within the binding pocket. As presented here, Lhr glycosylase activity results in the formation of an AP site two nucleotides up-stream of the d-uracil. Does residue 1192 also cause glycosidic bond breakage closer to the d-uracil or does it elicit a different function? An answer to this may be achieved through 3'-Cy5 or dual labelling to show the resultant product upstream from the d-uracil base. Further experimentation may also illuminate the dependence of uracil base, whether on the tracking or displayed strand, or for any preferences to DNA:RNA hybrids. Other 'damaged' base preference may also be investigated to fully determine Lhr's glycosylase specificity.

6.5 Lhr in a biological context

Lhr's role within a cell has been difficult to pin down. Examples in both bacteria and archaea display variable genetic phenotypes and expression responses when exposed to DNA damaging agents. This may further promote the idea of a diverse family of proteins with multiple functions and presence in multiple repair pathways, or roles in specific repair dependent on each organism. One thing that is certain is Lhr's association with DNA repair in both HR and at sites of replicative stress.

Genetic data presented here in section 5.2 confirms *EcoLhr*'s involvement in replicative stress associated with ssDNA gaps along with *EcoRadA*, displays a new

genetic repair phenotype in both growth rate and viability when exposed to a source of oxidative damage, and describes a very limited sensitivity to ICL's when grown in the presence of MMC. Generating genetic phenotypes of *E. coli* knockout cells have had limited fruition and great difficulty was seen in generating the data presented here^{1,119}. This may be due to the regulation control posed on *lhr* through a heat shock promoter or the percentage of rare codons seen throughout the gene, limiting expression to very specific circumstances (presented in section 3.2.3). These limitations have made it difficult to produce complementation assays to the phenotypes presented here but this is something that is needed to reaffirm the observed results. Furthering *EcoLhr*'s proposed roles in relieving replicative stress and in oxidative damage repair is the apparent involvement in a mutagenic repair pathway along with *EcoRadA* as displayed in Figure 5.6, through a reduction in spontaneous resistance to rifampicin. This phenotype is extremely rare and to my knowledge has only been displayed on two occasions involving altered expression of MMR proteins^{228,229}.

Piecing together these genetic observations and biochemical analysis of *Mth* and *EcoLhr* suggest a strong preference to the repair of DNA damage within the context of DNA replication. This may be as a role to unwind and excise inhibitory DNA lesions ahead of the replisome through DNA remodelling, which can be investigated by displacement of DNA road blocks and unwinding capabilities on chemically modified DNA^{114,156,251,252}. *EcoLhr*'s involvement in oxidative damage repair but inability to target 8-oxoguanine containing DNA substrates, to which is the main form of damage, appears counterintuitive. It has been reported that oxidative damage may indeed be a source of cytosine deamination to uracil (or uracil derivatives) but it is difficult to say

whether these are present during the genetic assays reported here, or whether Lhr would perform a direct or indirect role in repair^{50–52,238,253,254}. Further investigation into Lhr's glycosylase substrates is therefore needed. One possible way to justify *EcoLhr*'s involvement in oxidative damage repair is through consideration of its possible expression control by RpoE. Whilst 8-oxo-guanine may not be a substrate for *EcoLhr* its deletion may cause a wider disruption to the regulatory network coordinated by RpoE/ σ^{24} leading to an increase in sensitivity.

The observed loss of spontaneous mutation may be justified through the potential accumulation of endogenous cytosine deamination during normal cell growth and in entering stationary phase. Accumulation of d-uracil may be targeted by Lhr and RadA (Sms) which removes the adjacent bases, damage which is then repaired erroneously by downstream proteins. Further analysis of the effect of the *lhr* and *radA* (*sms*) knockouts on cell growth may also yield an explanation for the reduced observable mutation rate.

Further insight into *E. coli* Lhr's biological role may be afforded through investigation into its relationship with RNaseT. As presented in Figure 3.6, RNaseT and Lhr are located within the same operon. Here, we began to investigate functional interactions with Lhr glycosylase activity and RNaseT action on DNA but ran out of time for full characterisation (Figure 8.13). To fully appreciate RNaseT and Lhr association additional pull-down studies may be performed, although it must be noted that operonic proteins may not always confer cooperative functions. Investigation into Lhr's interactions with both RNaseT and RadA (Sms) may begin to build a larger repair pathway and start to link observed biochemical data to cellular function(s).

Chapter 7 References

1. Cooper, D. L., Boyle, D. C. & Lovett, S. T. Genetic analysis of Escherichia coli RadA: Functional motifs and genetic interactions. *Mol Microbiol* **95**, 769–779 (2015).
2. Alberts, B. *et al.* The Structure and Function of DNA. (2002).
3. Pagès, V. & Fuchs, R. P. How DNA lesions are turned into mutations within cells? *Oncogene* **21**, 8957–8966 (2002).
4. Mu, D. *et al.* DNA interstrand cross-links induce futile repair synthesis in mammalian cell extracts. *Mol Cell Biol* **20**, 2446–2454 (2000).
5. Wang, J. Y. J. DNA damage and apoptosis. *Cell Death & Differentiation* **2001** 8:11 **8**, 1047–1048 (2001).
6. Prorok, P. *et al.* Evolutionary Origins of DNA Repair Pathways: Role of Oxygen Catastrophe in the Emergence of DNA Glycosylases. *Cells* **10**, (2021).
7. Jiang, S., Lin, T., Xie, Q. & Wang, L. Network Analysis of RAD51 Proteins in Metazoa and the Evolutionary Relationships With Their Archaeal Homologs. *Front Genet* **9**, 1–9 (2018).
8. Rocha, E. P. C., Cornet, E. & Michel, B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genetics* vol. 1 0247–0259 Preprint at <https://doi.org/10.1371/journal.pgen.0010015> (2005).
9. Herschlag, D. & Pinney, M. M. Hydrogen Bonds: Simple after All? Special Issue: Current Topics in Mechanistic Enzymology. *Biochemistry* **57**, (2018).

10. Maffeo, C. *et al.* Close encounters with DNA. *J Phys Condens Matter* **26**, 413101 (2014).
11. Brosh, R. M. & Matson, S. W. History of DNA Helicases. *Genes* **2020**, Vol. *11*, Page 255 **11**, 255 (2020).
12. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **1953** *171*:4356 **171**, 737–738 (1953).
13. Gidron, Y., Russ, K., Tissarchondou, H. & Warner, J. The relation between psychological factors and DNA-damage: A critical review. *Biol Psychol* **72**, 291–304 (2006).
14. Bauer, N. C., Corbett, A. H. & Doetsch, P. W. The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res* **43**, 10083–10101 (2015).
15. Compound Interest. The Chemical Structure of DNA – Compound Interest. <https://www.compoundchem.com/2015/03/24/dna/> (2015).
16. Burnside, K. & Rajagopal, L. Regulation of prokaryotic gene expression by eukaryotic-like enzymes. *Curr Opin Microbiol* **15**, 125 (2012).
17. Foster, P. L. Stress-Induced Mutagenesis in Bacteria. doi:10.1080/10409230701648494.
18. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair and mutagenesis. *Environ Mol Mutagen* **58**, 235 (2017).
19. Foster, P. L. Stress-Induced Mutagenesis in Bacteria. *Crit Rev Biochem Mol Biol* **42**, 373 (2007).

20. O'Donnell, M. *et al.* Principles and Concepts of DNA Replication in Bacteria, Archea, and Eukarya. *Cold Spring Harb Perspect Biol* **5**, a010180 (2013).
21. Cooper, G. The Eukaryotic Cell Cycle. in *The Cell: A Molecular Approach. 2nd edition.* (Sinauer Associates, 2000).
22. Meselson, M. & Stahl, F. W. The replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences* **44**, 671–682 (1958).
23. Nachtomy, O., Shavit, A. & Yakhini, Z. Gene expression and the concept of the phenotype. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences* **38**, 238–254 (2007).
24. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science (1979)* **355**, 1330–1334 (2017).
25. Kelley, M. R., Logsdon, D. & Fishel, M. L. Targeting DNA repair pathways for cancer treatment: what's new? *Future Oncology* **10**, 1215–1237 (2014).
26. Langston, L. D. & O'Donnell, M. DNA Replication: Keep Moving and Don't Mind the Gap. *Mol Cell* **23**, 155–160 (2006).
27. Morin, J. A. *et al.* Mechano-chemical kinetics of DNA replication: identification of the translocation step of a replicative DNA polymerase. *Nucleic Acids Res* **43**, 3643–3652 (2015).
28. Yao, N. Y. & O'Donnell, M. The Replisome. *Cell* **141**, 1088-1088.e1 (2010).

29. Marians, K. J. Understanding how the replisome works. *Nat Struct Mol Biol* **15**, 125–127 (2008).
30. Zeman, M. K. & Cimprich, K. A. Causes and Consequences of Replication Stress. *Nat Cell Biol* **16**, 2 (2014).
31. Bruning, J. G., Howard, J. L. & McGlynn, P. Accessory replicative helicases and the replication of protein-bound DNA. *J Mol Biol* **426**, 3917–3928 (2015).
32. Edenberg, E. R., Downey, M. & Toczyski, D. Polymerase Stalling during Replication, Transcription and Translation. *Current Biology* **24**, R445–R452 (2014).
33. Yeeles, J. T. P., Poli, J., Marians, K. J. & Pasero, P. Rescuing stalled or damaged replication forks. *Cold Spring Harbor perspectives in biology* vol. 5 a012815 Preprint at <https://doi.org/10.1101/cshperspect.a012815> (2013).
34. Alexander, J. L. & Orr-Weaver, T. L. Replication fork instability and the consequences of fork collisions from rereplication. *Genes Dev* **30**, 2241 (2016).
35. Rudolph, C. J., Upton, A. L., Briggs, G. S. & Lloyd, R. G. Is RecG a general guardian of the bacterial genome? *DNA Repair (Amst)* **9**, 210–223 (2010).
36. Deng, L., Chen, C. L., Zhai, Y., Dong, Y. & Lou, H. Editorial: DNA Replication Stress and Cell Fate. *Front Cell Dev Biol* **9**, 2999 (2021).
37. Gwynn, E. J. *et al.* The Conserved C-Terminus of the PcrA / UvrD Helicase Interacts Directly with RNA Polymerase. *PLoS One* **8**, 1–11 (2013).

38. Ghosal, G. & Chen, J. DNA damage tolerance: a double-edged sword guarding the genome. *Transl Cancer Res* **2**, 107–129 (2013).
39. Grove, J. I., Harris, L., Buckman, C. & Lloyd, R. G. DNA double strand break repair and crossing over mediated by RuvABC resolvase and RecG translocase. *DNA Repair (Amst)* **7**, 1517–1530 (2008).
40. Bhattacharyya, B. *et al.* Structural mechanisms of PriA-mediated DNA replication restart. *Proceedings of the National Academy of Sciences* **111**, 1373–1378 (2014).
41. Heller, R. C. & Marians, K. J. Unwinding of the nascent lagging strand by Rep and PriA enables the direct restart of stalled replication forks. *Journal of Biological Chemistry* **280**, 34143–34151 (2005).
42. Lovett, S. T. Template-switching during replication fork repair in bacteria. *DNA Repair (Amst)* (2017) doi:10.1016/j.dnarep.2017.06.014.
43. Hoeijmakers, J. H. J. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366–374 (2001).
44. Lindahl, T. An N-Glycosidase from *Escherichia coli* That Releases Free Uracil from DNA Containing Deaminated Cytosine Residues. *Proc Natl Acad Sci U S A* **71**, 3649 (1974).
45. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **1993** 362:6422 **362**, 709–715 (1993).

46. Sancar, A. & Rupp, W. D. A novel repair enzyme: UVRABC excision nuclease of *Escherichia coli* cuts a DNA strand on both sides of the damaged region. *Cell* **33**, 249–260 (1983).
47. Lahue, R. S., Au, K. G. & Modrich, P. DNA Mismatch Correction in a Defined System. *Science (1979)* **245**, 160–164 (1989).
48. Schormann, N., Ricciardi, R. & Chattopadhyay, D. Uracil-DNA glycosylases—Structural and functional perspectives on an essential family of DNA repair enzymes. *Protein Sci* **23**, 1667 (2014).
49. Jacobs, A. L. & Schär, P. DNA glycosylases: in DNA repair and beyond. *Chromosoma* **121**, 1 (2012).
50. Cadet, J. & Richard Wagner, J. DNA Base Damage by Reactive Oxygen Species, Oxidizing Agents, and UV Radiation. *Cold Spring Harb Perspect Biol* **5**, (2013).
51. Almatarneh, M. H., Flinn, C. G. & Poirier, R. A. Mechanisms for the Deamination Reaction of Cytosine with H₂O/OH⁻ and 2H₂O/OH⁻: A Computational Study. *J Chem Inf Model* (2008) doi:10.1021/ci7003219.
52. Kreutzer, D. A. & Essigmann, J. M. Oxidized, deaminated cytosines are a source of C → T transitions in vivo. *Proc Natl Acad Sci U S A* **95**, 3578–3582 (1998).
53. Dodsons, M. L., Michaelis, M. L. & Lloyd, S. Unified Catalytic Mechanism for DNA Glycosylases*. *Journal of Biological Chemistry* **269**, 32709–32712 (1994).
54. Savva, R., McAuley-Hecht, K., Brown, T. & Pearl, L. The structural basis of specific base-excision repair by uracil-DNA glycosylase. *Nature* **373**, 487–493 (1995).

55. Aravind, L. & Koonin, E. v. The alpha/beta fold uracil DNA glycosylases: a common origin with diverse fates. *Genome Biol* **1**, research0007.1 (2000).
56. Mullins, E. A., Warren, G. M., Bradley, N. P. & Eichman, B. F. Structure of a DNA glycosylase that unhooks interstrand cross-links. *Proc Natl Acad Sci U S A* **114**, 4400–4405 (2017).
57. Bradley, N. P., Wahl, K. L., Steenwyk, J. L., Rokas, A. & Eichman, B. F. Resistance-Guided Mining of Bacterial Genotoxins Defines a Family of DNA Glycosylases. *mBio* **13**, (2022).
58. Bradley, N. P., Washburn, L. A., Christov, P. P., Watanabe, C. M. H. & Eichman, B. F. Escherichia coli YcaQ is a DNA glycosylase that unhooks DNA interstrand crosslinks. *Nucleic Acids Res* **48**, 7005–7017 (2020).
59. Krokan, H. E. & Bjørås, M. Base Excision Repair. *Cold Spring Harb Perspect Biol* **5**, 1–22 (2013).
60. Sung, J. S. & Demple, B. Roles of base excision repair subpathways in correcting oxidized abasic sites in DNA. *FEBS Journal* vol. 273 1620–1629 Preprint at <https://doi.org/10.1111/j.1742-4658.2006.05192.x> (2006).
61. Wallace, S. S., Murphy, D. L. & Sweasy, J. B. Base Excision Repair and Cancer. *Cancer Lett* **327**, 73 (2012).
62. Xie, Y. *et al.* Deficiencies in mouse Myh and Ogg1 result in tumor predisposition and G to T mutations in codon 12 of the K-ras oncogene in lung tumors. *Cancer Res* **64**, 3096–3102 (2004).

63. Cortázar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470**, 419–423 (2011).
64. Tolentino, J. H., Burke, T. J., Mukhopadhyay, S., McGregor, W. G. & Basu, A. K. Inhibition of DNA replication fork progression and mutagenic potential of 1, N⁶-ethenoadenine and 8-oxoguanine in human cell extracts. *Nucleic Acids Res* **36**, 1300–1308 (2008).
65. Farnell, D. Nucleotide Excision Repair in the Three Domains of Life. *WURJ Health and Natural Sciences* **2**, 1–6 (2011).
66. Kisker, C., Kuper, J. & van Houten, B. Prokaryotic Nucleotide Excision Repair. *Cold Spring Harb Perspect Biol* **5**, a012591 (2013).
67. Schärer, O. D. Nucleotide Excision Repair in Eukaryotes. *Cold Spring Harb Perspect Biol* **5**, (2013).
68. van Houten, B. Nucleotide excision repair in *Escherichia coli*. *Microbiol Rev* **54**, 18–51 (1990).
69. Nakano, T. *et al.* Nucleotide Excision Repair and Homologous Recombination Systems Commit Differentially to the Repair of DNA-Protein Crosslinks. *Mol Cell* **28**, 147–158 (2007).
70. White, M. F. & Allers, T. DNA repair in the archaea—an emerging picture. *FEMS Microbiol Rev* **42**, 514–526 (2018).
71. Orren, D. K. & Sancar, A. Formation and Enzymatic Properties of the UvrB*DNA Complex*. *Journal of Biological Chemistry* **265**, 15796–15803 (1990).

72. DellaVecchia, M. J. *et al.* Analyzing the handoff of DNA from UvrA to UvrB utilizing DNA-protein photoaffinity labeling. *Journal of Biological Chemistry* **279**, 45245–45256 (2004).
73. Moolenaar, G. F., Moorman, C. & Goosen, N. Role of the Escherichia coli nucleotide excision repair proteins in DNA replication. *J Bacteriol* **182**, 5706–5714 (2000).
74. Li, G.-M. Mechanisms and functions of DNA mismatch repair. *Cell Res* **18**, 85–98 (2008).
75. Wang, H. & Hays, J. B. Human DNA mismatch repair: Coupling of mismatch recognition to strand-specific excision. *Nucleic Acids Res* **35**, 6727–6739 (2007).
76. Modrich, P. Mechanisms in E. coli and Human Mismatch Repair (Nobel Lecture). *Angewandte Chemie International Edition* **55**, 8490–8501 (2016).
77. Kadyrov, F. A., Dzantiev, L., Constantin, N. & Modrich, P. Endonucleolytic Function of MutL α in Human Mismatch Repair. *Cell* **126**, 297–308 (2006).
78. Natrajan, G. *et al.* Structures of Escherichia coli DNA mismatch repair enzyme MutS in complex with different mismatches: a common recognition mode for diverse substrates. *Nucleic Acids Res* **31**, 4814 (2003).
79. Jiricny, J. Postreplicative Mismatch Repair. *Cold Spring Harb Perspect Biol* **5**, a012633 (2013).
80. Sung, P. & Klein, H. Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nat Rev Mol Cell Biol* **7**, 739–750 (2006).

81. Davis, A. J. & Chen, D. J. DNA double strand break repair via non-homologous end-joining. *Transl Cancer Res* **2**, 130–143 (2013).
82. Crickarda, J. B., Kaniecki, K., Kwon, Y., Sung, P. & Greene, E. C. Meiosis-specific recombinase Dmc1 is a potent inhibitor of the Srs2 antirecombinase. *Proc Natl Acad Sci U S A* **115**, E10041–E10048 (2018).
83. Li, X. & Heyer, W.-D. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res* **18**, 99–113 (2008).
84. Huang, Y. & Li, L. DNA crosslinking damage and cancer - a tale of friend and foe. *Translational Cancer Research* vol. 2 144–154 Preprint at <https://doi.org/10.3978/j.issn.2218-676X.2013.03.01> (2013).
85. Kuzminov, A. Recombinational Repair of DNA Damage in Escherichia coli and Bacteriophage λ . *Microbiology and Molecular Biology Reviews* **63**, 751 (1999).
86. Cortés-Ledesma, F. & Aguilera, A. Double-strand breaks arising by replication through a nick are repaired by cohesin-dependent sister-chromatid exchange. *EMBO Rep* **7**, 919 (2006).
87. Roman, L. J., Eggleston, A. K., Kowalczykowski, S. C., Eggleston, A. K. & Kowalczykowski, S. C. Processivity of the DNA helicase activity of Escherichia coli recBCD enzyme. *J Biol Chem* **267**, 4207–4214 (1992).
88. Spies, M., Dillingham, M. S. & Kowalczykowski, S. C. Translocation by the RecB motor is an absolute requirement for χ -recognition and RecA protein loading by RecBCD enzyme. *Journal of Biological Chemistry* **280**, 37078–37087 (2005).

89. Zhao, F., Kim, W., Kloeber, J. A. & Lou, Z. DNA end resection and its role in DNA replication and DSB repair choice in mammalian cells. *Experimental & Molecular Medicine* 2020 52:10 **52**, 1705–1714 (2020).
90. Bhat, K. P. & Cortez, D. RPA and RAD51: fork reversal, fork protection, and genome stability. *Nat Struct Mol Biol* **25**, 446–453 (2018).
91. 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**, 306–313 (2011).
92. Zou, Y., Liu, Y., Wu, X. & Shell, S. M. Functions of human replication protein A (RPA): From DNA replication to DNA damage and stress responses. *J Cell Physiol* **208**, 267–273 (2006).
93. Hishida, T. *et al.* Role of the Escherichia coli RecQ DNA helicase in SOS signaling and genome stabilization at stalled replication forks. *Genes Dev* **18**, 1886 (2004).
94. Jazayeri, A. *et al.* ATM- and cell cycle-dependent regulation of ATR in response to DNA double-strand breaks. *Nat Cell Biol* **8**, 37–45 (2006).
95. Maréchal, A. & Zou, L. DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb Perspect Biol* **5**, (2013).
96. Lin, Z., Kong, H., Nei, M. & Ma, H. Origins and evolution of the recA/RAD51 gene family: Evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc Natl Acad Sci U S A* **103**, 10328–10333 (2006).

97. Zhao, B. *et al.* ATPase activity tightly regulates RecA nucleofilaments to promote homologous recombination. *Cell Discov* **3**, 1–15 (2017).
98. Haldenby, S., White, M. F. & Allers, T. RecA family proteins in archaea: RadA and its cousins. *Biochem Soc Trans* **37**, 102–107 (2009).
99. Berti, M. *et al.* Sequential role of RAD51 paralog complexes in replication fork remodeling and restart. *Nat Commun* **11**, 1–12 (2020).
100. Tarsounas, M. *et al.* Telomere maintenance requires the RAD51D recombination/repair protein. *Cell* **117**, 337–347 (2004).
101. Malkova, A. & Ira, G. Break-induced replication: functions and molecular mechanism. *Curr Opin Genet Dev* **23**, 271 (2013).
102. Li, X. & Heyer, W. D. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Research 2008 18:1* **18**, 99–113 (2008).
103. Kogoma, T. Recombination by replication. *Cell* **85**, 625–627 (1996).
104. Van Gool, A. J., Hajibagheri, N. M. A., Stasiak, A. & West, S. C. Assembly of the Escherichia coli RuvABC resolvosome directs the orientation of Holliday junction resolution. *Genes Dev* **13**, 1861–1870 (1999).
105. 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).
106. Kogoma, T. Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiol Mol Biol Rev* **61**, 212–238 (1997).

107. 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).
108. 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).
109. Marini, F. & Wood, R. D. A human DNA helicase homologous to the DNA cross-link sensitivity protein Mus308. *Journal of Biological Chemistry* **277**, 8716–8723 (2002).
110. Buckley, R. J., Kramm, K., Cooper, C. D. O., Grohmann, D. & Bolt, E. L. Mechanistic insights into Lhr helicase function in DNA repair. *Biochemical Journal* **477**, 2935–2947 (2020).
111. Snider, J., Thibault, G. & Houry, W. A. The AAA+ superfamily of functionally diverse proteins. *Genome Biol* **9**, 1–8 (2008).
112. Byrd, A. K. & Raney, K. D. Superfamily 2 helicases. *Front Biosci (Landmark Ed)* **17**, 2070–88 (2012).
113. Lohman, T. M., Tomko, E. J. & Wu, C. G. Non-hexameric DNA helicases and translocases: mechanisms and regulation. *Nature Reviews Molecular Cell Biology* **2008 9:5** **9**, 391–401 (2008).
114. Byrd, A. K. & Raney, K. D. Protein displacement by an assembly of helicase molecules aligned along single-stranded DNA. *Nat Struct Mol Biol.* **11**, 531–538 (2004).

115. Fairman-Williams, M. E., Guenther, U.-P. & Jankowsky, E. SF1 and SF2 : family matters. **20**, 313–324 (2010).
116. Chamieh, H., Ibrahim, H. & Kozah, J. Genome-wide identification of SF1 and SF2 helicases from archaea. *Gene* **576**, 214–228 (2016).
117. Chen, C. Y., Liu, X., Boris-Lawrie, K., Sharma, A. & Jeang, K. T. Cellular RNA helicases and HIV-1: Insights from genome-wide, proteomic, and molecular studies. *Virus Res* **10:1**, 33–43 (2012).
118. Hall, M. C. & Matson, S. W. Helicase motifs: the engine that powers DNA unwinding. *Mol Microbiol* **34**, 867–877 (1999).
119. Reuven, N. B., Koonin, E. v., Rudd, K. E. & Deutscher, M. P. The gene for the longest known Escherichia coli protein is a member of helicase superfamily II. *J Bacteriol* **177**, 5393–5400 (1995).
120. Ordonez, H. & Shuman, S. Mycobacterium smegmatis Lhr is a DNA-dependent ATPase and a 3'-to-5' DNA translocase and helicase that prefers to unwind 3'-tailed RNA:DNA hybrids. *Journal of Biological Chemistry* **288**, 14125–14134 (2013).
121. Hajj, M. *et al.* Phylogenetic diversity of lhr proteins and biochemical activities of the thermococcales alhr2 dna/rna helicase. *Biomolecules* **11**, (2021).
122. Schütz, P. *et al.* Comparative Structural Analysis of Human DEAD-Box RNA Helicases. *PLoS One* **5**, 1–11 (2010).
123. Xia, J. *et al.* Bacteria-to-human protein networks reveal origins of endogenous DNA damage. *Cell* **176**, 127 (2019).

124. Abdelhaleem, M., Maltais, L. & Wain, H. The human DDX and DHX gene families of putative RNA helicases. *Genomics* **81**, 618–622 (2003).
125. Warren, G. M., Wang, J., Patel, D. J. & Shuman, S. Oligomeric quaternary structure of Escherichia coli and Mycobacterium smegmatis Lhr helicases is nucleated by a novel C-terminal domain composed of five winged-helix modules. *Nucleic Acids Res* **49**, 3876–3887 (2021).
126. Seier, T., Zilberberg, G., Zeiger, D. M. & Lovett, S. T. Azidothymidine and other chain terminators are mutagenic for template-switch-generated genetic mutations. *Proc Natl Acad Sci U S A* **109**, 6171 (2012).
127. Song, X., Huang, Q., Ni, J., Yu, Y. & Shen, Y. Knockout and functional analysis of two DExD/H-box family helicase genes in Sulfolobus islandicus REY15A. *Extremophiles* **20**, 537–546 (2016).
128. Chang, M., Bellaoui, M., Boone, C. & Brown, G. W. A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. *Proc Natl Acad Sci U S A* **99**, 16934–16939 (2002).
129. Boshoff, H. I. M., Reed, M. B., Barry, C. E. & Mizrahi, V. DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. *Cell* **113**, 183–193 (2003).
130. Rand, L. *et al.* The majority of inducible DNA repair genes in Mycobacterium tuberculosis are induced independently of RecA. *Mol Microbiol* **50**, 1031–1042 (2003).

131. Suzuki, S. *et al.* Genetic and Biochemical Characterizations of aLhr1 Helicase in the Thermophilic Crenarchaeon *Sulfolobus acidocaldarius*. *Catalysts* 2022, Vol. 12, Page 34 **12**, 34 (2021).
132. van Wolferen, M., Ma, X. & Albers, S. V. DNA Processing Proteins Involved in the UV-Induced Stress Response of Sulfolobales. *J Bacteriol* **197**, 2941 (2015).
133. Northall, S. J. *et al.* DNA binding and unwinding by Hel308 helicase requires dual functions of a winged helix domain. *DNA Repair (Amst)* **57**, (2017).
134. Ejaz, A., Ordonez, H., Jacewicz, A., Ferrao, R. & Shuman, S. Structure of mycobacterial 3'-to-5' RNA:DNA helicase Lhr bound to a ssDNA tracking strand highlights distinctive features of a novel family of bacterial helicases. *Nucleic Acids Res* **46**, 442 (2018).
135. de Felice, M. *et al.* A novel DNA helicase with strand-annealing activity from the crenarchaeon *Sulfolobus solfataricus*. *Biochem J* **408**, 87 (2007).
136. Ejaz, A. & Shuman, S. Characterization of Lhr-Core DNA helicase and manganese- dependent DNA nuclease components of a bacterial gene cluster encoding nucleic acid repair enzymes. *Journal of Biological Chemistry* **293**, 17491–17504 (2018).
137. Hishida, T. *et al.* Role of the *Escherichia coli* RecQ DNA helicase in SOS signaling and genome stabilization at stalled replication forks. *Genes Dev* **18**, 1886–1897 (2004).

138. Cherepanov, P. P. & Wackernagel, W. Gene disruption in *Escherichia coli*: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene* **158**, 9–14 (1995).
139. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A* **97**, 6640 (2000).
140. Saragliadis, A., Trunk, T. & Leo, J. C. Producing Gene Deletions in *Escherichia coli* by P1 Transduction with Excisable Antibiotic Resistance Cassettes. *J Vis Exp* **2018**, 58267 (2018).
141. Sanders, E. R. Aseptic Laboratory Techniques : Plating Methods 2 . Streak Plate Procedure : Isolation of Bacterial Colonies Using the Quadrant Method. *J. Vis. Exp* **63**, 1–18 (2012).
142. Promega. Wizard SV Gel and PCR Clean-Up System Technical Bulletin, TB308. *Promega* 1–13 (2010).
143. NEB. NEBuilder HiFi DNA Assembly Reaction (E2621). 2–3 (2015).
144. Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**, 465–469 (2008).
145. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
146. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

147. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564–577 (2007).
148. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–552 (2000).
149. Guindon, S. & Gascuel, O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst Biol* **52**, 696–704 (2003).
150. Chevenet, F., Brun, C., Bañuls, A. L., Jacq, B. & Christen, R. TreeDyn: Towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics* **7**, 439 (2006).
151. Kumar, S. R. & Imlay, J. A. How Escherichia coli Tolerates Profuse Hydrogen Peroxide Formation by a Catabolic Pathway. *J Bacteriol* **195**, 4569 (2013).
152. Wei, Y., Vollmer, A. C. & LaRossa, R. A. In Vivo Titration of Mitomycin C Action by Four Escherichia coli Genomic Regions on Multicopy Plasmids. *J Bacteriol* **183**, 2259 (2001).
153. Zenkin, N., Kulbachinskiy, A., Bass, I. & Nikiforov, V. Different rifampin sensitivities of Escherichia coli and Mycobacterium tuberculosis RNA polymerases are not explained by the difference in the β -subunit rifampin regions I and II. *Antimicrob Agents Chemother* **49**, 1587–1590 (2005).
154. Prism - GraphPad. <https://www.graphpad.com/scientific-software/prism/>.
155. GelAnalyzer. <http://www.gelanalyzer.com/?i=1>.

156. Buckley, R. J. Biochemical Analysis of Hel308 and HelQ Helicases. (University of Nottingham, 2017).
157. Invitrogen. SimplyBlue™ SafeStain For fast, sensitive, and safe Coomassie G-250 staining of proteins Catalog numbers LC6060, LC6065. (2012).
158. Sullivan, M. R. & Bernstein, K. A. RAD-ical new insights into RAD51 regulation. *Genes* vol. 9 Preprint at <https://doi.org/10.3390/genes9120629> (2018).
159. Brendel, V., Brocchieri, L., Sandler, S. J., Clark, A. J. & Karlin, S. Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms. *J Mol Evol* **44**, 528–541 (1997).
160. Guy, C. P. *et al.* Interactions of RadB, a DNA repair protein in archaea, with DNA and ATP. *J Mol Biol* **358**, 46–56 (2006).
161. Wardell, K. *et al.* RadB acts in homologous recombination in the archaeon *Haloferax volcanii*, consistent with a role as recombination mediator. *DNA Repair (Amst)* **55**, 7–16 (2017).
162. Diallo, A. B. *et al.* RadA, a Key Gene of the Circadian Rhythm of *Escherichia coli*. *Int J Mol Sci* **23**, (2022).
163. Beam, C. E., Saveson, C. J. & Lovett, S. T. Role for radA/sms in Recombination Intermediate Processing in *Escherichia coli*. *J Bacteriol* **184**, 6836 (2002).
164. Inoue, M. *et al.* The Lon protease-like domain in the bacterial RecA paralog RadA is required for DNA binding and repair. *Journal of Biological Chemistry* **292**, 9801–9814 (2017).

165. Marie, L. *et al.* Bacterial RadA is a DnaB-type helicase interacting with RecA to promote bidirectional D-loop extension. *Nature Communications* 2017 **8**:1 **8**, 1–14 (2017).
166. Lee, I. & Suzuki, C. K. Functional mechanics of the ATP-dependent Lon protease—lessons from endogenous protein and synthetic peptide substrates. *Biochimica et Biophysica Acta - Proteins and Proteomics* vol. 1784 727–735 Preprint at <https://doi.org/10.1016/j.bbapap.2008.02.010> (2008).
167. Cooper, D. L. & Lovett, S. T. Recombinational branch migration by the RadA/Sms paralog of RecA in *Escherichia coli*. *Elife* **5**, (2016).
168. Diver, W. P., Sargentini, N. J. & Smith, K. C. A Mutation (*radA100*) in *Escherichia Coli* That Selectively Sensitizes Cells Grown in Rich Medium to X- or U.V.-radiation, or Methyl Methanesulphonate. <https://doi.org/10.1080/09553008214551251> **42**, 339–346 (2009).
169. Persky, N. S. & Lovett, S. T. Mechanisms of Recombination: Lessons from *E. coli*. <http://dx.doi.org/10.1080/10409230802485358> **43**, 347–370 (2009).
170. Carrasco, B., Cozar, M. C., Lurz, R., Alonso, J. C. & Ayora, S. Genetic Recombination in *Bacillus subtilis* 168: Contribution of Holliday Junction Processing Functions in Chromosome Segregation. *J Bacteriol* **186**, 5557 (2004).
171. Cohen, S. E. & Golden, S. S. Circadian Rhythms in Cyanobacteria. *Microbiol Mol Biol Rev* **79**, 373 (2015).
172. Ishiura, M. *et al.* Expression of a Gene Cluster *kaiABC* as a Circadian Feedback Process in Cyanobacteria. *Science* (1979) **281**, 1519–1523 (1998).

173. Ye, J., Osborne, A. R., Groll, M. & Rapoport, T. A. RecA-like motor ATPases - Lessons from structures. *Biochimica et Biophysica Acta - Bioenergetics* vol. 1659 1–18 Preprint at <https://doi.org/10.1016/j.bbabi.2004.06.003> (2004).
174. Hajredini, F. & Ghose, R. A Conserved Structural Role for the Walker-A Lysine in P-Loop Containing Kinases. *Front Mol Biosci* **8**, 948 (2021).
175. The PyMOL Molecular Graphics System; Version 1.2r3pre; Schrödinger; LLC. PyMOL | pymol.org. <https://pymol.org/2/>.
176. Guo, P. *et al.* Controlling the Revolving and Rotating Motion Direction of Asymmetric Hexameric Nanomotor by Arginine Finger and Channel Chirality. *ACS Nano* **13**, 6207–6223 (2019).
177. Grigorescu, A. A. *et al.* Inter-subunit interactions that coordinate Rad51's activities. *Nucleic Acids Res* **37**, 557 (2009).
178. Nagy, G. N. *et al.* Structural Characterization of Arginine Fingers: Identification of an Arginine Finger for the Pyrophosphatase dUTPases. *J Am Chem Soc* **138**, 15035–15045 (2016).
179. Kannan, L. & Wheeler, W. C. Maximum Parsimony on Phylogenetic networks. *Algorithms for Molecular Biology* **7**, 1–10 (2012).
180. Baum, D. Phylogenetic Trees and Monophyletic Groups | Learn Science at Scitable. *Nature Education* 190 <https://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956/> (2009).

181. Bitomský, M., Mládková, P., Pakeman, R. J. & Duchoslav, M. Clade composition of a plant community indicates its phylogenetic diversity. *Ecol Evol* **10**, 3747 (2020).
182. De Moraes Russo, C. A. & Selvatti, A. P. Bootstrap and Rogue Identification Tests for Phylogenetic Analyses. *Mol Biol Evol* **35**, 2327–2333 (2018).
183. Kioy, D., Jannin, J. & Mattock, N. Human African trypanosomiasis. *Nat Rev Microbiol* **2**, 186–187 (2004).
184. Masson, J. Y., Stasiak, A. Z., Stasiak, A., Benson, F. E. & West, S. C. Complex formation by the human RAD51C and XRCC3 recombination repair proteins. *Proc Natl Acad Sci U S A* **98**, 8440–8446 (2001).
185. Huang, S. & Deutscher, M. P. Sequence and transcriptional analysis of the *Escherichia coli* rnt gene encoding RNase T. *Journal of Biological Chemistry* **267**, 25609–25613 (1992).
186. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* **20**, 1085–1093 (2019).
187. Assaf, R., Xia, F. & Stevens, R. Detecting operons in bacterial genomes via visual representation learning. *Scientific Reports* 2021 11:1 **11**, 1–10 (2021).
188. Burkhardt, D. H. *et al.* Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *Elife* **6**, (2017).
189. Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. & Koonin, E. v. Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context. *Genome Res* **11**, 356–372 (2001).

190. Matange, N., Podobnik, M. & Visweswariah, S. S. Metallophosphoesterases: structural fidelity with functional promiscuity. *Biochemical Journal* **467**, 201–216 (2015).
191. Ghosh, S., Ejaz, A., Repeta, L. & Shuman, S. Pseudomonas putida MPE, a manganese-dependent endonuclease of the binuclear metallophosphoesterase superfamily, incises single-strand DNA in two orientations to yield a mixture of 3'-PO₄ and 3'-OH termini. *Nucleic Acids Res* **49**, 1023–1032 (2021).
192. Tyagi, R., Shenoy, A. R. & Visweswariah, S. S. Characterization of an Evolutionarily Conserved Metallophosphoesterase That Is Expressed in the Fetal Brain and Associated with the WAGR Syndrome. *Journal of Biological Chemistry* **284**, 5217–5228 (2009).
193. Padmanabha, K. P. & Deutscher, M. P. RNase T affects Escherichia coli growth and recovery from metabolic stress. *J Bacteriol* **173**, 1376 (1991).
194. Bechhofer, D. H. & Deutscher, M. P. Bacterial ribonucleases and their roles in RNA metabolism. *Crit Rev Biochem Mol Biol* **54**, 242 (2019).
195. Viswanathan, M., Lanjuin, A. & Lovett, S. T. Identification of RNase T as a High-Copy Suppressor of the UV Sensitivity Associated With Single-Strand DNA Exonuclease Deficiency in Escherichia coli. (1999).
196. Kelly, K. O. & Deutscher, M. P. The presence of only one of five exoribonucleases is sufficient to support the growth of Escherichia coli. *J Bacteriol* **174**, 6682 (1992).

197. Hsiao, Y. Y., Fang, W. H., Lee, C. C., Chen, Y. P. & Yuan, H. S. Structural Insights Into DNA Repair by RNase T—An Exonuclease Processing 3' End of Structured DNA in Repair Pathways. *PLoS Biol* **12**, 1001803 (2014).
198. Maclåg, A. *et al.* In vitro transcription profiling of the σ S subunit of bacterial RNA polymerase: re-definition of the σ S regulon and identification of σ S-specific promoter sequence elements. *Nucleic Acids Res* **39**, 5338 (2011).
199. Maeda, H., Fujita, N. & Ishihama, A. Competition among seven Escherichia coli σ subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Res* **28**, 3497–3503 (2000).
200. Ades, S. E., Grigorova, I. L. & Gross, C. A. Regulation of the alternative sigma factor σ E during initiation, adaptation, and shutoff of the extracytoplasmic heat shock response in Escherichia coli. *J Bacteriol* **185**, 2512–2519 (2003).
201. Hiratsu, K., Amemura, M., Nashimoto, H., Shinagawa, H. & Makino, K. The rpoE gene of Escherichia coli, which encodes σ (E), is essential for bacterial growth at high temperature. *J Bacteriol* **177**, 2918–2922 (1995).
202. Bianchi, A. A. & Baneyx, F. Hyperosmotic shock induces the sigma32 and sigmaE stress regulons of Escherichia coli. *Mol Microbiol* **34**, 1029–1038 (1999).
203. Rouviere, P. E. *et al.* rpoE, the gene encoding the second heat-shock sigma factor, sigma E, in Escherichia coli. *EMBO J* **14**, 1032–1042 (1995).
204. al Mamun, A. A. M. *et al.* Identity and function of a large gene network underlying mutagenic repair of DNA breaks. *Science* **338**, 1344–1348 (2012).

205. Egler, M., Grosse, C., Grass, G. & Nies, D. H. Role of the Extracytoplasmic Function Protein Family Sigma Factor RpoE in Metal Resistance of *Escherichia coli*. *J Bacteriol* **187**, 2297 (2005).
206. Fasnacht, M. & Polacek, N. Oxidative Stress in Bacteria and the Central Dogma of Molecular Biology. *Front Mol Biosci* **8**, 392 (2021).
207. Rare Codon Caltor, Programmed by Edmund Ng. <https://people.mbi.ucla.edu/sumchan/caltor.html>.
208. Codon usage table. <https://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=83333>.
209. Gavriilidou, A. *et al.* Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nature Microbiology* **2022 7:5 7**, 726–735 (2022).
210. Altschul. BLAST Genome Taxa Query Page. Preprint at <http://www-archbac.u-psud.fr/genomics/phylotaxBlast.html> (1997).
211. Allers, T., Ngo, H. P., Mevarech, M. & Lloyd, R. G. Development of Additional Selectable Markers for the Halophilic Archaeon *Haloferax volcanii* Based on the *leuB* and *trpA* Genes. *Appl Environ Microbiol* **70**, 943 (2004).
212. Johnson, S. J. & Jackson, R. N. Ski2-like RNA helicase structures Common themes and complex assemblies. *RNA Biology* **10**, 33–43 (2013).
213. Blouin, S., Craggs, T. D., Lafontaine, D. A. & Penedo, J. C. Functional studies of DNA-protein interactions using FRET techniques. *Methods in Molecular Biology* **543**, 475–502 (2009).

214. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006.0008 (2006).
215. Cooper, D. L. & Lovett, S. T. Toxicity and tolerance mechanisms for azidothymidine, a replication gap-promoting agent, in Escherichia coli. *DNA Repair (Amst)* **10**, 260 (2011).
216. Olivero, O. A. Mechanisms of genotoxicity of nucleoside reverse transcriptase inhibitors. *Environ Mol Mutagen* **48**, 215–223 (2007).
217. Lennicke, C., Rahn, J., Lichtenfels, R., Wessjohann, L. A. & Seliger, B. Hydrogen peroxide – production, fate and role in redox signaling of tumor cells. *Cell Communication and Signaling 2015 13:1* **13**, 1–19 (2015).
218. Auten, R. L. & Davis, J. M. Oxygen Toxicity and Reactive Oxygen Species: The Devil Is in the Details. *Pediatric Research 2009 66:2* **66**, 121–127 (2009).
219. Hazra, T. K. *et al.* Repair of hydantoin, one electron oxidation product of 8-oxoguanine, by DNA glycosylases of Escherichia coli. *Nucleic Acids Res* **29**, 1967 (2001).
220. Cheng, S. Y., Seo, J., Huang, B. T., Napolitano, T. & Champeil, E. Mitomycin C and decarbamoyl mitomycin C induce p53-independent p21WAF1/CIP1 activation. *Int J Oncol* **49**, 1815–1824 (2016).
221. Borowy-Borowski, H., Lipman, R. & Tomasz, M. Recognition between Mitomycin C and Specific DNA Sequences for Cross-Link Formation[^]. *Proc. Natl. Acad. Sci. U.S.A* **29**, 37 (1990).

222. Dapa, T., Fleurier, S., Bredeche, M. F. & Matic, I. The SOS and RpoS Regulons Contribute to Bacterial Cell Robustness to Genotoxic Stress by Synergistically Regulating DNA Polymerase Pol II. *Genetics* **206**, 1349 (2017).
223. Trautinger, B. W. & Lloyd, R. G. Modulation of DNA repair by mutations flanking the DNA channel through RNA polymerase. *EMBO J* **21**, 6944 (2002).
224. Campbell, E. A. *et al.* Structural mechanism for rifampicin inhibition of bacterial RNA polymerase. *Cell* **104**, 901–912 (2001).
225. Garibyan, L. *et al.* Use of the rpoB gene to determine the specificity of base substitution mutations on the Escherichia coli chromosome. *DNA Repair (Amst)* **2**, 593–608 (2003).
226. Chib, S., Ali, F. & Seshasayee, A. S. N. Genomewide Mutational Diversity in Escherichia coli Population Evolving in Prolonged Stationary Phase. *mSphere* **2**, (2017).
227. Harris, R. S. *et al.* Mismatch repair protein MutL becomes limiting during stationary-phase mutation. *Genes Dev* **11**, 2426 (1997).
228. Galán, J. C. *et al.* Mutation rate is reduced by increased dosage of mutL gene in Escherichia coli K-12. *FEMS Microbiol Lett* **275**, 263–269 (2007).
229. Yuan On, Y. Development and road-testing of an inducible hypermutator strain of Pseudomonas aeruginosa. (University of Cambridge, 2022).
230. Rosano, G. L., Ceccarelli, E. A., Neubauer, P., Bruno-Barcena, J. M. & Schweder, T. Recombinant protein expression in Escherichia coli: advances and challenges. *Front Microbiol.* (2014) doi:10.3389/fmicb.2014.00172.

231. Harris, E. & Angal, S. *Protein Purification Techniques: A Practical Approach. Protein Purification Techniques* (Oxford University Press, 1989). doi:10.1093/OSO/9780199636747.001.0001.
232. Bornhorst, J. A. & Falke, J. J. [16] Purification of Proteins Using Polyhistidine Affinity Tags. *Methods Enzymol* **326**, 245 (2000).
233. Hellman, L. M. & Fried, M. G. Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein-Nucleic Acid Interactions. *Nat Protoc* **2**, 1849 (2007).
234. Summer, H., Grämer, R. & Dröge, P. Denaturing Urea Polyacrylamide Gel Electrophoresis (Urea PAGE). *J Vis Exp* (2009) doi:10.3791/1485.
235. Vidal, A. E., Hickson, I. D., Boiteux, S. & Radicella, J. P. Mechanism of stimulation of the DNA glycosylase activity of hOGG1 by the major human AP endonuclease: bypass of the AP lyase activity step. *Nucleic Acids Res* **29**, 1285 (2001).
236. Hölz, K., Pavlic, A., Lietard, J. & Somoza, M. M. Specificity and Efficiency of the Uracil DNA Glycosylase-Mediated Strand Cleavage Surveyed on Large Sequence Libraries. *Sci Rep* **9**, (2019).
237. Jiang, Y., Liu, C. & Huang, A. EDTA-Functionalized Covalent Organic Framework for the Removal of Heavy-Metal Ions. *ACS Appl Mater Interfaces* **11**, 32186–32191 (2019).
238. Krokan, H. E., Drabløs, F. & Slupphaug, G. Uracil in DNA – occurrence, consequences and repair. *Oncogene* 2002 21:58 **21**, 8935–8948 (2002).

239. 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).
240. Rubinson, E. H., Gowda, A. S. P., Spratt, T. E., Gold, B. & Eichman, B. F. An Unprecedented Nucleic Acid Capture Mechanism for Excision of DNA Damage. *Nature* **468**, 406 (2010).
241. Mullins, E. A., Warren, G. M., Bradley, N. P. & Eichman, B. F. Structure of a DNA glycosylase that unhooks interstrand cross-links. *Proc Natl Acad Sci U S A* **114**, 4400–4405 (2017).
242. Haushalter, K. A., Stukenberg, P. T., Kirschner, M. W. & Verdine, G. L. Identification of a new uracil-DNA glycosylase family by expression cloning using synthetic inhibitors. *Current Biology* **9**, 174–185 (1999).
243. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343–345 (2009).
244. New England BioLabs. Q5[®] Site-Directed Mutagenesis Kit Protocol (E0554). 7–8 (2019).
245. Proudfoot, C. & McCulloch, R. Distinct roles for two RAD51-related genes in *Trypanosoma brucei* antigenic variation. *Nucleic Acids Res* **33**, 6906 (2005).
246. Dobson, R., Stockdale, C., Lapsley, C., Wilkes, J. & McCulloch, R. Interactions among *Trypanosoma brucei* RAD51 paralogues in DNA repair and antigenic variation. *Mol Microbiol* **81**, 434–456 (2011).

247. Swartley, J. S. *et al.* Capsule switching of *Neisseria meningitidis*. *Proc Natl Acad Sci U S A* **94**, 271–276 (1997).
248. Deitsch, K. W., Lukehart, S. A. & Stringer, J. R. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. *Nat Rev Microbiol* **7**, 493 (2009).
249. Chevigny, N. *et al.* RADA-dependent branch migration has a predominant role in plant mitochondria and its defect leads to mtDNA instability and cell cycle arrest. *PLoS Genet* **18**, e1010202 (2022).
250. Mishra, A., Saxena, S., Kaushal, A. & Nagaraju, G. RAD51C/XRCC3 Facilitates Mitochondrial DNA Replication and Maintains Integrity of the Mitochondrial Genome. *Mol Cell Biol* **38**, (2018).
251. Raney, K. D. Chemical modifications of DNA for study of helicase mechanisms. *Bioorg Med Chem* **22**, 4399–4406 (2014).
252. Jenkins, T. *et al.* The HelQ human DNA repair helicase utilizes a PWI-like domain for DNA loading through interaction with RPA, triggering DNA unwinding by the HelQ helicase core. *NAR Cancer* **3**, (2021).
253. Nabel, C. S., Manning, S. A. & Kohli, R. M. The Curious Chemical Biology of Cytosine: Deamination, Methylation and Oxidation as Modulators of Genomic Potential. *ACS Chem Biol* **7**, 20 (2012).
254. Zuo, S., Boorstein, R. J. & Teebor, G. W. Oxidative damage to 5-methylcytosine in DNA. *Nucleic Acids Res* **23**, 3239 (1995).

Chapter 8 : Appendices

8.1 Appendix 1

8.1.1 *M. thermautotrophicus* Lhr protein sequence

Mth_1802

MIKKQERMYTSGEIHSILHPWVSEWFRRTFDDFTEAQRYAIMDIHRGRNVLVSSPTGSGKTLTAFLSIISEL
TRLADDGELEDSVYCIYISPLKALDNDIERNLEEPLSAIRDIAAGEGRDLEIRKAVRTGDTTSYERSRMLKK
PPHILITTPETLSILLVAPKFREKLSTVRYVIVDEIHS LADNKRGVHLSLSLERLQHLVGDFTRIGLSATVH
PLERVARFLVGYSGSERECLIVDVSYLKELDIDLICPVDDIVAADPEEIGNALYDILHDLIEDHRTTLIFT
NTRSGTESVVYNLKS RFPESYSDSNIMAHSSLSREIRLETEEKLKRGELKAVVSSTSLELGIDIGYIDLVV
LLSSPKSVSRALQRIGRSGHQLHQRSKGRIVVVD RDDLVECSLILKNALEGKIDS IKVPENCLDVLAQHIYG
MAIENPWDIDHALAVIRNSYCYRNLSREDYLSVLSY LAGEYVELEERYVYAKIWVDYDKNQFGKRGKLARML
YSTNIGTIPDRSAAVVKCGGKVVGRIEEDFMEKLRKGD T FVLGGRIYRFNYARGMTVNVTPASGPPTIPSWF
SEQPLSFDLALDIQRFRDIMDGK FQYGRSRDEIMEFIMS YLHVDERAASSIYEYFREQYLYAGIP SIRRML
VEYYTGFGGRKFIVFHSLFGRRVND AISRAVAYVIARR YRDVMI SVSDNGFYLSSEGKMGGLEAFRELEPE
NLRNVLKKALDR TETLASRFRHCAGRALMILRRYR GEEKSVGRQQVRGKILLKFVSELDDKFP ILEEARREV
MEDYMDIENAIRVLEWIRDGDMEIKQINTRIPSPFAFNLVAQGYLDV **LKYE**DRIEFIRRMHQAIIDEIKR

Figure 8.1 *MthLhr* amino acid sequence as used.

SUMO-1 motif sequence (**bold**).

8.1.2 *E. coli* Lhr protein sequences

Eco_b1653

MADNPDPSSLLPDVFSPATRDWFLRAFKQPTAVQPQTWHVAARSEHALVIAPTGSCKTLAAFLYALDRLFRE
GGEDTREAHKRKTSRILYISPIKALGTDVQRNLQIPLKGIADERRRRGETEVNLRVGI RTGDTPAQERSKLT
RNPPDILITTPESLYLMLTSRARETLRGVETVIIDEVHAVAGSKRG AHLALSLERLDALLHTSAQRIGLSAT
VRSASDVAAFLGGDRPVTVVNPAMP RHPQIRIVVPVANMDDVSSVASGTGEDSHAGREGSIWPYIETGILDE
VLRHRSTIVFTNSRGLAEKLTARLNELYAARLQRSPIAVDAAHFESTSGATSNRVQSSDVFIARSHHGSVS
KEQRAITEQALKSGELRCVVATSSLELGIDMGAVDLVIQVATPLSVASGLQRIGRAGHQVGGVSKGLFFPRT
RRDLVDSAVIVECMFAGRLENLTPPHNPLDVLAQQTVAAAAMDALQVDEWYSRVRAAPWKDLPRRVFDATL
DMLSGRYPSGDFSAFRPKLVWNRETGILTARPGAQLLAVTSGGTIPDRGMYSVLLPEGE EKAGSRRVGE LDE
EMVYESRVNDIITLGATSWRIQQITRDQVIVTPAPGRSARLPFWRGE GNGRPAELGEMIGDFLHLLADGAFF
SGTIPPWLAEENTIANIQGLIEEQRNATGIVPGSRHLVLERCRDEIGDWRIILHSPYGRRVHEPWAVAIAGR
IHALWGADASVVASDDGIVARIPD TDGKLPDAAIFLFEPEKLLQIVREAVGSSALFAARFRECAARALLMPG
RTPGHRTPLWQQRLRASQ LLEIAQGYPDFPVILETLRECLQDVYDLPALERLMRR LINGGEIQISDVTTTTPS
PFATSLFLGYVAEFMYQSDAPLAERRASVLSL DSELLRNLLGQVDPGELLDPQVIRQVEEELQRLAPGRRAK
GEEGLFDLLRELGPMTVEDLAQRHTGSSEEVASYLENLLAVKRIFPAMISGQERLACMDDAARLRDALGVRL
PESLPEIYLHRVSYPLRDLFLRYLRAHALVTAEQLAHEFSLGIAIVEEQQLQQLREQGLVMNLQQDIWVSDEV
FRRLRLRS LQAAREATRPVAATTYARLLLERQGVLPATDGSPALFASTSPGVYEGVDGVMRVIEQLAGVGLP
ASLWESQILPARVRDYSSEMLDELLATGAVIWSGQKKGEDDGLVALHLQEYAAESFTPAEADQANRSALQQ
AIVAVLADGGAWFAQQISQRIRDKIGESVDLSALQEALWALVWQGVITSDIWAPLRALTRSSSNARTSTRRS
HRARRGRPVYAQPVS PRVSYNTPNLAGRWSLLQVEPLN DTERMLALAENMLDRYGIISRQAVIAENIPGGFP
SMQTLCRSMEDSGRIMGRFVEGLGGAQFAERLTI DRLRLATQATQTRHYTPVALSANDPANVWGNLLPWP
AHPATLVPTRRAGALVVVSGGKLLLYLAQGGKMLVWQEKEELLAPEV FHALTTALRREPRLRFTLLEVN DL
PVRQTPMFTLLREAGFSSSPQGLDWG

Figure 8.2 *EcoLhr* protein sequence untagged as used.

Eco_b1653 with His₆ tag

MHHHHHHADNPDPSLLPDVFSPATRDWFLRAFKQPTAVQPQTWHVAARSEHALVIAPT**GSGKT**LAAFLYAL
DRLFREGGEDTREAHKRKTSRILYISPIKALGTDVQRNLQIPLKGIADERRRRGETEVNLRVGI RTGDTPAQ
ERSKLTRNPPDILITTPESLYMLTTSRARETLRGVE**TVIIDE**VHAVAGSKRG AHLALS LERLDALLHTSAQR
IGLSATVRSASDVAAFLGGDRPVTVVNPPAMRHPQIRIVVPVANMDDVSSVASGTGEDSHAGREGSIWPYIE
TGILDEVLRHRSTIVFTNSRGLAEKLTAR**L**NELYAARLQRSPSIAVDAAHFESTSGATSNRVQSSDVFIARS
HHGSVSKEQRAITEQALKSGELRCVVATSSLELGIDMGAVDLVIQVATPLSVASGLQRI GRAGHQVGGVSKG
LFFPRTRRDLVDSAVIVECMFAGRLENLTPPHNPLDVLAQQTVAAAAMDALQVDEWYSRVRRAPWKDLPRR
VFDATLDMLSGRYPGDFSAFRPKLVWNRETGILTARPGAQLLAVTSGGTIPDRGMYSVLLPEGEEKAGSRR
VGELDEEMVYESRVNDIITLGATSWRIQQITRDQVIVTPAPGRSARLPFWRGEGNGRPAELGEMIGDFLHLL
ADGAFFSGTIPWLAEEENTIANIQGLIEEQRNATGIVPGSRHLVLERCRDEIGDWRIILHSPYGRRVHEPWA
VAIAGRIHALWGADASVVASDDGIVARIPD TDGKLPDAAIFLFEPEKLLQIVREAVGSSALFAARFRECAAR
ALLMPGRTPGHRTPLWQQLRASQLEIAQGYPDFPVIETLRECLQDVYDLPALERLMRRLNGGEIQISDV
TTTTSPPFATSLFLGYVAE**FMYQ**SDAPLAERRASVLSLDELLRNLLGQVDPGELLDPQVIRQVEEELQRLA
PGRRAKGEEGLFDLLRELGPMTVEDLAQRHTGSSEEVASYLENLLAVKRIFPAMISGQERLACMDDAARLRD
ALGVRLPELPEIYLRVSYPLRDLFLRYLRAHALVTAEQLAHEFSLGIAIVEEQQLQQLREQGLVMNLQQDI
WVSDEVFRRLRLRSLQAAREATRPVAATTYARLLLLERQGVLPATDGPALFASTSPGVYEGVDGVMRVIEQL
AGVGLPASLWESQILPARVRDYSSEMLDELLATGAVIWSGQKKG**E**DDGLVALHLQEYAAESFTPAEADQAN
RSALQQAI VAVLADGGAWFAQQISQRIRDKIGESVDLSALQEALWALVWQGVITSDIWAPLRALTRSSSNAR
TSTRRSHR**ARRGR**PVYAQPVS PRVSYNTPNLAGRWSLLQVEPLNDRMLALAENMLDRYGIISRQAVIAEN
IPGGFPSMQTLCRSMEDSGRIMRGRFVEGLGGAQFAERLTIDRLRDLATQATQTRHYTPVALSANDPANVWG
NLLPWPAPATLVPTRRAGALVVVSGGKLLLYLA**Q**GGKMLVWQEKEELLAPEVFHALTTALRREPRL**RFTL**
TEVNDLPVRQTPMFTLLREAGFSSSPQGL**DWG**

Figure 8.3 *EcoLhr* with cloned in His₆ tag and important residues highlighted.

N-terminal histidine tag (orange), Walker A (red) and Walker B (purple) with target catalytic residues (green), proposed site of pre-mature transcription termination (highlight), SUMO-1 motif (bold) and glycosylase active site [arginine](#) R, [aspartic acid](#) D (catalytic residue), [glutamic acid](#) E, [glutamine](#) Q, [leucine](#) L, and [tryptophan](#) W.

Eco_b1653 Lhr-CTD, amino acids 876-1538

MHHHHHSSGMSDKIIHLTDDSFDTDLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNID
QNPGTAPKYGIRGIPTLLLFKNGEVAATKVGALSQGQLKEFLDANLAGT**ENLYFQSM**GYVAEFMYQSDAPLA
ERRASVLSLDSELLRNLLGQVDPGELLDPQVIRQVEEELQRLAPGRRAKGEEGLFDLLRELGPMTVEDLAQR
HTGSSEEVASYLENLLAVKRIFPAMISGQERLACMDDAARLRDALGVRLPESLPEIYLHRVSYPLRDLFLRY
LRAHALVTAEQLAHEFSLGIAIVEEQQLREQGLVMNLQQDIWVSDEVFRRLRLRSLQAAREATRPVAATT
YARLLERQGVLPATDGSALFASTSPGVYEGVDGVMRVEQLAGVGLPASLWESQILPARVRDYSSEMLDE
LLATGAVIWSGQKKGEDDGLVALHLQEYAAESFTPAEADQANRSALQQAIVAVLADGGAWFAQQISQRIRD
KIGESVDLSALQEALWALVWQGVITSDIWAPLRALTRSSSNARTSTRRSHRARRGRPVYAQPVS PRVSYNTP
NLAGRWSLLQVEPLNDRMLALAENMLDRYGIISRQAVIAENIPGGFPSMQTLCRSMEDSGRIMRGRFVEG
LGGAQFAERLTIDRLRLATQATQTRHYTPVALSANDPANVWGNLLPWPAPATLVPTRRAGALVVVSGGKL
LLYLAQGGKMLVWQEKEELLAPEVFHALTTALRREPRRLRFTLTVNDLPVRQTPMFTLLREAGFSSSPQGL
DWG

Figure 8.4 *Eco*Lhr-CTD protein sequence as used.

Lhr amino acids 876-1538 with N-terminal His₆ tag (orange), linker regions (underlined), *E. coli* TrxA (gold), TEV cleavage site (bold), Lhr protein start (highlight) and catalytic aspartic acid (red).

8.1.3 *E. coli* RadA (Sms) protein sequence

Eco_b4389

MGSSHHHHHSQDPAKAPKRAFVCNECGADYPRWQQQCSACHAWNTITEVRLAASPMVARNERLSGYAGSAG
VAKVQKLSDISLEELPRFSTGFKEFDRVLGGGVVPGSAILIGGNPGAGKSTLLLQTLCKLAQQMKTLYVTGE
ESLQQVAMRAHRLGLPTDNLNMLSETSIEQICLIAEEEEQPKLMVIDSIQVMHMADVQSSPGSVAQVRETAAY
LTRFAKTRGVAIVMVGHVTKDGSLAGPKVLEHCIDCSVLLDGDADSRFRTLRSKRNRFGAVNELGVFAMTEQ
GLREVSNP S A I F L S R G D E V T S G S S V M V V W E G T R P L L V E I Q A L V D H S M M A N P R R V A V G L E Q N R L A I L L A V L H R
H G G L Q M A D Q D V F V N V V G G V K V T E T S A D L A L L L A M V S S L R D R P L P Q D L V V F G E V G L A G E I R P V P S G Q E R I S E A
A K H G F R R A I V P A A N V P K K A P E G M Q I F G V K K L S D A L S V F D D L

Figure 8.5 *E. coli* RadA (Sms) as used here.

N-terminal histidine tag (orange).

8.1.4 *E. coli* RNaseT protein sequence

Eco_b1652

MGSSHHHHHSSSGLVPRGSHMSDNAQLTGLCDRFRGFYPVVIDVETAGFNAKTDALLEIAAITLKMDEQGWL
MPD^TTLHFHVEPFVGANLQPEALAFNGIDPNDPDRGAVSEYEALHEIFKVVRKGIKASGCNRAIMVAHNANF
DHSFMMAAERASLKRNPFPFATFDTAALAGLALGQTVLSKACQTAGMDFDSTQAHSALYDTERTAVLFCE
IVNRWKRLGGWPLSAAEEV

Figure 8.6 *E. coli* RNaseT as used here.

N-terminal histidine tag (orange), linker (underlined) and thrombin cleavage site (grey).

8.2 Distribution of Lhr among bacteria

Table 8.1 Results from Blastp search of *E. coli* Lhr against each bacterial phyla.

Only the top hit species is displayed and ascension numbers for those with an extended CTD. Colours indicate the strength of match **very good** <1E-100, **good** <1e-20 and **poor** >1e-20. Tally as follows **30**, **13** and **12** to a total of **54** bacterial phyla where it was possible to search through the NCBI database.

Phylum	Species	Top hit information			
		Query cover	E value	% identity	GTDB ascension No.
Acidobacteriota	Acidobacteria bacterium	96%	1.00E-163	37.97%	MBW4032335.1
Actinobacteria	Actinobacteria bacterium 13_1_20CM_4_69_9	97%	7.00E-144	35.95%	OLD99479.1
Aerophobota	Candidatus Aerophobetes bacterium	39%	5.00E-31	31.14%	
Aquificota	Sulfurihydrogenibium sp.	47%	1.00E-27	26.96%	
Armatimonadetes	Armatimonadetes bacterium	79%	0.00E+00	43.55%	MBI3926547.1
Bacteroidota	Bacteroidetes bacterium	95%	5.00E-164	37.46%	NUN68509.1
Bipolaricaulota	Candidatus Acetothermum autotrophicum	99%	0.00E+00	54.55%	Lhr core

Caldatribacteriota	Candidatus Atribacteria bacterium	96%	1.00E-132	34.91%	TFH10364.1
Caldisericota	Caldiserica bacterium	39%	3.00E-30	31.14%	
Calditrichota	Calditrichaeota bacterium	88%	4.00E-155	38.48%	KAA3613013.1
Calescibacterota	Candidatus Calescamantes bacterium	37%	7.00E-11	24.80%	
Chloroflexota	Anaerolineae bacterium	95%	9.00E-164	39.13%	CAG0947541.1
Chrysiogenetota	Chrysiogenales bacterium	53%	3.00E-25	28.03%	
Cloacimonadota	Candidatus Cloacimonetes bacterium	80%	7.00E-130	35.25%	NLG61818.1 (N), NLG63523.1 (C)
Coprothermobacterota	Coprothermobacter sp.	5%	3.00E-03	32.65%	
Cyanobacteria	Leptolyngbya sp. Heron Island J	96%	3.00E-143	34.71%	WP_023077153.1
Dadabacteria	Candidatus Dadabacteria bacterium	96%	2.00E-158	36.15%	NIP32256.1
Deferribacterota	Geovibrio thiophilus	46%	3.00E-34	30.88%	
Deinococcota	Deinococcus-Thermus bacterium	60%	2.00E-52	30.61%	
Deinobacteria	Candidatus Deinobacteria bacterium	42%	2.00E-07	21.98%	
Desantisbacteria	Candidatus Desantisbacteria bacterium CG1_02_49_89	43%	4.00E-30	28.42%	

Dictyoglomota	Dictyoglomi bacterium	40%	4.00E-16	25.93%	
Edwardsbacteria	Candidatus Edwardsbacteria bacterium	9%	1.90E+00	35.05%	
Elusimicrobiota	Elusimicrobia bacterium	37%	4.00E-11	23.76%	
Fermentibacterota	Candidatus Fermentibacteria bacterium	20%	6.00E-08	25%	
Fibrobacterota	Fibrobacter sp.	95%	7.00E-122	33.88%	NLL14634.1
Firestonebacteria	Candidatus Firestonebacteria bacterium RIFOXYA2_FULL_40_8	39%	3.00E-28	30.03%	
Firmicutes	Firmicutes bacterium ADurb.Bin419	99%	0.00E+00	77.33%	Lhr core
Fusobacteriota	Fusobacteria bacterium	97%	0.00E+00	47.65%	Lhr core
Gemmatimonadota	Gemmatimonadetes bacterium	96%	1.00E-154	36.72%	NJD08932.1
Goldbacteria	Candidatus Goldbacteria bacterium HGW-Goldbacteria-1	39%	4.00E-24	28.90%	
Hydrogenedentota	Candidatus Hydrogenedentes bacterium	94%	1.00E-130	34.73%	NLT59800.1

Latescibacterota	Candidatus Latescibacteria bacterium	94%	3.00E+152	38.20%	MBN1292824.1
Lindowbacteria	Candidatus Lindowbacteria bacterium RIFCSPLOWO2_12_FULL_62_27	5%	2.60E-01	31.37%	
Margulisbacteria	Candidatus Margulisbacteria bacterium	98%	0.00E+00	51.93%	Lhr core
Marinisomatota	Candidatus Marinimicrobia bacterium	99%	0.00E+00	44.32%	Lhr core
Methylomirabilota	candidate division NC10 bacterium	96%	2.00E-138	34.07%	PWB45355.1
Nitrospinota	Nitrospinae bacterium	98%	7.00E-137	34.32%	MYA97065.1
Nitrospirota	Nitrospiraceae bacterium	99%	0.00E+00	52.25%	NOT21783.1
Omnitrophota	Candidatus Omnitrophica bacterium	96%	7.00E-134	34.04%	NPU96047.1
Patescibacteria	Parcubacteria group bacterium	99%	0.00E+00	54%	Lhr core
Planctomycetota	Phycisphaerae bacterium	98%	2.00E-172	37.88%	MBC8088584.1
Poribacteria	Candidatus Poribacteria bacterium	41%	3.00E-74	40.98%	
Proteobacteria	Gammaproteobacteria bacterium	79%	0.00E+00	47.10%	MBU2049169.1
Rifluebacteria	Candidatus Rifluebacteria bacterium	18%	4.00E-08	31.68%	

Schekmanbacteria	Candidatus Schekmanbacteria bacterium RBG_16_38_11	3%	2.00E-01	55.17%	
Spirochaetota	Spirochaetes bacterium RBG_16_49_21	98%	5.00E-139	34.41%	OHD70879.1
Sumerlaeota	Candidatus Sumerlaea chitinovorans	45%	2.00E+25	26.63%	
Synergistota	Synergistaceae bacterium	40%	1.00E-33	31.68%	
Tectomicrobia	Candidatus Entotheonella palauensis	40%	3.00E-27	29.41%	
Thermotogota	Thermotogae bacterium	65%	4.00E-106	37.74%	lhr core
Unclassified	Parcubacteria group bacterium	99%	0.00E+00	54%	Lhr core
Verrucomicrobiota	Verrucomicrobiales bacterium	97%	5.00E-180	36.09%	Lhr core
Wallbacteria	Candidatus Wallbacteria bacterium	37%	2.00E-10	25.83%	MBW4032335.1
Zixibacteria	Candidate division Zixibacteria bacterium	40%	3.00E-26	27.05%	OLD99479.1

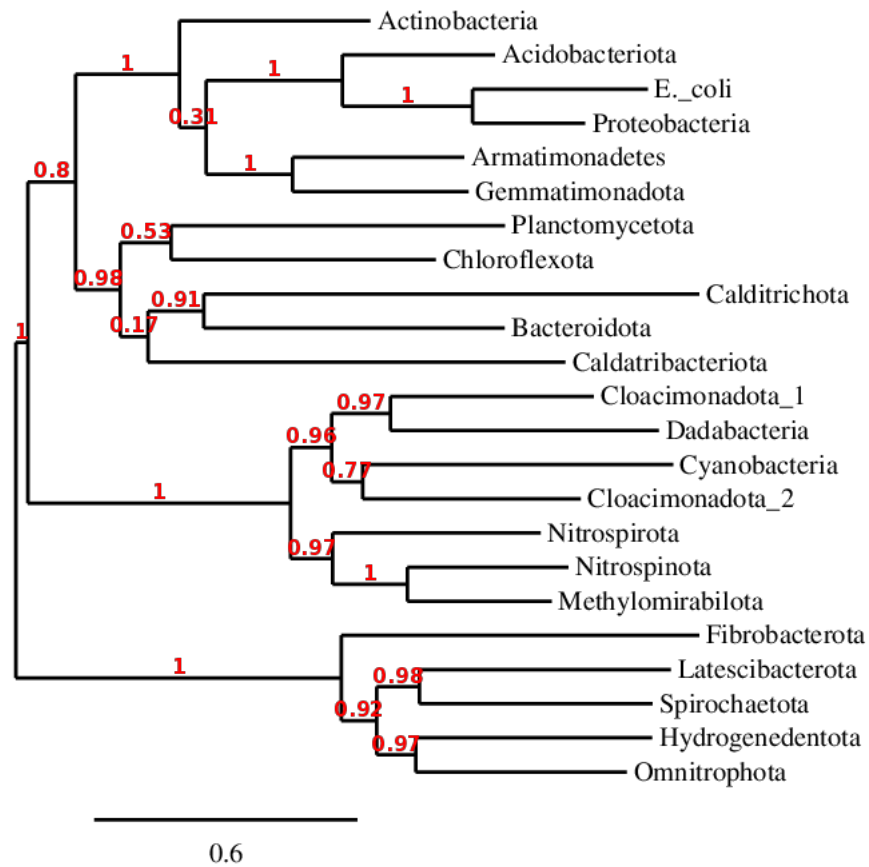


Figure 8.7 Analysis of Lhr-extended proteins among bacterial phyla when present.

Blastp identified Lhr proteins of interest were subject to “one click” phylogenetic analysis on ‘phylogeny.fr’¹⁴⁴. Analysis performed using proteins with quoted ascension numbers in table 8.1. Tree generation facilitated by PhyML¹⁴⁹ and annotated by TreeDyn¹⁵⁰.

8.3 H₂O₂ growth curves

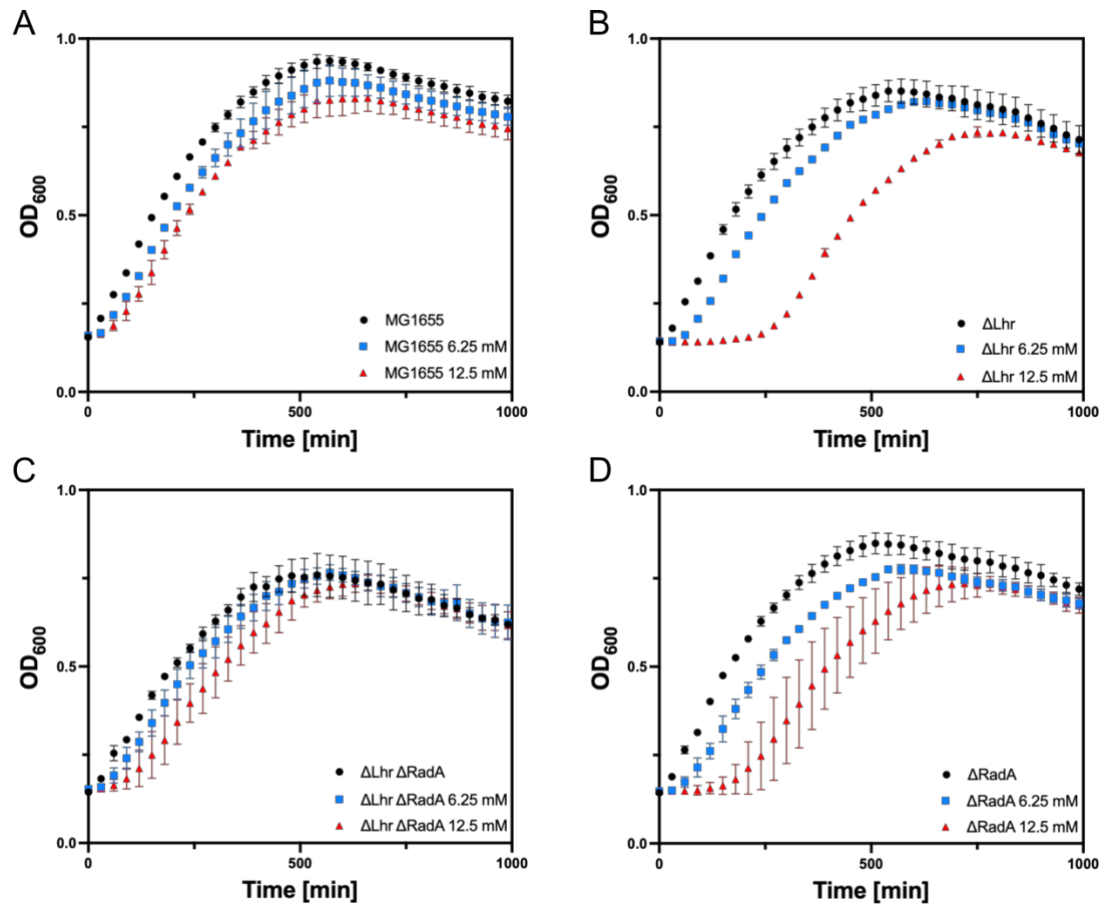


Figure 8.8 Effect of hydrogen peroxide on *E. coli* knockout cell strains full results.

For all growth in absence of H₂O₂ (**black circle**), 6.25 mM H₂O₂ (**blue square**) and 12.5 mM H₂O₂ (**red triangle**). (A) MG1655 wild type, (B) Δ lhr, (C) Δ lhr Δ radA, and (D) Δ radA. Graphs show data plotted from at least two repeats with error bars depicting standard error from mean.

8.4 *In vitro* analysis of *E. coli* Lhr gel examples

8.4.1 *E. coli* Lhr activity on d-uracil duplex DNA

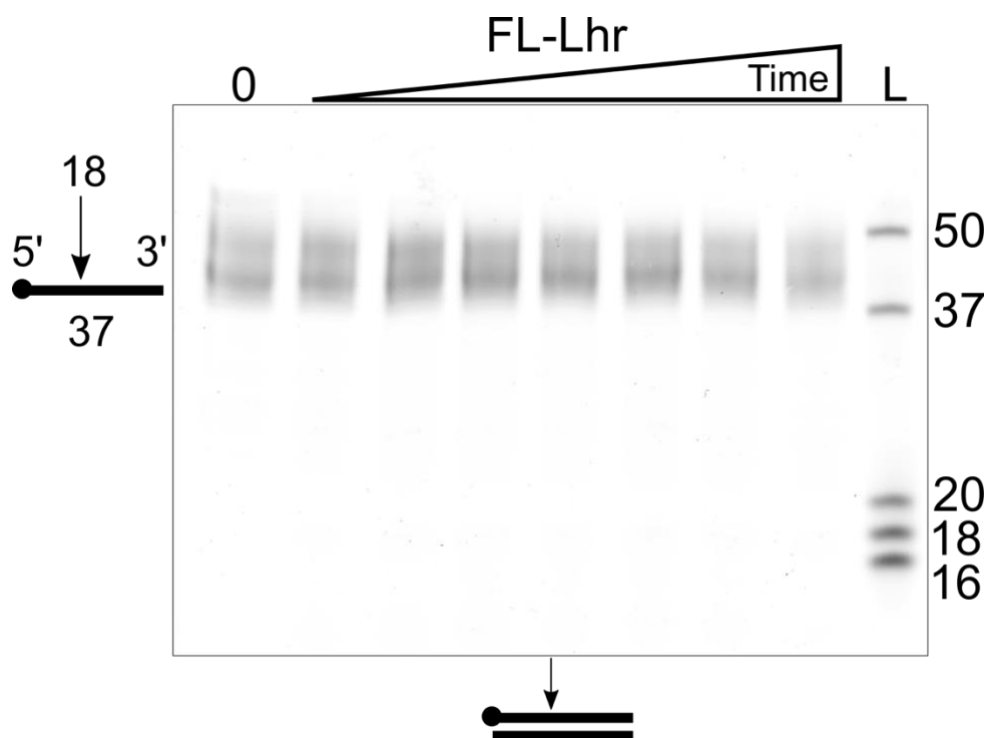


Figure 8.9 *E. coli* Lhr glycosylase activity on a d-uracil duplex DNA substrate.

15% denaturing acrylamide TBE gel showing FL-Lhr (80 nM) glycosylase activity on 12.5 nM of 5' Cy5 labelled d-uracil containing duplex DNA. Reaction was measured over 30 minutes with samples taken at 5 minute intervals.

8.4.2 *E. coli* Lhr-CTD activity on d-uracil flayed duplex and dsDNA

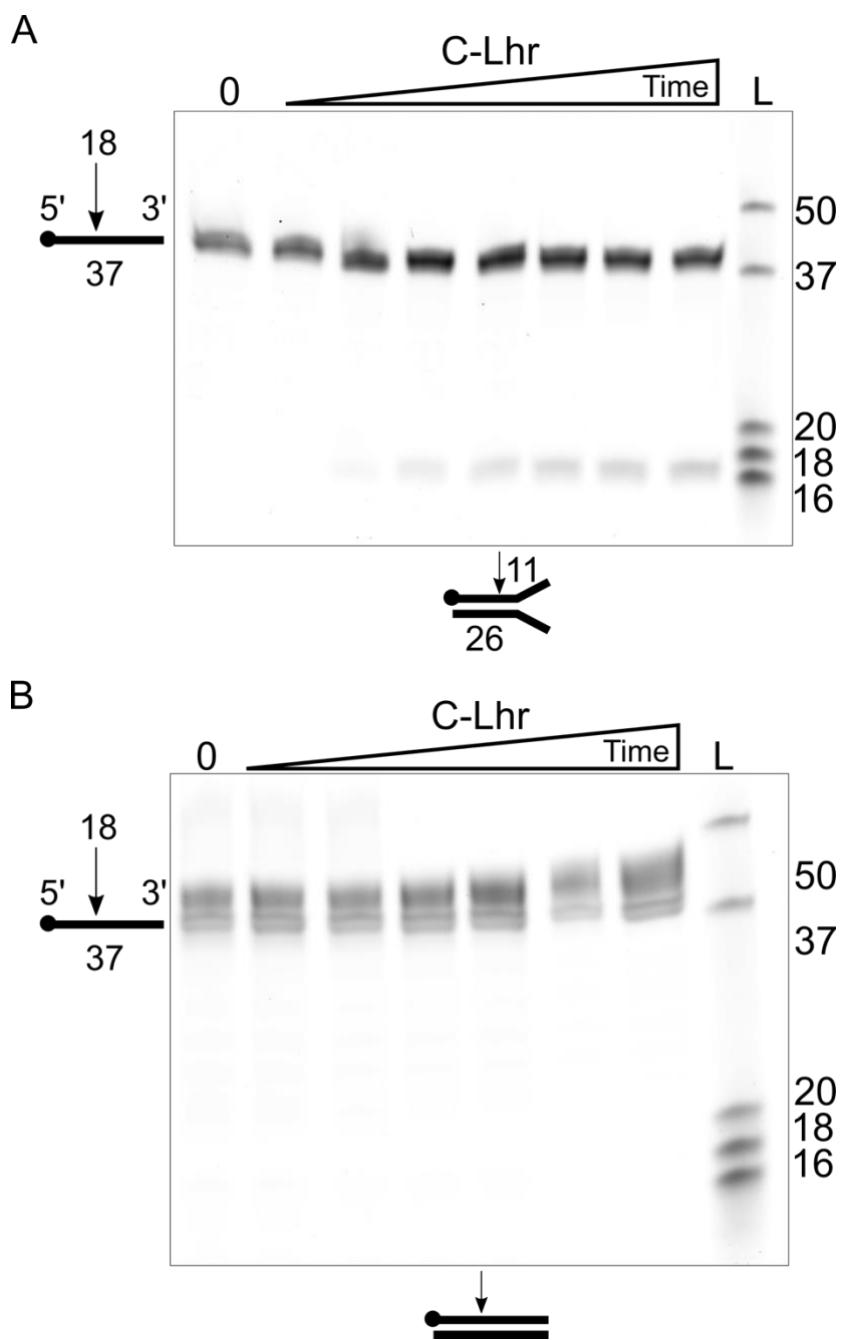


Figure 8.10 *E. coli* Lhr-CTD glycosylase activity on damaged flayed duplex and dsDNA substrates.

15% denaturing acrylamide TBE gels showing Lhr-CTD (80 nM) glycosylase activity on 12.5 nM of 5' Cy5 labelled d-uracil containing flayed duplex (A) and dsDNA (B). Reaction was measured over 30 minutes with samples taken at 5 minute intervals.

8.5 *E. coli* RadA (Sms) protein purification

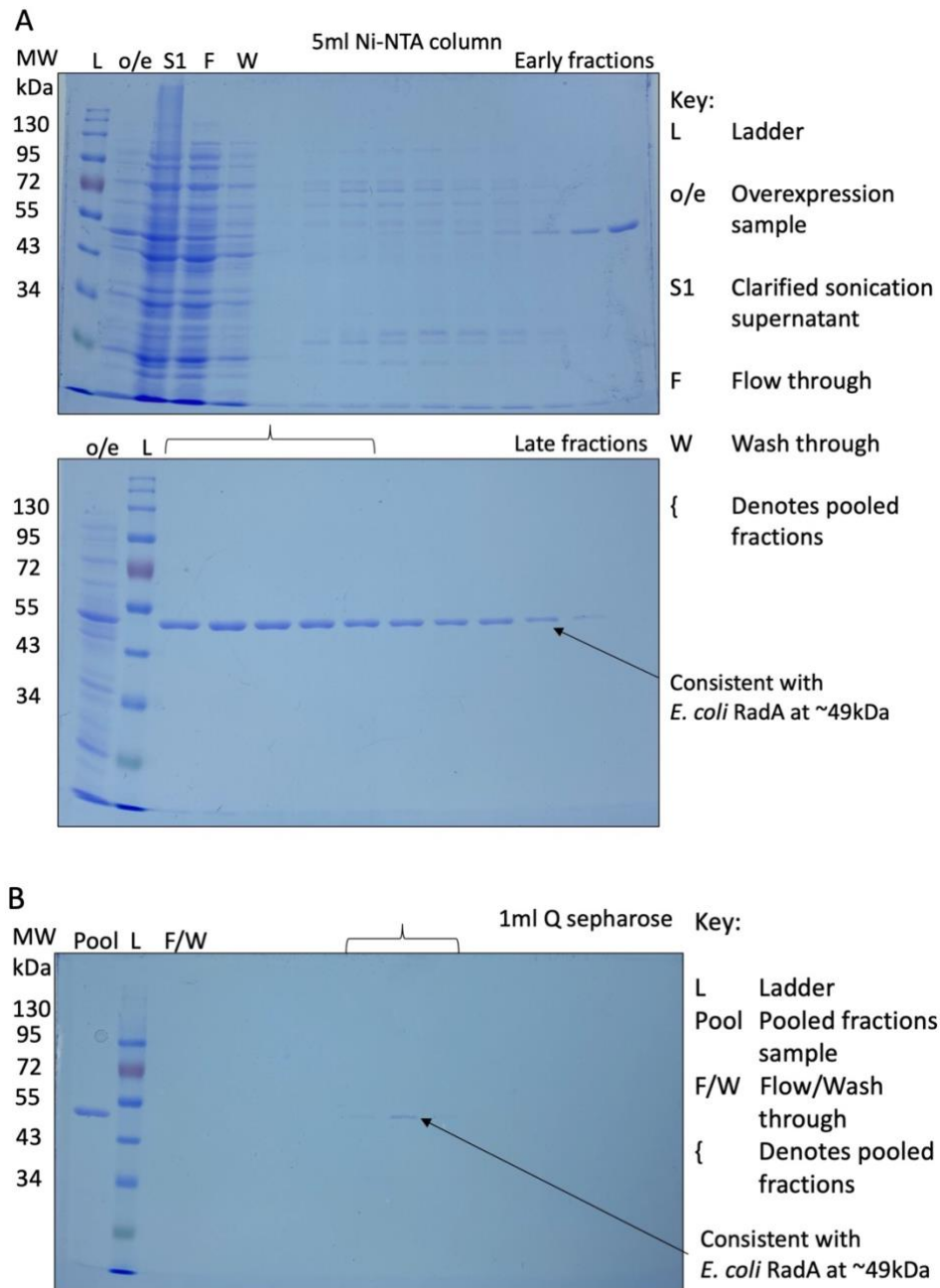


Figure 8.11 Purification of *E. coli* RadA (Sms).

Coomassie stained 10% acrylamide SDS PAGE analysis of RadA (**Sms**) purification using affinity chromatography. (**A**) Initial separation by Ni-NTA affinity chromatography. Purest fractions were pooled for dialysis and concentrated by Q-sepharose ion exchange chromatography. (**B**) Q-sepharose affinity chromatography purified RadA (**Sms**). Fractions were pooled and dialysed into storage buffer. RadA (**Sms**) of expected 49 kDa as highlighted.

8.6 *E. coli* RNaseT protein purification

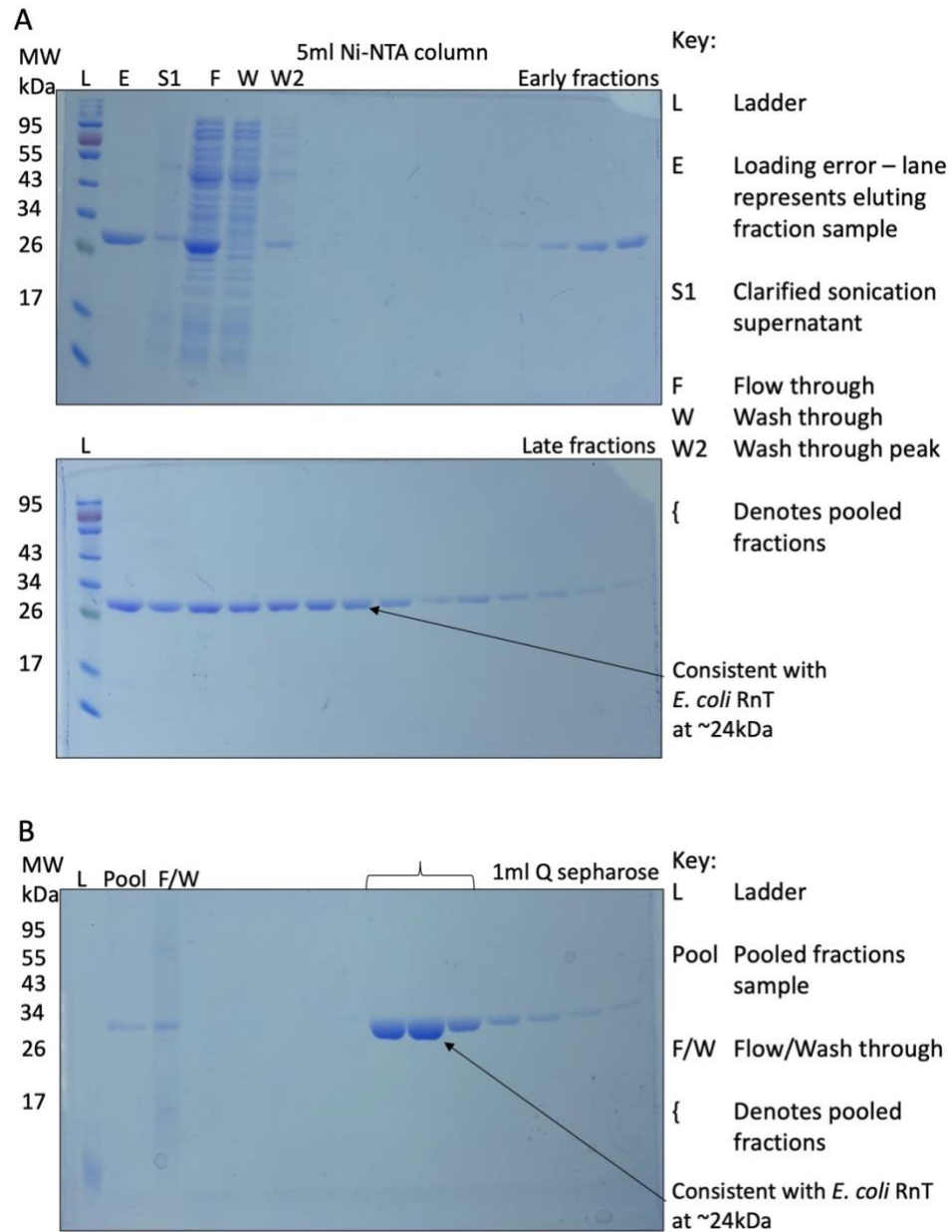


Figure 8.12 Purification of *E. coli* RNaseT.

Coomassie stained 12% acrylamide SDS PAGE analysis of RNaseT (RnT) purification using affinity chromatography. **(A)** Initial separation by Ni-NTA affinity chromatography. Purest fractions were pooled for dialysis and concentrated by Q-sepharose ion exchange chromatography. **(B)** Q-sepharose affinity chromatography purified RNaseT. Fractions were pooled and dialysed into storage buffer. RNaseT of expected 24 kDa as highlighted, please note RNaseT protein runs at 27 kDa due to the N-terminal His₆-thrombin site.

8.7 *In vitro* analysis of *E. coli* RadA (Sms) and RNaseT

8.7.1 *E. coli* RNaseT is a DNA nuclease

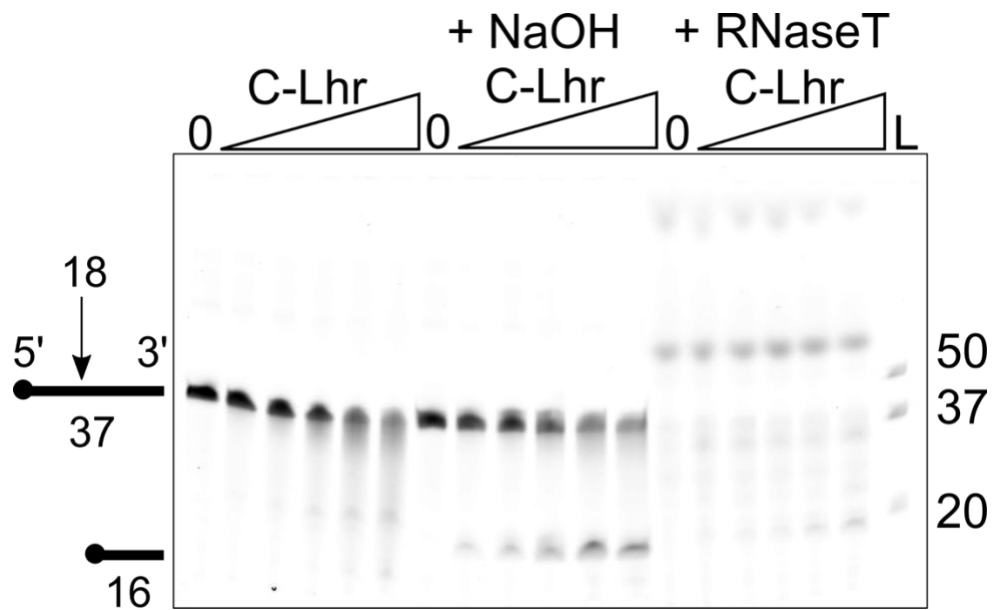


Figure 8.13 *E. coli* RNaseT acting as a potent nuclease.

Investigation into RNaseT ability to act as an AP ligase when presented with Lhr-glycosylated product. RNaseT addition (right third) causes complete degradation of DNA by its nuclease activity. Additional bandage of variable sizes is seen as Lhr-CTD activity/protein concentration increases. 15% denaturing acrylamide TBE gel containing 12.5 nM of 5' Cy5 labelled d-uracil containing ssDNA substrate. Lhr-CTD concentration increases from 50 to 800 nM linearly. RNaseT was added at 100 nM after 30 minutes of Lhr-CTD activity when present. Reactions occurred at 37°C.

8.7.2 *E. coli* RadA (Sms) EMSA

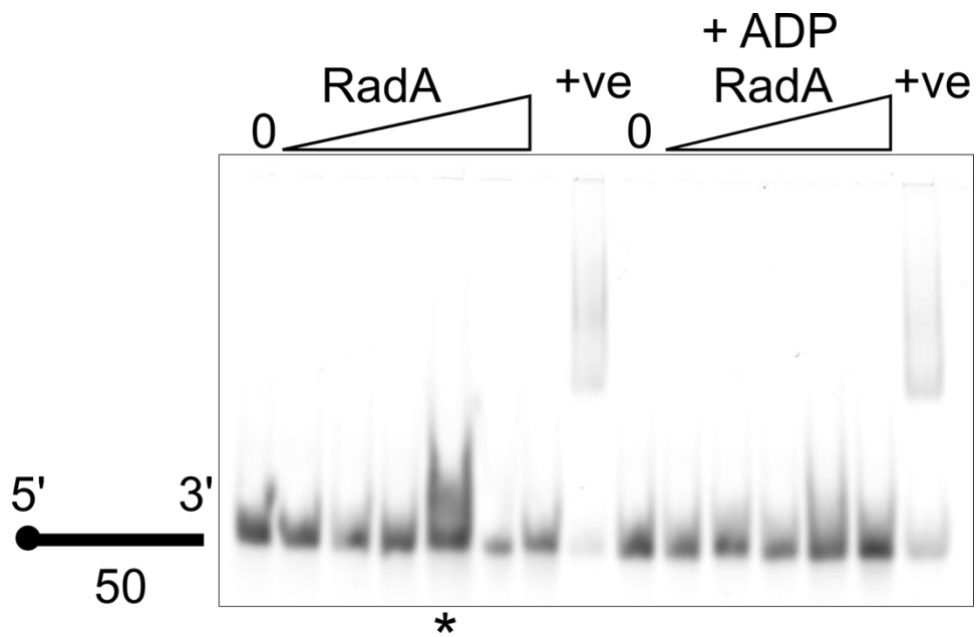


Figure 8.14 *E. coli* RadA (Sms) EMSA in the presence and absence of ADP.

5% Native acrylamide EMSA gel showing binding abilities of RadA (Sms) on 12.5 nM of 5' Cy5 labelled ssDNA substrates. Protein concentrations increase from 25 to 200 nM linearly. DNA binding is limited with only mild streaking at 150 and 200 nM protein. '+ve' is *MthHel308* positive DNA binding control. '*' indicates a double loading error. Reactions occurred at 37°C of 20 minutes.

Chapter 9 Publications

Proof Delivery Form

Journal and Article number: BCJ-2020-0379

Number of pages (not including this page): 13

Biochemical Journal

Please check your proof carefully to ensure (a) accuracy of the content and (b) that no errors have been introduced during the production process.

- You are responsible for ensuring that any errors contained in this proof are marked for correction before publication. Errors not marked may appear in the published paper.
- Corrections should only be for typographical errors in the text or errors in the artwork; substantial revision of the substance of the text is not permitted.
- Please answer any queries listed below.
- If corrections are required to a figure, please supply a new copy.

Your proof corrections and query answers should be returned as soon as possible (ideally within 48 hours of receipt). Please upload your corrected proof and any additional files (e.g. artwork) via the online proof review page from which you downloaded this file. You can also provide any specific instructions or comments in the 'Response Comments' box on the online page.

Notes:

1. Please provide the paper's reference number in any correspondence about your article
2. If you have any queries, please contact the publisher by email (production@portlandpress.com) or by telephone +44 (0)20 7685 2410

Supplementary Material:



This proof does not contain any supplementary material. If supplementary content is associated with this article it will be published, in the format supplied by the authors, with the online version of the article.

Queries for author:

- Q1: Please review the highlighting of the author surnames. Bearing in mind that this will direct how the authors are indexed by the journal and PubMed, please confirm if this is correct or indicate any changes that are needed.
- Q2: Please provide city name for affiliations 1 and 3.
- Q3: Please confirm that this statement is an accurate reflection of any competing interests (or lack thereof) of the author(s).
- Q4: Please confirm that all sources of funding (including all relevant grant numbers) have been acknowledged in Funding section.
- Q5: Please provide doi or PMID for ref. [14].
-

Research Article

Mechanistic insights into Lhr helicase function in DNA repair

Ryan J. Buckley^{1,*}, Kevin Kramm^{2,*},  Christopher D. O. Cooper³, Dina Grohmann² and  Edward L. Bolt¹

¹School of Life Sciences, University of Nottingham, U.K.; ²Institute of Microbiology and Archaea Centre Biochemistry, Single-Molecule Biochemistry Lab, University of Regensburg, 93053 Regensburg, Germany; ³Department of Biological and Geographical Sciences, School of Applied Sciences, University of Huddersfield, U.K.

Correspondence: Edward L. Bolt (ed.bolt@nottingham.ac.uk) or Dina Grohmann (dina.grohmann@ur.de)



The DNA helicase *Large helicase-related* (Lhr) is present throughout archaea, including in the Asgard and Nanoarchaea, and has homologues in bacteria and eukaryotes. It is thought to function in DNA repair but in a context that is not known. Our data show that archaeal Lhr preferentially targets DNA replication fork structures. In a genetic assay, expression of archaeal Lhr gave a phenotype identical to the replication-coupled DNA repair enzymes Hel308 and RecQ. Purified archaeal Lhr preferentially unwound model forked DNA substrates compared with DNA duplexes, flaps and Holliday junctions, and unwound them with directionality. Single-molecule FRET measurements showed that binding of Lhr to a DNA fork causes ATP-independent distortion and base-pair melting at, or close to, the fork branchpoint. ATP-dependent directional translocation of Lhr resulted in fork DNA unwinding through the ‘parental’ DNA strands. Interaction of Lhr with replication forks *in vivo* and *in vitro* suggests that it contributes to DNA repair at stalled or broken DNA replication.

Introduction

Lhr (*Large helicase-related*) protein is an ATP-dependent DNA translocase and helicase that forms a distinct group within Superfamily 2 helicases [1,2]. Lhr was discovered and named in bacteria [2], in which it is present in eight of ~30 phyla [2,3]. It is widespread in archaea [4], and the archaeal Lhr is a sequence homologue of the DDX-family of uncharacterized putative helicases found in eukaryotes including in humans [5–7]. Archaeal and bacterial Lhr proteins show high amino acid sequence identity (typically ~30%) between their N-terminal 800–900 amino acids, which is referred to as the ‘Lhr-Core’, that comprises their helicase domains [8]. Bacterial Lhr is extended to 1300–1500 amino acids by a region of unknown function that lacks obvious sequence homologues. Biochemical analysis of the Lhr-Core from the bacteria *Mycobacterium smegmatis* and *Pseudomonas putida* identified ATP-dependent ssDNA translocation with 3′ to 5′ directionality [9,10]. A crystal structure of bacterial Lhr-Core highlights significant similarities with the archaeal DNA repair helicase Hel308 [9,11], most notably in the orientation and interaction of its winged helix domain (WHD) with RecA-like domains typical of Ski2-like helicases [12,13].

Lhr-Core is conserved in many archaea and bacteria, in a genomic context adjacent to a manganese-dependent phosphodiesterase (MPE), an enzyme with active site architecture resembling Mre11 [8]. In other bacteria, full-length Lhr frequently occurs adjacent to the gene encoding RNaseT, which has roles in DNA repair and RNA maturation [14,15]. Deletion of the Lhr-Core gene (Saci_1500) in the archaeon *Sulfolobus acidocaldarius* resulted in a mild, ~4-fold, sensitivity to UV irradiation in comparison with wild type cells [16]. In contrast, genetic analysis of Lhr in *E. coli* revealed a phenotype in cells treated with the replication inhibitor AZT — deletion of gene *lhr* was synergistic with deletion of the gene encoding the replication-recombination-repair protein RadA [17]. These observations, and reported 4-fold up-regulation in transcription of *lhr* in *M. tuberculosis* in response to mitomycin C [18], suggest that Lhr may be part of a prokaryotic replication-coupled

*These authors contributed equally to this work.

Received: 12 May 2020
Revised: 23 July 2020
Accepted: 23 July 2020

Accepted Manuscript online:
24 July 2020
Version of Record published:
0 Month 2020

DNA repair pathway. In this work we investigated the properties of Lhr protein from archaea, a homologue of the eukaryotic DDX proteins. We provide evidence that archaeal Lhr interacts with stalled DNA replication, and that the purified Lhr protein has a preference for targeting forked DNA, remodelling it at the fork branch-point prior to its dissociation.

Materials and methods

Molecular cloning of archaeal Lhr

The *lhr* gene (open reading frame mt_1802) from the euryarchaeon *Methanothermobacter thermautotrophicus* (Mth) was first cloned into pBluescript using *SalI* and *XbaI* restriction endonuclease sites (pEB307) after PCR amplification from Mth genomic DNA (a kind gift from Prof. James Chong, University of York). The Mth *lhr* gene contains an internal *NdeI* restriction site that was altered by silent mutation using QuikChange II site-directed mutagenesis (Agilent). This allowed sub-cloning through a second PCR amplification into pET22b and pT7-7 using *NdeI* and *EcoRI* restriction sites (respectively, pEB352 and pEB353). DNA sequences of these constructs were verified to confirm that plasmids were suitable for protein expression and genetic analysis in *E. coli*.

Genetic analysis of archaeal Lhr

The basis and details for the genetic assay using *E. coli* strain *dnaE486 ΔrecQ* (Figure 1) are detailed in reference [19]. *E. coli* cells were transformed with empty plasmid vector pT7-7, or with pT7-7 constitutively expressing either bacterial RecQ as a control [20], verified helicases from *M. thermautotrophicus* — Hel308 [19], Cas3 [21] and Hef [22] – or putative archaeal helicases, also from *M. thermautotrophicus* — mt1347 and mt0203. Transformed cells were grown in a shaking water bath at 30°C from colonies inoculated in LB broth containing ampicillin (50 µg/ml), until OD₆₀₀ of 0.5. Then 100 µl of culture was spread onto a sector of each agar ampicillin plate for incubation at 30°C, 37°C or 42°C.

Purification of archaeal Lhr protein

Plasmid pEB352 was transformed into *E. coli* strain BL21 Codon+ (Agilent) for overexpression of Lhr protein from the archaeon *Methanothermobacter thermautotrophicus*. An overnight culture of this (20 ml) was added to 2 L of LB-ampicillin (50 µg/ml) and chloramphenicol (10 µg/ml) and grown at 30°C with shaking in baffled flasks. At an O.D₆₀₀ of 0.5, Lhr expression was induced by addition of isopropyl-β-D-thiogalactopyranoside (IPTG, 0.8 mM) and growth was continued for a further two hours. Harvested cells were resuspended in buffer C (20 mM Tris.HCl pH 8.0, 10% glycerol, 100 mM potassium chloride and 2 mM DTT) for –80°C storage.

To purify Lhr protein the biomass was thawed on ice, sonicated and clarified by centrifugation. Soluble protein supernatant was loaded in buffer C into a 5 ml HiTrap Heparin column, and Lhr was eluted in a linear gradient of 0.1–1.5 M potassium chloride in buffer C at ~0.7–0.9 M. Peak Lhr fractions were pooled and loaded directly onto a 16/60 sephacryl S200 column in buffer C, and peak fractions were pooled and dialyzed overnight in buffer C. Dialyzed Lhr was loaded onto a 1 ml HiTrap Q sepharose column and was eluted in a linear gradient of 0.1–1.5 M potassium chloride in buffer C, at ~0.6–0.8 M potassium chloride. Peak Lhr fractions were pooled and dialyzed into buffer C containing 35% glycerol, and stored as aliquots after flash-freezing for storage at –80°C.

Preparation of DNA substrates for helicase and DNA binding assays

Nucleotide sequences used to generate all substrates are given in Supplementary Table S1. One DNA strand (900 ng in a 20 µl reaction volume) for each substrate was 5'-end labelled with ³²P using T4 polynucleotide kinase and γ³²-P-ATP. The radio-labelled DNA strand was separated from unincorporated γ³²-P-ATP using a BioSpin 6 column and the resulting labelled DNA was mixed with 900 ng of each appropriate unlabelled strand in 1× SSC buffer (150 mM sodium chloride, 15 mM sodium citrate at pH 7.0), heated to 95°C for 5 min and allowed to anneal by cooling overnight to room temperature. Resulting DNA was mixed with gel loading dye and loaded onto a 10% TBE gel for electrophoresis at 150 volts for 2 h. The gel was then exposed to autoradiography film and the developed film revealed the positions of the desired substrates for excision from the gel. DNA was eluted from excised gel slices by soaking overnight at 4°C in 20 mM Tris.HCl pH 7.5 containing 20 mM sodium chloride. DNA in buffer recovered from gel debris was quantified by scintillation counting using as standards the scintillation counts of samples taken throughout the procedure that were of known

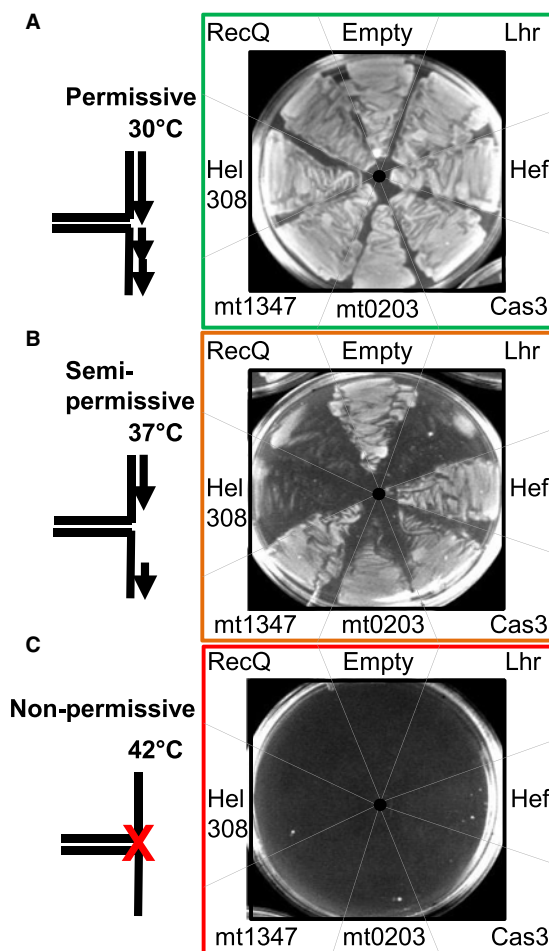


Figure 1. Archaeal Lhr interacts with stalling replication forks in *E. coli dnaE486 ΔrecQ* cells.

Panels are colour-coded to illustrate temperatures at which replication is unhindered (permissive, 30°C), destabilized (semi-permissive, 37°C) or stopped (non-permissive, 42°C). For each temperature cells were spread onto ampicillin agar after expressing the protein indicated from a plasmid. (A) At 30°C cells replicate normally resulting in fully viable growth in each sector. (B) At 37°C replication is destabilized by the *dnaE486* allele [20]. Hel308 and RecQ, gave inviability phenotypes as expected [19,20,31], and Lhr gave the same phenotype. (C) At 42°C the *dnaE486* allele makes cells inviable — this is used as a control that *dnaE486* suppressor mutations have not arisen.

DNA mass (ng). This established the final yield of substrate DNA in ng that was converted to a final concentration of DNA (nM) for use in assays.

Helicase assays and EMSAs

See Supplementary Table S1 for substrates. Helicase reactions were in buffer HB (20 mM Tris.HCl pH 7.5, 2 mM DTT, 100 µg/ml BSA and 7% glycerol) supplemented with 2 mM ATP (at pH 7.5) and 1 mM magnesium chloride. Helicase assays were at 45°C for either 20 min or in reactions over a time course as shown. Reactions were stopped by addition of de-proteinising buffer (1× is 0.625% SDS, 50 mM EDTA and 2.5 mg/ml proteinase K) and gel loading dye was added prior to electrophoresis at 150 volts for 1 h through a 10% acrylamide TBE gel. Assay products were imaged on a storm™ scanner (Amersham) from phosphorimaging screens, after drying the gels under a vacuum on a flatbed gel dryer. Assay products were quantified from TIF files of gel images using the GelEval software. For EMSAs, Lhr (100 nM) was mixed with DNA (10 nM) in buffer HB at room temperature with reactions loaded directly onto a 5% acrylamide TBE gel and were imaged using the ChemiDoc MP imaging system (Bio-Rad).

Assays using fluorescent DNA fork-2 and confocal single-molecule FRET measurements

Fluorescent fork-2 DNA was formed from the four fork-1 oligonucleotides (Supplementary Table S1) mixed in equimolar concentration (10 μ M) in annealing buffer (10 mM Tris-HCl pH 7.8, 50 mM NaCl, 1 mM EDTA), heated to 95°C for 3 min and cooled to room temperature (23°C) over 1.5 h. DNA was stored at -20°C. For EMSAs, Lhr (100 nM) was mixed with DNA (10 nM) in buffer HB at room temperature with addition of ATP and magnesium chloride (1 : 2 mM) as indicated in Figure 3, and reactions loaded directly onto a 5% acrylamide TBE gel. Gels were imaged using the ChemiDoc MP imaging system (Bio-Rad).

Prior to FRET measurements, the sample chambers (Cellview slide, Greiner Bio-One) were passivated with 2 mg/ml BSA in 10 mM Tris-HCl pH 8 for 10 min and washed once with Millipore water. For formation of complexes, 1 nM DNA, 1 μ M LHR, 1 mM MgCl₂ and 2 mM ATP were mixed in H78 buffer (20 mM NaHEPES pH 7.8, 10% (v/v) glycerol, 100 mM potassium acetate, 1 mM EDTA, 2 mM DTT) and incubated for up to 20 min at room temperature or 45°C. Afterwards, samples were diluted by a factor of 10 in H78 buffer and added to the sample chamber.

Single-molecule fluorescence of diffusing complexes was detected with a MicroTime 200 confocal microscope (PicoQuant) equipped with pulsed laser diodes (532 nm: LDH-P-FA-530B; 636 nm: LDH-D-C-640; PicoQuant/cleanup filter: zet635; Chroma). The fluorophores were excited at 20 μ W using pulsed interleaved excitation (40 MHz). Emitted fluorescence was collected using a 1.2 NA, \times 60 microscope objective (UplanSApo \times 60/1.20W; Olympus) and focused through a 50 μ m confocal pinhole. A dichroic mirror (T635lpxr; Chroma) was used to separate donor and acceptor fluorescence. Additional bandpass filters (donor: ff01-582/64; Chroma; acceptor: H690/70; Chroma) completed spectral separation of the sample fluorescence. Each filtered photon stream was detected by an individual APD (SPCM-AQRH-14-TR, Excelitas Technologies) and recorded by a time-correlated single photon counting (TCSPC) capable HydraHarp 400 (PicoQuant).

FRET data analysis

Data analysis of confocal FRET measurements was performed with the software package PAM [23]. Photon bursts of diffusing molecules were selected based on an all-photon burst search (APBS, parameters: L = 100, M = 10, and T = 500 μ s) and an additional dual-channel burst search (DCBS, parameters: L = 100, M_{GG+GR} = 20, M_{RR} = 20, and T = 500 μ s).

For an APBS, the FRET efficiency of each burst (calculated as proximity ratio E_{PR}) and the raw stoichiometry factor S_{raw} was calculated as:

$$E_{PR} = \frac{N_{DA}}{N_{DD} + N_{DA}} \quad (1)$$

$$S_{raw} = \frac{N_{DD} + N_{DA}}{N_{DD} + N_{DA} + N_{AA}} \quad (2)$$

where N_{DD} , N_{DA} and N_{AA} are the number of detected photons. Indices refer to donor donor emission upon donor excitation (DD), acceptor emission upon donor excitation (DA) and acceptor emission upon acceptor excitation (AA). These were used to calculate the donor leakage and direct excitation correction factors. For DCBS, the FRET efficiency E and the stoichiometry factor S of each burst were calculated as:

$$E = \frac{N_{DA} - (c_{leak} \cdot N_{DD} + c_{dir} \cdot N_{AA})}{\gamma \cdot N_{DD} + N_{DA} - (c_{leak} \cdot N_{DD} + c_{dir} \cdot N_{AA})} \quad (3)$$

$$S = \frac{\gamma \cdot N_{DD} + N_{DA} - (c_{leak} \cdot N_{DD} + c_{dir} \cdot N_{AA})}{\gamma \cdot N_{DD} + N_{DA} + \beta \cdot N_{AA} - (c_{leak} \cdot N_{DD} + c_{dir} \cdot N_{AA})} \quad (4)$$

where c_{leak} is the correction factor for donor leakage, c_{dir} is the correction factor for direct excitation of the acceptor, γ and β are the detection and excitation correction factors. Burst data were corrected for donor leakage and direct excitation of the acceptor (determined from APBS according to [24], as well as γ and β (determined from DCBS ES-histograms using an internal fit on multiple E/S separated FRET populations). The data were binned (bin size = 0.025), plotted as E histogram and fitted with a single (DNA) or multiple Gaussian fits using the Origin software.

The inter-fluorophore distance r was calculated from corrected E values according to:

$$r = R_0 \sqrt[6]{\frac{1-E}{E}} \quad (5)$$

using the following Förster radius: $R_0 = 5.9$ nm of the ATTO 532-ATTO 647N dye pair.

Analysis of Lhr and DDX52 structures

Protein sequence homology was assessed using BLASTP [25] against sequences with a Protein DataBank [26] record, using the *Methanothermobacter thermautotrophicus* Δ H open reading frame Mth1802 (UniProt: O27830) and human DDX52 (UniProt: Q9Y2R4) helicase protein sequences as search queries. Protein fold, secondary structure and structural homology searches were performed with Phyre2 [27] under Intensive mode. Predicted structure models were analyzed, superimposed and RMSD calculated with DALI [28], superimposing against the *M. smegmatis* Lhr [9] (PDB: 5V9X) helicase structure. Protein secondary structure was predicted in PSIPRED [29]. Structural models rendered in PyMOL were superimposed using the C_α chain.

Results

Genetic analysis of archaeal Lhr indicates interaction with stalled DNA replication

Lhr is distributed throughout the archaeal domain, including in all classes of the Asgardarchaeota that is most closely related to eukaryotes, and in the extremely reduced genomes of Nanoarchaeota — details are presented as Supplementary Data in Supplementary Table S2. We utilized Lhr from the euryarchaeal species *Methanothermobacter thermautotrophicus* (Mth), and first analyzed this Lhr using genetics. Two previous studies in archaea had deleted the *lhr* gene — *Haloflex volcanii* this gave no discernible phenotype in response to UV or γ irradiation [30], and in *Sulfolobus acidocaldarius* there was very modest (4-fold) UV sensitivity [16]. Here, we observed a robust phenotype for Mth Lhr in a genetic assay that detects interaction with stalled DNA replication [20] — it uses *E. coli* cells with a conditional mutation in the gene encoding DNA polymerase III (*dnaE*), the replicative polymerase. This particular mutation, *dnaE486*, causes structural instability of DNA polymerase III at 37°C that triggers stalling of DNA replication, mimicking DNA damage. Cells survive this by activating replication-coupled DNA repair, therefore 37°C is called a ‘semi-permissive’ temperature. However, interference with de-stabilised replication at 37°C by heterologously expressed protein causes low cell viability because native replication-coupled repair is impeded. This assay had previously identified DNA repair phenotypes for archaeal Hel308 and RecQ [19,20,31], and was re-visited to assess other putative archaeal helicases including Lhr (Figure 1). As expected from previous findings [19,20], expression of bacterial RecQ or Hel308 in these cells at permissive temperature (30°C) had no effect on viability (Figure 1A), indicating that these proteins are not toxic when expressed in *E. coli* cells replicating normally, but both caused inviability at 37°C indicating interaction with unstable replication (Figure 1B). Expression of Lhr also caused cell inviability at 37°C, and the normal viability of cells at 30°C confirmed that Lhr protein does not confer toxicity to normal replication. Expression of other known or putative archaeal helicases had no observable effect on cell viability at 37°C (Figure 1A,B). All cells were inviable at 42°C (Figure 1C), a temperature at which the replisome cannot function because of the *dnaE486* mutation — this ensures that suppressor mutations have not arisen to give false positive results at 37°C. This genetic analysis suggests that Lhr, like archaeal Hel308 and bacterial RecQ, interacts with de-stabilised replication forks. This information was taken forward for biochemical analysis of the Mth Lhr protein.

Archaeal Lhr protein preferentially targets fork-DNA for DNA translocation

The bacterial ‘core’ Lhr (Lhr-Core), which lacks a 700 amino acid C-terminal region present in the bacterial but not archaeal Lhr enzymes, is a ssDNA-stimulated ATPase that translocates ssDNA with 3′ to 5′ directionality [10]. Purified full-length archaeal Lhr (Supplementary Figure S1) was challenged with a gapped DNA duplex substrate to determine if it had similar properties (Figure 2A). In this assay, loading of Lhr onto ssDNA revealed 3′ to 5′ translocation directionality by displacement of the 32 nt strand in preference to the 21 nt strand (Figure 2B lanes 2 and 3). DNA unwinding of the gapped duplex by Lhr *in vitro* was most effective at 2 mM ATP and 1 mM magnesium chloride (Supplementary Figure S2), conditions that were used for

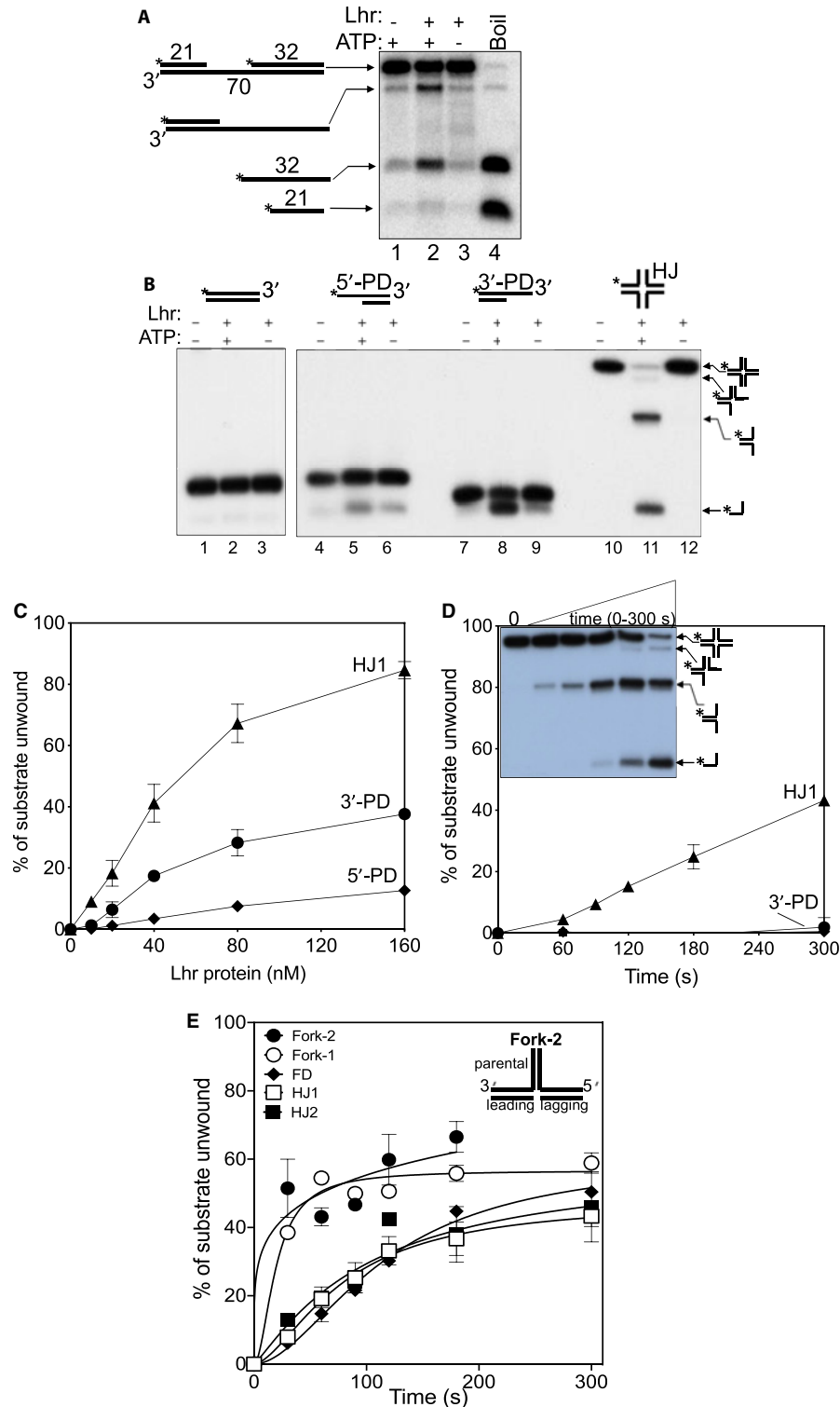


Figure 2. Lhr is most effective at unwinding branched DNA molecules.

Part 1 of 2

All parts show results of Lhr helicase reactions observed in TBE 10% acrylamide gels. Asterisks indicate 5'-³²P end labelling of a DNA strand and DNA was used at 10 nM unless stated. (A) Lhr (100 nM) gave ATP-dependent displacement of the 32 nt strand from the gapped duplex (1 nM) indicating 3' to 5' directionality. (B) Lhr (100 nM) did not significantly unwind fully base paired DNA duplex or a partial duplex with a 5' ssDNA tail (5'-PD, lanes 4–6), but unwound a partial duplex DNA with a 3'-ssDNA-tail (3'-PD, lanes 7–9). A Holliday junction (HJ) was unwound more effectively in this assay to generate three-strand,

Figure 2. Lhr is most effective at unwinding branched DNA molecules.

Part 2 of 2

two strand and ssDNA products as indicated at the side of the gel panel. The apparent proficiency of Lhr in unwinding the Holliday junction compared with partial duplex DNA Holliday junction was confirmed in part (C), in which Lhr was added to DNA at 10, 20, 80 and 160 nM as indicated. Reactions were repeated three times — the range of standard error is shown. (D) Holliday junction DNA (HJ) was unwound by Lhr (40 nM) at least 10-fold more effectively than the 3'-tailed partial duplex (3'-PD) over the course of time (0–300 s). The inset gel summarizes that Lhr unwound Holliday junctions into products that were further unwound, indicating that Lhr is not specific for targeting Holliday junctions. Reactions were carried out twice, and bars show the standard error. (E) Lhr (40 nM) unwound fork-1 and fork-2 DNA most effectively over time, compared with Holliday junctions. A cartoon of the fork-2 structure is shown for reference to the fork parental, leading and lagging strands. For comparison, the graph also shows a flayed duplex (FD) DNA, comprising the fork parental duplex but neither lagging nor leading strands. Details of the substrates are given in the Supplementary data along with representative gels. Reactions were done three times and bars show standard error from mean.

subsequent assays. We next assessed unwinding of different model synthetic DNA substrates to establish if Lhr had a substrate preference that could be used to gain insight into its DNA unwinding mechanism. In agreement with a requirement for ssDNA to trigger DNA translocation, Lhr did not unwind DNA in a fully base-paired DNA duplex (Figure 2B lanes 1–3). It was weakly active at unwinding a partial duplex with 25 nt of 5' tailed ssDNA (5'-PD, lanes 4–6) but substantially unwound a partial duplex with a 3' ssDNA tail (3'-PD, lanes 7–9). This is in agreement with the 3' to 5' directionality observed when unwinding the gapped duplex (Figure 2A), but some dissociation of the 5' tailed substrate suggested that Lhr may more generally distort DNA base-pairing, leading to low levels DNA strand dissociation, when bound to DNA – further investigation

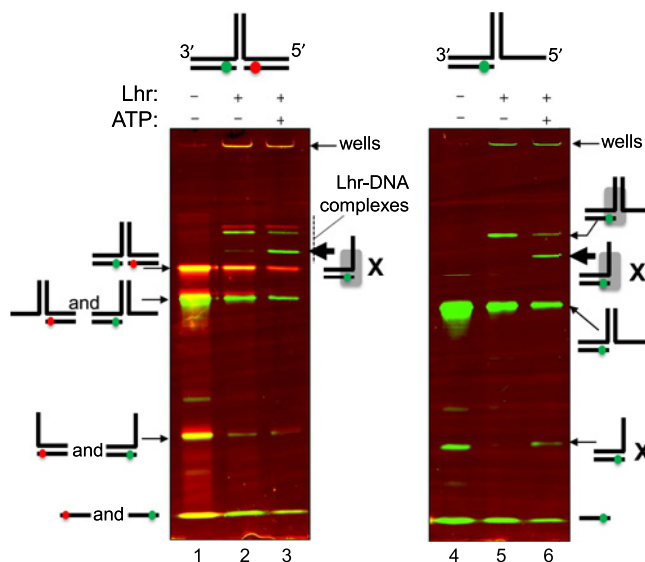


Figure 3. Evidence for directionality of fork dissociation by Lhr.

Both panels are from the same native acrylamide EMSA gel, divided to be able to annotate each part with substrate and product DNA. Fork-2 was labelled with ATTO 532 (green) and ATTO 647N (red) fluorophores at the indicated positions in the cartoon representations shown above the gels, a full fork corresponding to lanes 1–3, and a partial fork lacking a lagging strand corresponding to lanes 4–6. The fluorescence signal of the fluorophores was detected using a fluorescence scanner. Green and red bands in the gel correspond to fork-2 and fork-2-Lhr bound reaction products that contain one or both of the labelled DNA strands. Lanes 1 and 4 show bands corresponding to each full substrate as naked DNA, and its component intermediate DNA molecules; each form as shown to the side of the panel. Addition of Lhr (100 nM) and ATP-Mg²⁺ to reactions is indicated above each panel. Free substrate (lanes 1 and 4) was bound by Lhr (lanes 2 and 5), and indicated by the label for Lhr-DNA complexes and a grey rectangle denoting DNA-bound protein. Addition of ATP triggered fork dissociation into the products indicated with the letter X. Helicase products could remain bound to Lhr protein, as indicated by the grey rectangles representing Lhr. ATP-dependent formation of Lhr-bound two-strand DNA product is highlighted in lanes 3 and 6.

of this is presented later (Figure 4). Lhr also unwound a partial duplex comprising an RNA–DNA hybrid with a 3' single stranded 'tail' as well as the corresponding tailed DNA duplex (Supplementary Figure S3). These data indicate that Lhr requires single-stranded DNA (ssDNA) to trigger directional translocation/helicase activity.

Unwinding of the 3' tailed partial duplex (3'-PD) was quite modest — maximally 30% of substrate was unwound when Lhr was used at 10-fold molar excess over DNA (Figure 2C). Lhr unwound an equivalent branched substrate, a Holliday junction (indicated as HJ in the figures), 3-fold to 10-fold more effectively than tailed duplexes measured in, respectively, endpoint (Figure 2C) and time course assays (Figure 2D). This Holliday junction (HJ1) was generated by annealing of the same DNA strand, and its complements, that was used to generate the 3' tailed partial duplex to ensure DNA sequences were consistent, as detailed in Supplementary Table S1. Lhr generated two major products from unwinding of HJ1 — these products were identifiable as labelled in Figure 2D by comparing them with the single forked product generated by the Holliday junction specific helicase RuvAB (Supplementary Figure S4A), and with ssDNA product of Lhr unwinding the 3' tailed duplex (Figure 2B lanes 8 and 11, and Supplementary Figure S4B). The structural specificity of RuvAB for unwinding Holliday junctions to only a fork without further unwinding of the fork into ssDNA [32,33], therefore contrasted with Lhr, suggesting that Lhr may be able to target forked DNA for unwinding.

To narrow down the substrate preferences of Lhr *in vitro* we compared unwinding of forked DNA with Holliday junction DNA as a function of time (Figure 2E). Two different Holliday junctions were compared with equivalent forked DNA that comprised a fully base-paired 'parental' DNA duplex and leading and lagging strand duplexes of the same DNA sequences as Holliday junctions (Figure 2E and Supplementary Table S1). These assays, using 20 nM of DNA and 40 nM of Lhr protein, indicated modest preference for forked DNA compared with Holliday junctions (Figure 2E). Multiple products from Lhr unwinding Holliday junctions were again apparent (Supplementary Figure S5).

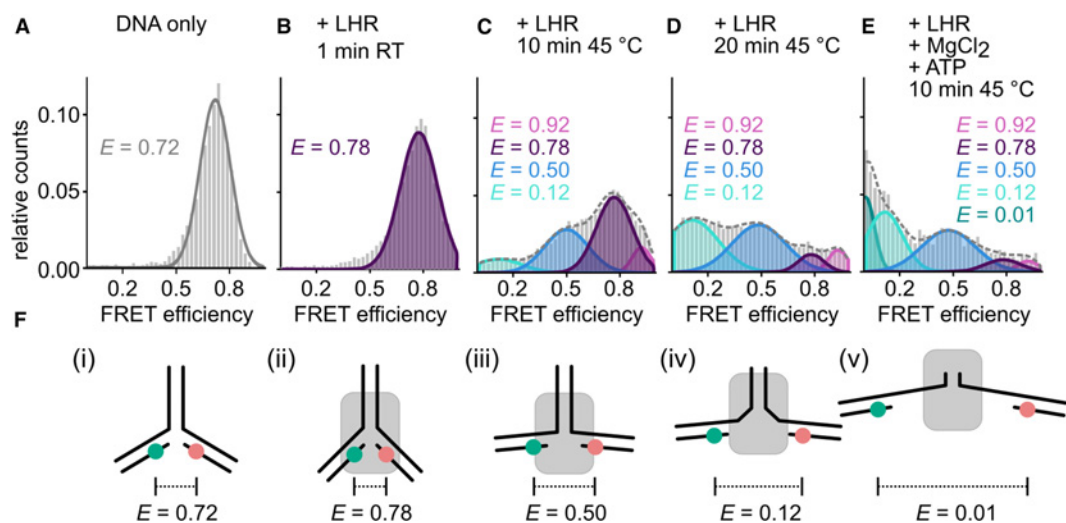


Figure 4. Single-molecule FRET analysis of conformational changes induced by Lhr on fork-2 DNA.

Measurements were performed on freely diffusing DNA/protein complexes to monitor Lhr-induced conformational changes on forked DNA substrate. (A) ATTO532 (donor) and ATTO647N (acceptor) labelled fork-2 in the absence of protein, (B) after addition of Lhr (1 μ M) at room temperature (RT), then (C and D) after 10 min and 20 min of incubation at 45°C. (E) After 10 min at 45°C with addition of 1 mM MgCl₂ and 2 mM ATP. FRET populations were fitted with multiple Gaussian distributions. The mean FRET efficiency *E* of the fitted peaks are shown. The dashed line is the cumulative fit curve. Each measurement was carried out at least three times — see Supplementary Figure S6). (F) Putative model for the mechanism of Lhr-dependent fork DNA unwinding. Conformations are based on inter-fluorophore distances derived from the measured FRET efficiencies that are presented as a data table in Supplementary Figure S6: i. relaxed conformation of fork DNA labelled with donor (green, leading strand) and acceptor (red, lagging strand); ii. compacted fork bound by Lhr (grey); iii. stretched conformation after heat activation of LHR; iv. partially melted fork DNA after ATP-Mg²⁺ addition; v. mostly unwound fork still bound to Lhr.

The preference of Lhr for forked DNA that we observed *in vitro* is consistent with Lhr targeting replication forks in genetic assays (Figure 1). But it raised the question, how does Lhr most effectively unwind fully base-paired forks, when it requires access to ssDNA for translocation leading to DNA unwinding? We reasoned targeting of a fork branch-point by Lhr may disrupt base pairing allowing ssDNA loading and translocation, which we investigated using single-molecule Förster resonance energy transfer (smFRET) measurements.

smFRET measurements reveal ATP-independent remodeling of a DNA fork by Lhr, and ATP-dependent dissociation of the fork-lagging strand

Lhr unwound model DNA forks most effectively in ensemble reactions *in vitro* (Figure 2E), therefore the same fork-2 substrate was used for smFRET analysis. Here, a donor-acceptor dye-pair was positioned in the fork lagging strand (ATTO 647N) and leading strand (ATTO 532) (Supplementary Table S1). We began by assessing Lhr binding and unwinding of this fork-2 in EMSAs, exploiting the dual ATTO labelling that allows for greater differentiation of reaction products than the single ³²P end-radiolabel (Figure 3 lanes 1–3). The reactions were not de-proteinised and consequently LHR in complex with either the complete fork substrate or unwinding intermediates was detected. In reactions lacking ATP, Lhr-fork DNA complexes were observed (Figure 3 lane 2). With ATP, Lhr helicase products primarily result from unwinding of the fork ‘parental’ DNA not fork lagging or leading strands, visible as a single product. The resulting green fluorescing DNA-LHR complex is consistent with the two-strand molecule indicated that would be generated by 3′ to 5′ directionality of Lhr (product X in lane 3). To verify this, we repeated the Lhr binding and unwinding reactions using a partial fork-2 that lacked the red fluorescing lagging strand (Figure 3 lanes 4–6). As expected in the absence of ATP, Lhr bound to the partial fork-2 resulting in a single complex representing Lhr-DNA binding (lane 5). Addition of ATP gave the same two-strand DNA product both bound with Lhr and not bound (both marked X in Figure 3 lanes 4–6), also consistent with the green ATTO labelled partial fork-2 being unwound 3′ to 5′ through the ‘parental’ duplex.

Having gained some qualitative insight into unwinding of the ATTO labelled fork-2 by Lhr we next assessed the effect of Lhr on DNA conformation within the fork at the single-molecule level (Figure 4), by determining the efficiency of energy transfer from donor to acceptor (E). Higher FRET efficiency (E) values denote shorter inter-dye distances giving a readout of fork conformation at the branchpoint. In the absence of Lhr, the fork DNA gave a single population (E = 0.72) (Figure 4A,Fi) representing a relaxed state with angles of ~130° between the parental, lagging and leading strand DNA. Addition of Lhr at room temperature in buffer without ATP-Mg²⁺ shifted the signal to E = 0.78, representing a shortening of the inter-dye distance due to fork compaction or DNA rotation induced by Lhr (Figures 4B and 3F ii). Activating Lhr at 45°C (but without ATP-Mg²⁺) resulted in significant additional FRET populations corresponding to fork DNA undergoing changes into both stretched (E = 0.50) and further compacted (E = 0.92) conformations (Figure 4C,F iii). In these conditions, we also observed decreased signal intensities for compacted forks (E = 0.92 and E = 0.78) that corresponded with an increase in the low FRET efficiency population (E = 0.12), representing a highly stretched or partially unwound fork DNA conformation (Figure 4D,F iv) — the increased inter-dye distance indicated disruption of multiple base pairs close to the fork branch-point. The data indicate that fork DNA binding by Lhr in the absence of ATP causes multiple changes in fork conformation, including partial melting of DNA close to the fork branch-point. Addition of ATP-Mg²⁺ resulted in disappearance of the stretched fork signal (E = 0.5, Figure 4E) and appearance of a population with E ~ 0 that results from the fork being mostly unwound, fully separating the FRET dye pair (Figure 4E,F part v).

Discussion

Lhr protein is highly conserved throughout archaea and has sequence homology with DDX damage repair proteins found in humans and other eukaryotes [7]. Lhr proteins form two sub-groups, Lhr and Lhr-Core, the latter including the archaeal proteins of 800–900 amino acids arranged into RecA-like and accessory domains required for helicase activity. Bacterial *lhr* and bacterial/archaeal *lhr-core* are often located in a conserved genome context with at least one gene encoding a nuclease enzyme; *lhr* with *rnt* that encodes a 3′ to 5′ exonuclease implicated in DNA repair [14,15], and *lhr-core* with MPE, a manganese dependent exonuclease [10]. Our observation of a replication phenotype from expression of archaeal Lhr (Figure 1) is consistent with a role in replication-coupled DNA repair suggested from genetic analyses of Lhr from *E. coli* and *M. tuberculosis*

[17,18]. It is also consistent with our data from *in vitro* helicase assays (Figures 2 and 3) and smFRET (Figure 4) that purified Lhr protein targets and unwinds DNA forks.

The 3' to 5' directional DNA translocation of archaeal Lhr is the same as bacterial Lhr [10], and in addition we observe a strong preference for unwinding of DNA within three- or four-stranded forked and Holliday

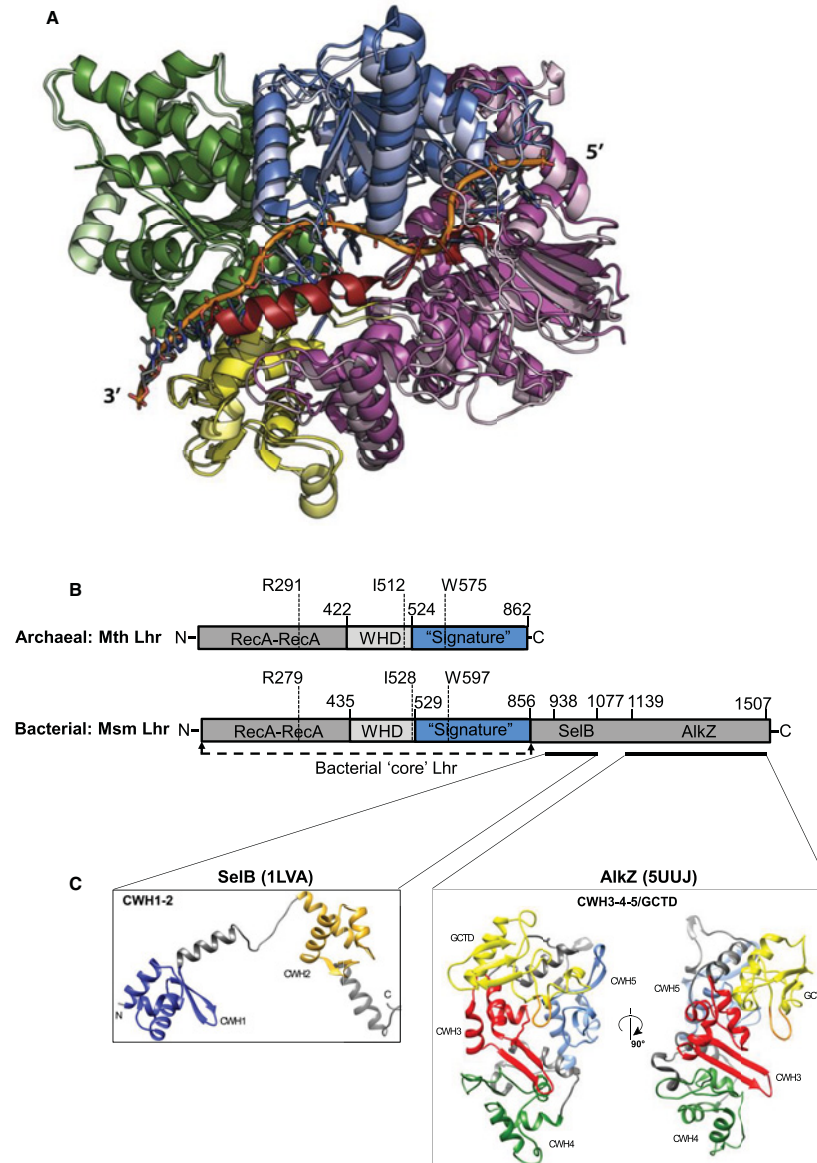


Figure 5. Novel features of Lhr protein structures.

(A) Structural model of the archaeal Lhr used in this work (Mth1802) superimposed onto *M. smegmatis* Lhr (PDB: 5V9X). Lhr is orientated and coloured according to the original description [9] (green, RecA domain 1; blue, RecA domain 2; yellow, winged helix; pink, domain of unknown function), with the Mth_1802 model in lighter shades. The ssDNA in the Lhr structure is shaded orange, and the *ab initio* modelled Mth1802 C-terminal 30 residues referred to in the text is shaded red. (B) Cartoon summary of the domain organization of Lhr proteins from archaea and bacteria. Labelled are the tandem RecA-like domains, winged helix domains (WHD) and a 'signature' domain of Lhr proteins that is of unknown functions. Amino acid positions are indicated, including invariant amino acids that are required for helicase activity of the bacterial Lhr [9]. Also highlighted is the 'core' helicase of the bacterial Lhr protein, and the C-terminal region of bacterial Lhr that is absent in archaea. (C) Summarizes two parts of the bacterial C-terminal Lhr region that match with structural folds of AlkZ and SelB proteins: CWH, C-terminal winged helix-turn-helix motif; GCTD, glycosylase C-terminal domain.

junction molecules, compared with ssDNA tailed-duplexes. Mycobacterial Lhr-Core was most active on RNA–DNA hybrids that have a 3′ ssDNA tail, although 3- or 4-strand forks or Holliday junctions were not tested [1]. Lhr does not seem to be a bona fide Holliday junction ‘branch migration’ helicase because it unwound model forked DNA better than model Holliday junctions, and because the products formed by Lhr unwinding Holliday junctions differed from the RuvAB branch migration complex. In addition, previous genetic studies on bacterial Lhr showed no strong phenotypes for Lhr associated with RuvABC or RecG-promoted recombination-repair, either epistatic or synergistic.

Our data showed more efficient unwinding of DNA forks by Lhr compared with unwinding of DNA from 3′ ssDNA tail provided to load Lhr for 3′ to 5′ translocation. This was despite the forked substrates being fully base-paired. Using single-molecule FRET we observed substantial melting and remodeling of the fork-2 substrate that would yield the ssDNA needed to trigger the ATP-dependent DNA translocation, thus unwinding the fork. The crystal structure of a mycobacterial Lhr-Core helicase bound to ssDNA most closely resembles the DNA repair helicase Hel308 [9,11,34], another Ski2-like helicase which has the same genetic phenotype as Lhr reported in this work and in previous studies [19,31]. The Lhr crystal structure represents the active translocation stage of Lhr, and the archaeal Lhr used in this work superimposes well when structurally modelled against it (RMSD 0.8 Å). including a region of the core bacterial and archaeal Lhr proteins, approximately amino acids 520–860, that is of unknown function that has been referred to as a ‘signature’ domain for Lhr proteins ([9] and Figure 5A). In addition, PHYRE2 *ab initio* modelling and PSIPRED searches of archaeal Lhr both predicted additional alpha helical content that was not resolved in the mycobacterial structure, including a 30-residue alpha helical extension intriguingly positioned relative to RecA-like domains and the translocating DNA strand (Figure 5A). We speculate that this may be significant for additional Lhr-DNA interactions, including with forked DNA, although it has not been possible to model a forked DNA structure onto these structures. Lhr is widespread across archaeal phyla (Supplementary Table S2) and can be easily identified in 30 bacterial phyla (Supplementary Excel File), although bacterial Lhr is distinguished from archaeal Lhr by the addition of a C-terminal 500–600 amino acids of unknown function that lacks obvious sequence homology to other proteins (Figure 5B). Structural homology searches and modelling using bacterial Lhr C-terminal residues against the PHYRE2 and DALI servers identified a region strongly matching protein folds in the DNA glycosylase enzyme AlkZ that contributes to replication-coupled DNA repair [35], and a smaller region matching tandem winged helix domains of the elongation factor SelB [36] (1.3 Å and 6.9 Å RMSD, respectively). We also noted interesting structural similarities between Lhr proteins and the human putative helicase DDX52, data that is presented in supplementary results (Supplementary Figure S7).

We conclude that our analyses indicate that archaeal Lhr proteins most likely target DNA arising at compromised replication forks, which may include RNA–DNA hybrids present as lagging strand Okazaki fragments. We propose that remodeling of fork DNA after binding by Lhr generates ssDNA for ATP-dependent DNA translocation to unwind the fork as part of DNA repair.

Competing Interests

Q3 The authors declare that there are no competing interests associated with the manuscript.

Q4 Funding

This work was supported by the Wellcome Trust (RCDF grant 023219 to ELB) and the Biotechnology and Biological Sciences Research Council (grant BB/M020541/1 to ELB).

Author Contributions

E.L.B. and D.G. designed the project; E.L.B., D.G., K.K. and C.D.O.C. wrote the manuscript; E.L.B., K.K., R.J.B. and C.D.O.C. carried out experiments.

Abbreviations

AA, acceptor excitation; APBS, all-photon burst search; DCBS, dual-channel burst search; HJ, Holliday junction; Lhr, large helicase-related; MPE, manganese-dependent phosphodiesterase; WHD, winged helix domain.

References

- 1 Ordonez, H. and Shuman, S. (2013) *Mycobacterium smegmatis* Lhr is a DNA-dependent ATPase and a 3′-to-5′ DNA translocase and helicase that prefers to unwind 3′-tailed RNA:DNA hybrids. *J. Biol. Chem.* **288**, 14125–14134 <https://doi.org/10.1074/jbc.M113.466854>

- 2 Reuven, N.B., Koonin, E.V., Rudd, K.E. and Deutscher, M.P. (1995) The gene for the longest known *Escherichia coli* protein is a member of helicase superfamily II. *J. Bacteriol.* **177**, 5393–5400 <https://doi.org/10.1128/JB.177.19.5393-5400.1995> 595
- 3 Ambur, O.H., Davidsen, T., Frye, S.A., Balasingham, S.V., Lagesen, K., Rognes, T. et al. (2009) Genome dynamics in major bacterial pathogens. *FEMS Microbiol. Rev.* **33**, 453–470 <https://doi.org/10.1111/j.1574-6976.2009.00173.x> 597
- 4 Chamieh, H., Ibrahim, H. and Kozah, J. (2016) Genome-wide identification of SF1 and SF2 helicases from archaea. *Gene* **576**, 214–228 <https://doi.org/10.1016/j.gene.2015.10.007> 598
- 5 Schutz, P., Karlberg, T., van den Berg, S., Collins, R., Lehtio, L., Högbohm, M. et al. (2010) Comparative structural analysis of human DEAD-box RNA helicases. *PLoS ONE* **5**, e12791 <https://doi.org/10.1371/journal.pone.0012791> 600
- 6 Xia, J., Chiu, L.Y., Nehring, R.B., Bravo Nunez, M.A., Mei, Q., Perez, M. et al. (2019) Bacteria-to-human protein networks reveal origins of endogenous DNA damage. *Cell* **176**, 127–143 e124 <https://doi.org/10.1016/j.cell.2018.12.008> 601
- 7 Abdelhaleem, M., Maltais, L. and Wain, H. (2003) The human DDX and DHX gene families of putative RNA helicases. *Genomics* **81**, 618–622 [https://doi.org/10.1016/S0888-7543\(03\)00049-1](https://doi.org/10.1016/S0888-7543(03)00049-1) 602
- 8 Ejaz, A., Goldgur, Y. and Shuman, S. (2019) Activity and structure of *Pseudomonas putida* MPE, a manganese-dependent single-strand DNA endonuclease encoded in a nucleic acid repair gene cluster. *J. Biol. Chem.* **294**, 7931–7941 <https://doi.org/10.1074/jbc.RA119.008049> 603
- 9 Ejaz, A., Ordonez, H., Jacewicz, A., Ferrao, R. and Shuman, S. (2018) Structure of mycobacterial 3'-to-5' RNA:DNA helicase Lhr bound to a ssDNA tracking strand highlights distinctive features of a novel family of bacterial helicases. *Nucleic Acids Res.* **46**, 442–455 <https://doi.org/10.1093/nar/gkx1163> 604
- 10 Ejaz, A. and Shuman, S. (2018) Characterization of Lhr-Core DNA helicase and manganese-dependent DNA nuclease components of a bacterial gene cluster encoding nucleic acid repair enzymes. *J. Biol. Chem.* **293**, 17491–17504 <https://doi.org/10.1074/jbc.RA118.005296> 605
- 11 Buttner, K., Nehring, S. and Hopfner, K.P. (2007) Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nat. Struct. Mol. Biol.* **14**, 647–652 <https://doi.org/10.1038/nsmb1246> 606
- 12 Northall, S.J., Buckley, R., Jones, N., Penedo, J.C., Soultanas, P. and Bolt, E.L. (2017) DNA binding and unwinding by Hel308 helicase requires dual functions of a winged helix domain. *DNA Repair (Amst)* **57**, 125–132 <https://doi.org/10.1016/j.dnarep.2017.07.005> 607
- 13 Johnson, S.J. and Jackson, R.N. (2013) Ski2-like RNA helicase structures: common themes and complex assemblies. *RNA Biol.* **10**, 33–43 <https://doi.org/10.4161/rna.22101> 608
- 14 Viswanathan, M., Lanjuin, A. and Lovett, S.T. (1999) Identification of RNase T as a high-copy suppressor of the UV sensitivity associated with single-strand DNA exonuclease deficiency in *Escherichia coli*. *Genetics* **151**, 929–934 609
- 15 Hsiao, Y.Y., Fang, W.H., Lee, C.C., Chen, Y.P. and Yuan, H.S. (2014) Structural insights into DNA repair by RNase T—an exonuclease processing 3' end of structured DNA in repair pathways. *PLoS Biol.* **12**, e1001803 <https://doi.org/10.1371/journal.pbio.1001803> 610
- 16 van Wolferen, M., Ma, X. and Albers, S.V. (2015) DNA processing proteins involved in the UV-induced stress response of sulfobacterales. *J. Bacteriol.* **197**, 2941–2951 <https://doi.org/10.1128/JB.00344-15> 611
- 17 Cooper, D.L., Boyle, D.C. and Lovett, S.T. (2015) Genetic analysis of *Escherichia coli* RadA: functional motifs and genetic interactions. *Mol. Microbiol.* **95**, 769–779 <https://doi.org/10.1111/mmi.12899> 612
- 18 Rand, L., Hinds, J., Springer, B., Sander, P., Buxton, R.S. and Davis, E.O. (2003) The majority of inducible DNA repair genes in *Mycobacterium tuberculosis* are induced independently of RecA. *Mol. Microbiol.* **50**, 1031–1042 <https://doi.org/10.1046/j.1365-2958.2003.03765.x> 613
- 19 Guy, C.P. and Bolt, E.L. (2005) Archaeal Hel308 helicase targets replication forks in vivo and in vitro and unwinds lagging strands. *Nucleic Acids Res.* **33**, 3678–3690 <https://doi.org/10.1093/nar/gki685> 614
- 20 Hishida, T., Han, Y.W., Shibata, T., Kubota, Y., Ishino, Y., Iwasaki, H. et al. (2004) Role of the *Escherichia coli* RecQ DNA helicase in SOS signaling and genome stabilization at stalled replication forks. *Genes Dev.* **18**, 1886–1897 <https://doi.org/10.1101/gad.1223804> 615
- 21 Howard, J.A., Delmas, S., Ivancic-Bace, I. and Bolt, E.L. (2011) Helicase dissociation and annealing of RNA-DNA hybrids by *Escherichia coli* Cas3 protein. *Biochem. J.* **439**, 85–95 <https://doi.org/10.1042/BJ20110901> 616
- 22 Komori, K., Hidaka, M., Horiuchi, T., Fujikane, R., Shinagawa, H. and Ishino, Y. (2004) Cooperation of the N-terminal helicase and C-terminal endonuclease activities of archaeal Hef protein in processing stalled replication fork. *J. Biol. Chem.* **279**, 53175–53185 <https://doi.org/10.1074/jbc.M409243200> 617
- 23 Schrimpf, W., Barth, A., Hendrix, J. and Lamb, D.C. (2018) PAM: a framework for integrated analysis of imaging, single-molecule, and ensemble fluorescence data. *Biophys. J.* **114**, 1518–1528 <https://doi.org/10.1016/j.bpj.2018.02.035> 618
- 24 Hellenkamp, B., Schmid, S., Doroshenko, O., Opanasyuk, O., Kuhnemuth, R., Rezaei Adariani, S. et al. (2018) Precision and accuracy of single-molecule FRET measurements—a multi-laboratory benchmark study. *Nat. Methods* **15**, 669–676 <https://doi.org/10.1038/s41592-018-0085-0> 619
- 25 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) 620
- 26 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H. et al. (2000) The protein data bank. *Nucleic Acids Res.* **28**, 235–242 <https://doi.org/10.1093/nar/28.1.235> 621
- 27 Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 <https://doi.org/10.1038/nprot.2015.053> 622
- 28 Holm, L. and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545–W549 <https://doi.org/10.1093/nar/gkq366> 623
- 29 Buchan, D.W.A. and Jones, D.T. (2019) The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 <https://doi.org/10.1093/nar/gkz297> 624
- 30 Allers, T., Ngo, H.P., Mevarech, M. and Lloyd, R.G. (2004) Development of additional selectable markers for the halophilic archaeon *Haloflex volcanii* based on the leuB and trpA genes. *Appl. Environ. Microbiol.* **70**, 943–953 <https://doi.org/10.1128/AEM.70.2.943-953.2004> 625
- 31 Fujikane, R., Shinagawa, H. and Ishino, Y. (2006) The archaeal Hjm helicase has recQ-like functions, and may be involved in repair of stalled replication fork. *Genes Cells* **11**, 99–110 <https://doi.org/10.1111/j.1365-2443.2006.00925.x> 626
- 32 Dickman, M.J., Ingleston, S.M., Sedelnikova, S.E., Rafferty, J.B., Lloyd, R.G., Grasby, J.A. et al. (2002) The RuvABC resolvosome. *Eur. J. Biochem.* **269**, 5492–5501 <https://doi.org/10.1046/j.1432-1033.2002.03250.x> 627
- 33 West, S.C. (1997) Processing of recombination intermediates by the RuvABC proteins. *Annu. Rev. Genet.* **31**, 213–244 <https://doi.org/10.1146/annurev.genet.31.1.213> 628

Q5

- 34 Richards, J.D., Johnson, K.A., Liu, H., McRobbie, A.M., McMahon, S., Oke, M. et al. (2008) Structure of the DNA repair helicase hel308 reveals DNA binding and autoinhibitory domains. *J. Biol. Chem.* **283**, 5118–5126 <https://doi.org/10.1074/jbc.M707548200>
- 35 Mullins, E.A., Warren, G.M., Bradley, N.P. and Eichman, B.F. (2017) Structure of a DNA glycosylase that unhooks interstrand cross-links. *Proc. Natl Acad. Sci. U.S.A.* **114**, 4400–4405 <https://doi.org/10.1073/pnas.1703066114>
- 36 Selmer, M. and Su, X.D. (2002) Crystal structure of an mRNA-binding fragment of *Moorella thermoacetica* elongation factor SelB. *EMBO J.* **21**, 4145–4153 <https://doi.org/10.1093/emboj/cdf408>

649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702

1 ***E. coli* DNA Repair Helicase Lhr is also a Uracil-DNA Glycosylase**

2 Ryan J. Buckley¹, Karl M. Hanson², Nadia R. Ahmed², Christopher D. O. Cooper²† and Edward L.
3 Bolt^{1*}

4

5 ¹University of Nottingham, School of Life Sciences, UK

6 ²School of Biological and Geographical Sciences, School of Applied Sciences, University of
7 Huddersfield, UK.

8 *Correspondence to ed.bolt@nottingham.ac.uk

9

10

11 **Summary**

12 DNA glycosylases protect genetic fidelity during DNA replication by removing potentially mutagenic
13 chemically damaged DNA bases. Bacterial Lhr proteins are well-characterized DNA repair helicases
14 that are fused to additional 600-700 amino acids of unknown function, but with structural homology
15 to SecB chaperones and AlkZ DNA glycosylases. Here we identify that *E. coli* Lhr is a uracil-DNA
16 glycosylase that depends on an active site aspartic acid residue. We show that the Lhr DNA helicase
17 activity is functionally independent of the uracil-DNA glycosylase activity, but that the helicase
18 domains are required for fully active uracil DNA glycosylase activity. Consistent with uracil DNA
19 glycosylase activity, deletion of *lhr* from the *E. coli* chromosome sensitized cells to oxidative stress
20 that triggers cytosine deamination to uracil. The ability of Lhr to translocate single-stranded DNA
21 and remove uracil bases suggests a surveillance role to seek and remove potentially mutagenic base
22 changes during replication stress.

23

24 **Keywords:** Helicase; glycosylase; DNA repair; uracil; DNA replication

25

26

27

28

29

30

31

32 **Introduction**

33 Lhr (*Large helicase-related*) proteins are ATP-dependent 3' to 5' DNA translocases within the
34 Superfamily 2 helicases [1]. The founder member of Lhr proteins was identified in bacteria [2], and
35 subsequently Lhr was found to be widely distributed across all clades of archaea [3]. High amino
36 acid sequence identity (typically about 30%) between archaeal and bacterial Lhr proteins is limited
37 to 800-900 amino acids that form helicase domains from the Lhr N-terminus—called the 'Lhr-Core'
38 [4]. Biochemical analyses of the Lhr-Core from the bacteria *Mycobacterium smegmatis* and
39 *Pseudomonas putida* and from the archaeon *Methanothermobacter thermautotrophicus* have
40 characterized Lhr translocation and helicase mechanism [5-7], and crystal structures of Lhr-Cores
41 highlight similarities with translocation by the archaeal DNA repair helicase Hel308 [6, 8], especially
42 in interactions between their winged helix and RecA-like domains [9, 10].

43 In addition to the Lhr-Core, bacterial Lhr proteins extend to 1400-1600 amino acids, in a C-terminal
44 protein region of unknown function, called Lhr-CTD (Lhr-C-terminal domains). Structural modelling
45 of the bacterial Lhr-CTD [11] and a subsequent cryo-EM structure [12], provided intriguing clues to
46 Lhr-CTD function, including the presence of an array of tandem winged helix domains characteristic
47 of the HTH_42 superfamily of proteins that have structural homology to the DNA glycosylase AlkZ
48 [11]. Genetic analyses of the effects on bacterial and archaeal cells of deleting the *lhr* gene revealed
49 mild sensitivities to agents that cause replication stress—UV irradiation [13] and azidothymidine
50 (AZT) [14]—and transcriptional up-regulation of *lhr* in response to mitomycin C [15]. In this work
51 we report new insights about how Lhr contributes to DNA repair in bacteria. We demonstrate that
52 the *E. coli* Lhr protein has uracil-DNA glycosylase activity, in addition to its well-characterized ATP-
53 dependent DNA translocase functions, and that cells lacking Lhr are sensitive to oxidative stress.

54

55 **Results**

56 ***E. coli* Lhr is an uracil DNA glycosylase requiring an active site aspartate**

57 We investigated whether *E. coli* Lhr is capable of DNA glycosylase activity, as suggested from
58 structural similarities between glycosylases and the uncharacterized C-terminal region of *E. coli* Lhr
59 (Lhr-CTD, LHR amino acids 876-1538) [11, 12] ([Figure 1A](#)). *E. coli* Lhr-CTD protein fragment and full-
60 length Lhr protein (1538 amino acid) were purified ([Figure 1B](#))—when Lhr-CTD (50 – 800 nM) mixed
61 with a Cy5-end labelled 37-nt ssDNA molecule modified to contain a single uracil nucleotide located
62 18 nucleotides from the 5' ssDNA end (uracil-ssDNA) a single product was observed on alkaline

63 treatment of reactions (Figure 1C, compare lanes 1-6 with 7-12), indicating DNA strand breakage at
64 a DNA abasic site consistent with glycosylase activity. No product was observed from Lhr-CTD mixed
65 with the same ssDNA lacking chemical modification (Figure 1C lanes 13-24).

66 To validate the uracil-DNA glycosylase activity of *E. coli* Lhr-CTD we sought to identify single amino
67 acid substitutions that would inactivate it. Sequence alignment of *E. coli* Lhr-CTD and AlkZ, with
68 which it has structural similarity [11, 12], were unproductive at identifying highly conserved residues
69 because Lhr-CTD lacks the 'QxQ' motif characteristic of AlkZ protein active sites [16], therefore
70 suggesting an alternate catalytic mechanism in LHR. Alternatively, potential active site amino acids
71 were identified through visual scrutiny of the Phyre2 [17] predicted model of the *E. coli* Lhr-CTD,
72 and in particular the positioning of side chains proximal to a proposed glycosylase active site (Figure
73 1D). Purified Lhr-CTD^{D1536A} (50 – 800 nM), gave no product when titrated into the uracil modified
74 ssDNA after alkaline treatment, compared with Lhr-CTD (Figure 1E). We then tested whether
75 substitution of the Lhr Asp-1536 residue inactivated uracil-DNA glycosylase active site chemistry or
76 had some other effect on the protein that perturbed DNA binding. Unmutated Lhr-CTD was unable
77 to form stable complex with DNA in EMSAs, when compared with full length Lhr (Figure 2A),
78 therefore we purified and tested full Lhr^{D1536A}. Lhr was also active as an uracil-DNA glycosylase
79 compared with Lhr-CTD (Figure 2B), but the Lhr^{D1536A} mutation inactivated glycosylase activity in
80 agreement with inactive Lhr-CTD^{D1536A} fragment (Figure 2C). In EMSAs Lhr^{D1536A} formed stable
81 complex with DNA similarly to Lhr (Figure 2D), therefore we conclude that Lhr is a uracil DNA
82 glycosylase that requires an active site aspartic acid residue.

83

84 **DNA glycosylase activity of *E. coli* Lhr is independent from its DNA helicase activity**

85 Full length Lhr was substantially more active than Lhr-CTD as a uracil-DNA glycosylase when
86 measured in assays as a function of time (Figure 3A) — this may be explained by much more stable
87 DNA binding by full length Lhr compared with Lhr-CTD that was observed in EMSAs (Figure 2A). We
88 therefore continued to use full-length Lhr to further investigate uracil-DNA glycosylase function
89 against flayed duplex DNA molecules that are substrates for unwinding by the Lhr 3' to 5' DNA
90 helicase activity [11]. For this work the duplex substrate was formed from annealing uracil-
91 containing ssDNA with its unmodified complementary DNA strand, with uracil positioned 8nt from
92 the fork branchpoint, 18 nt from the Cy5-DNA 5' end. Measured as a function of time, Lhr generated

93 glycosylase product from the uracil duplex at least 5-fold more effectively than when incubated with
94 uracil-ssDNA (Figure 3B), and Lhr was more active than Lhr-CTD on the uracil-fork DNA (Figure 3C).
95 Neither Lhr or Lhr-CTD gave any glycosylase product when uracil was substituted for a single 8-
96 oxoguanine residue at the same position in DNA (Figure 3D). The product of Lhr from uracil-DNA
97 single strands or duplex migrated close to the 16 nt marker, indicating that the same glycosylase
98 product was formed from both substrates. Glycosylase assay conditions in Figures 1-3 included Mg^{2+}
99 in the reaction buffer, but Lhr was also active in buffers containing EDTA, Mn^{2+} and Ca^{2+} instead of
100 Mg^{2+} (Figure 3E lanes 1-4) but was inactive as a glycosylase on DNA lacking a uracil residue (Figure
101 3E lanes 5-8). Lhr^{D1536} that is inactive as a uracil-DNA glycosylase was proficient at fork DNA
102 unwinding (Figure 3F), and we therefore conclude that the uracil glycosylase activity of Lhr is
103 functionally distinct from helicase activity, but we observe that glycosylase activity is enhanced
104 when the helicase domains are present, by contributing to DNA binding (Figure 2A).

105

106 ***E. coli* cells lacking LHR are sensitive to oxidative stress**

107 Oxidative damage to DNA in *E. coli* cells includes deamination of cytosine to uracil and further
108 oxidized uracil derivatives [18-21], triggering cytosine to adenine transversion mutations. We
109 therefore assessed for a contribution from Lhr to DNA repair in *E. coli* cells that is consistent with *in*
110 *vitro* uracil-DNA glycosylase activity. The *lhr* gene was deleted in *E. coli* MG1655 (Δlhr) by
111 recombineering, and we removed the inactivating antibiotic resistance marker, verified by
112 sequencing across the deletion site. We first tested Δlhr cells for sensitivity to azidothymidine (AZT),
113 a previously reported phenotype for Lhr in *E. coli* cells [14]. In a viability plate assay after growing
114 cells in broth (LB) containing a fixed 7.5 $\mu g/mL$ AZT we observed 10-fold reduced viability of Δlhr
115 cells compared with wild type cells (Figure 4A), and similar moderate sensitivity of Δlhr cells across
116 AZT concentrations (Figure 4B), agreeing with the previous study [14]. We next measured survival
117 of cells when grown in media containing hydrogen peroxide as a potent oxidizing agent. Hydrogen
118 peroxide (12.5 mM) added to growth media after cells had reached OD_{600} of 0.3 resulted in
119 significantly reduced growth of Δlhr cells in exponential phase compared with wild type cells (Figure
120 4C). This agreed with 10-100-fold reduced cell viability compared with wild type cells when Δlhr cells
121 grown in the same way, but without hydrogen peroxide, were then spotted onto LB agar containing
122 increasing concentrations of hydrogen peroxide to count their viability (Figure 4D).

123 **Discussion**

124 We provide biochemical evidence that *E. coli* Lhr is a uracil-DNA glycosylase (Lhr-UDG), a new
125 function for bacterial Lhr proteins alongside their well-characterized 3' to 5' single DNA
126 translocation activity that is stimulated by forked or flayed DNA substrates [5, 7, 11]. We show that
127 the Lhr-UDG activity requires an active site aspartate residue (Asp-1536), similarly to the active site
128 aspartate general base (Asp-62) that is essential for major groups of UNG/UDG proteins [22]. The
129 Lhr UDG function is positioned in the previously uncharacterized Lhr-CTD — though this fragment
130 of Lhr was proficient as a 'stand-alone' uracil-DNA glycosylase, its activity was significantly increased
131 by the presence of the Lhr helicase domains, probably by the helicase domains providing more
132 stable DNA binding compared with Lhr-CTD, observed in EMSAs. Inactivating the Lhr-UDG activity
133 did not inactivate DNA unwinding by Lhr, providing further support for the DNA binding functions
134 of Lhr being concentrated in the helicase domains.

135 Loss of Lhr from bacterial cells (Δlhr) causes mild sensitivity to AZT [14], a phenotype we also
136 observed after generating Δlhr cells and removing the inserted antibiotic resistance marker. We
137 identified that Δlhr cells were also significantly more sensitive than *lhr*⁺ cells to oxidative stress
138 induced by hydrogen peroxide, which is one of several routes causing genetic damage by cytosine
139 deamination in bacterial cells. This therefore supports our *in vitro* observations of Lhr-UDG function.
140 UDGs are ubiquitous in nature, although this is the first report of a UDG fused to a DNA helicase. *E.*
141 *coli* has a canonical UDG enzyme that functions in global DNA repair coupled with stable DNA
142 replication — upregulation of Mycobacterial Lhr in response to mitomycin C treatment [15], and
143 the sensitivity of *E. coli* cells to the polymerase inhibitor AZT when they lack *lhr*, may indicate that
144 Lhr is activated as part of bacterial responses to specific forms of replication-stress. In this context
145 removal of uracil from DNA by Lhr may protect genetic fidelity at sites that are overcoming blocked
146 DNA replication. We cannot exclude that Lhr may be able to remove other lesions or chemical
147 modifications from DNA, although we observed that it was inactive as a glycosylase against 8-
148 oxoguanine, suggesting that it has at least specificity for recognizing pyrimidine damage over
149 purines.

150

151

152

153 **Experimental Procedures**

154 **Proteins**

155 DNA sequences of primers and substrates, plasmids and *E. coli* strains are detailed in Supplementary
156 data. *E. coli* MG1655 gene b1653, encoding Lhr, was PCR-amplified from genomic DNA, and cloned
157 into pET14b using *Nde*I and *Hin*DIII restriction sites generating pRJB28 for expression of hexa-
158 histidine tagged Lhr. Lhr-CTD (amino acids 876-1538) was amplified and cloned by ligation
159 independent cloning (LIC) The resulting plasmid based on vector pNH-TrxT (GU269914.1 [23]). These
160 plasmids were used to generate Lhr^{D1536A} using mutagenic primers in PCR by Q5 hot start
161 polymerase, and resulting reactions were treated with *Dpn*I, T4 polynucleotide kinase, and DNA
162 ligase. Plasmid DNA was extracted and sequenced from colonies after transforming reaction
163 mixtures into *E. coli*.

164 Lhr and C-Lhr proteins were over-expressed in *E. coli* Rosetta 2 cells grown in MU broth with
165 ampicillin and chloramphenicol selection. Cells were grown with shaking at 37°C to OD₆₀₀ of 1.2 and
166 transferred to an ice slurry for cooling before addition of IPTG (0.8 mM). Growth was continued for
167 10 hours at 18°C, cells were harvested and resuspended in 20 mM HEPES pH 8.0, 1.5 M ammonium
168 sulfate, 20 mM imidazole, 10% (w/v) glycerol (Ni-NTA buffer A) containing 0.1 mM
169 phenylmethylsulphonyl fluoride (PMSF). This process and purification was also followed for
170 obtaining Lhr^{D1536A} protein. Cells were thawed and sonicated on ice, clarified by centrifugation, and
171 soluble proteins were loaded into a 5 ml butyl sepharose column equilibrated with 20 mM HEPES
172 pH 8.0, 1.5 M ammonium sulfate, 10% (w/v) glycerol (hydrophobic salt buffer A). The column was
173 washed with 20 mM HEPES pH 8.0, 900 mM ammonium sulfate. Then a 5 ml Ni-NTA column pre-
174 equilibrated with Ni-NTA buffer A was attached in tandem with the butyl sepharose column and
175 columns washed with 20 mM HEPES pH 8.0 and 10% glycerol until no proteins were detectable by
176 UV monitoring as eluting from the columns. The butyl sepharose column was removed and Lhr was
177 eluted from the Ni-NTA column by increasing imidazole to 500 mM in 20 mM HEPES and 10%
178 glycerol. Lhr-containing fractions were pooled and dialyzed overnight into 20 mM HEPES pH 8.0, 150
179 mM NaCl, 10% (w/v) glycerol (low salt buffer A) and loaded into a 1 ml Q-sepharose column. Lhr
180 eluted in an increasing gradient of NaCl to 1.5 M. Lhr fractions were pooled and dialyzed overnight
181 for storing in 20 mM HEPES pH 8.0, 150 mM NaCl and 35% (w/v) glycerol for aliquoting, flash
182 freezing, and storage at -80°C. *E. coli* uracil-DNA glycosylase and formamidopyrimidine DNA
183 glycosylase (Fpg) control proteins (Figure 3) were purchased from New England Biolabs.

184

185 ***In vitro* DNA binding, unwinding and glycosylase assays**

186 DNA strands for substrate formation (supplemental data) were synthesized with Cy5 end-label. DNA
187 binding was assessed using electrophoretic mobility shift assays (EMSAs). Reactions were incubated
188 at 37°C for 20 minutes in helicase buffer (HB); 20 mM Tris pH 7.5, 10% (v/v) glycerol, 100 µg/ml BSA,
189 using 12.5 nM Cy5-fluorescently labelled DNA substrate, 25 mM DTT and 5 mM EDTA, and then
190 placed on to ice for 10 minutes. Orange G and 80% (v/v) glycerol (OG) was added to load reactions
191 onto a 5% acrylamide TBE gel that was electrophoresed for 1 hour 30 minutes at 140 V. Gels were
192 imaged using a Typhoon phosphor-imager (Amersham) at 633 nm using a R765 filter for Cy5
193 detection.

194 DNA unwinding assays were at 37°C in reactions containing buffer HB, 12.5 nM Cy5-fluorescently
195 labelled DNA substrate, 25 mM DTT, 1.25 µM unlabeled 'trap' DNA, 5 mM ATP and 5 mM CaCl₂.
196 Reactions were pre-incubated at 37°C for 5 minutes without the 'trap' or ATP before they were
197 added together to start the reactions for 30 minutes at 37°C, stopped by addition of stock stop
198 solution (4 µl per 20 µl reaction); 2 mg/mL proteinase K in 200 mM EDTA and 2.5% (w/v) SDS. OG
199 dye was added for electrophoresis through a 10% acrylamide TBE gel for 45 minutes at 150 V. Gels
200 were imaged using a Typhoon phosphor-imager (Amersham) at 633 nm using a R765 filter for Cy5
201 detection.

202 DNA glycosylase reactions were at 37°C in reaction mixtures containing buffer HB, 12.5 nM Cy5-
203 fluorescently labelled DNA substrate, 25 mM DTT, 5 mM ATP, 4 mM MnCl₂ and 4 mM CaCl₂.
204 Reactions were pre-incubated at 37°C before being initiated by addition of Lhr protein and (unless
205 in a time course assay) allowed to continue for 30 minutes before addition of stock stop solution
206 and 4 µl of 1 M NaOH. Reaction samples were boiled for 5 minutes and formamide added before
207 loading into a 15% denaturing (8 M urea) acrylamide TBE gel for 4 hours at 5 watts per gel. Gels
208 were imaged using a Typhoon phosphor-imager (Amersham) at 633 nm using a R765 filter for Cy5
209 detection, generating TIFFs that were measured using Gel Analyzer 19.1 (Lazar) software. Graphs of
210 glycosylase activity were generated using Prism (GraphPad).

211

212 **Generation of a chromosomal deletion of *E. coli* *lhr***

213 DNA constructs and strain genotypes are presented in the Supplementary material. *Lhr* deletion was
214 by recombineering [24] and P1 transduction of an FRT (FLP recognition target) flanked Kan^r marker.

215 To generate an effective P1 stock the overnight culture was used to inoculate 8 ml of Mu broth
216 containing 6 mM CaCl₂. A sample of the cells (0.1 mL) grown at 37°C to OD₆₀₀ 0.8-1.0 in a shaking
217 water bath was added to four overlay tubes each containing 3 mL of 0.4% w/v Mu broth agar held
218 at 42°C. P1 phage stock was diluted 10-100-fold in MC buffer (100 mM MgSO₄, 5 mM CaCl₂) and
219 0.05 mL, 0.1 mL or 0.2 mL of this diluted phage was added to the overlay tubes containing cells and
220 molten agar and gently mixed. The remaining tube was left without phage as a control. The contents
221 of each overlay tube was poured onto P1 agar plates and left to set for overnight growth at 37°C for
222 18 hours. Soft agar from phage-lysed plates was added to 1 ml of MC buffer (100 mM MgSO₄ and 5
223 mM CaCl₂) and 0.5 ml of chloroform for vigorous mixing before centrifugation at 5752 rcf for 20
224 minutes at 4°C. The supernatant was retrieved and mixed with chloroform (0.5 mL) for storage at
225 4°C as a P1 phage stock. MG1655 recipient strain was grown in a Mu Broth to OD₆₀₀ 0.8 using a
226 shaking water bath. Cells were pelleted, resuspended in 1 ml of MC buffer, and left at 25°C for 10
227 minutes. 0.2 ml of cells were added into 3 overlay tubes containing 0 ml, 0.05 ml and 0.2 ml of P1
228 lysate produced previously and incubated for 30 minutes at 37°C. Cell/P1 lysate mix was added to
229 2.5 ml of 0.6% agar, mixed gently and poured onto Mu Broth agar plates containing 30 µg/ml
230 kanamycin and left to set. Plates were grown for 1-2 days, lid-up, at 37°C to allow colonies to
231 develop. Colonies were then picked and purified by streaking onto Mu broth agar plates containing
232 no antibiotic. This was repeated 3 times before plating again onto agar containing 30 µg/ml
233 kanamycin for confirmation of gene knockout and Kan^r-FTR insertion.

234 Successful P1 treated MG1655 cells were transformed with pCP20. Transformants were picked
235 and used to inoculate 8 ml of Mu broth containing no antibiotic. Culture was grown overnight at
236 45°C in a shaking water bath FLP recombinase expression and plasmid curation. Cells were then
237 streaked onto Mu Broth agar plates to produce single colonies and grown at 37°C overnight.
238 Colonies were re-streaked 3 times before replica plating onto Mu Broth agar plates containing 50
239 µg/ml ampicillin, 30 µg/ml kanamycin and then no antibiotic to confirm loss of the pCP20 plasmid.
240 Isolates which only grew on the no antibiotic agar plates were grown overnight for glycerol stock
241 production and streaked a further time for colony PCR diagnostic confirmation.

242

243 ***E. coli* viability spot assays**

244 Cell viabilities were measured from liquid cultures grown to OD₆₀₀ 0.3-0.4 in a shaking water bath
245 at 37°C monitored in the growth tubes by using a Spectronic 20+. Cells were then treated by addition

246 to the growth media of hydrogen peroxide (H₂O₂) or AZT at concentrations stated in the results.
247 Cells were grown for a further 30 minutes and then serially diluted into 1x M9 medium to arrest
248 growth for spotting (10 ul) on to agar plates grown overnight in a 37°C incubator. For comparing
249 growth curves cells were grown to OD₆₀₀ 0.3-0.4 and then transferred into a 24-well flat-bottomed
250 plate and H₂O₂ was added to appropriate wells to the given concentration from a 0.98 M stock.
251 Growth in the plates was monitored with orbital shaking in a FLUOstar microplate reader (BMG
252 Labtech). OD₆₀₀ readings were taken every 30 minutes in this time, and data was extracted and
253 analyzed using Prism (GraphPad) software.

254

255 **Author Contributions**

256 ELB and CDOC designed the project and with RJB wrote the paper. RJB, KMH, NA and CDOC
257 performed experiments, analyzed data, and generated images.

258

259 **Acknowledgments**

260 The work was supported by BBSRC grant BB/T006625-1 (ELB) and the BBSRC Doctoral Training
261 Partnership (ELB/RJB).

262

263 **References**

- 264 1. Hajj, M., et al., *Phylogenetic Diversity of Lhr Proteins and Biochemical Activities of the*
265 *Thermococcales aLhr2 DNA/RNA Helicase*. *Biomolecules*, 2021. **11**(7).
- 266 2. Reuven, N.B., et al., *The gene for the longest known Escherichia coli protein is a member of*
267 *helicase superfamily II*. *J Bacteriol*, 1995. **177**(19): p. 5393-400.
- 268 3. Chamieh, H., H. Ibrahim, and J. Kozah, *Genome-wide identification of SF1 and SF2 helicases*
269 *from archaea*. *Gene*, 2016. **576**(1 Pt 2): p. 214-28.
- 270 4. Ejaz, A., Y. Goldgur, and S. Shuman, *Activity and structure of Pseudomonas putida MPE, a*
271 *manganese-dependent single-strand DNA endonuclease encoded in a nucleic acid repair*
272 *gene cluster*. *J Biol Chem*, 2019. **294**(19): p. 7931-7941.

- 273 5. Ordonez, H. and S. Shuman, *Mycobacterium smegmatis Lhr Is a DNA-dependent ATPase*
274 *and a 3'-to-5' DNA translocase and helicase that prefers to unwind 3'-tailed RNA:DNA*
275 *hybrids*. J Biol Chem, 2013. **288**(20): p. 14125-34.
- 276 6. Ejaz, A., et al., *Structure of mycobacterial 3'-to-5' RNA:DNA helicase Lhr bound to a ssDNA*
277 *tracking strand highlights distinctive features of a novel family of bacterial helicases*.
278 Nucleic Acids Res, 2018. **46**(1): p. 442-455.
- 279 7. Ejaz, A. and S. Shuman, *Characterization of Lhr-Core DNA helicase and manganese-*
280 *dependent DNA nuclease components of a bacterial gene cluster encoding nucleic acid*
281 *repair enzymes*. J Biol Chem, 2018. **293**(45): p. 17491-17504.
- 282 8. Buttner, K., S. Nehring, and K.P. Hopfner, *Structural basis for DNA duplex separation by a*
283 *superfamily-2 helicase*. Nat Struct Mol Biol, 2007. **14**(7): p. 647-52.
- 284 9. Northall, S.J., et al., *DNA binding and unwinding by Hel308 helicase requires dual functions*
285 *of a winged helix domain*. DNA Repair (Amst), 2017. **57**: p. 125-132.
- 286 10. Johnson, S.J. and R.N. Jackson, *Ski2-like RNA helicase structures: common themes and*
287 *complex assemblies*. RNA Biol, 2013. **10**(1): p. 33-43.
- 288 11. Buckley, R.J., et al., *Mechanistic insights into Lhr helicase function in DNA repair*. Biochem J,
289 2020. **477**(16): p. 2935-2947.
- 290 12. Warren, G.M., et al., *Oligomeric quaternary structure of Escherichia coli and*
291 *Mycobacterium smegmatis Lhr helicases is nucleated by a novel C-terminal domain*
292 *composed of five winged-helix modules*. Nucleic Acids Res, 2021. **49**(7): p. 3876-3887.
- 293 13. van Wolferen, M., X. Ma, and S.V. Albers, *DNA Processing Proteins Involved in the UV-*
294 *Induced Stress Response of Sulfolobales*. J Bacteriol, 2015. **197**(18): p. 2941-51.
- 295 14. Cooper, D.L., D.C. Boyle, and S.T. Lovett, *Genetic analysis of Escherichia coli RadA:*
296 *functional motifs and genetic interactions*. Mol Microbiol, 2015. **95**(5): p. 769-79.
- 297 15. Rand, L., et al., *The majority of inducible DNA repair genes in Mycobacterium tuberculosis*
298 *are induced independently of RecA*. Mol Microbiol, 2003. **50**(3): p. 1031-42.
- 299 16. Mullins, E.A., et al., *Structure of a DNA glycosylase that unhooks interstrand cross-links*.
300 Proc Natl Acad Sci U S A, 2017. **114**(17): p. 4400-4405.
- 301 17. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis*. Nat
302 Protoc, 2015. **10**(6): p. 845-58.

- 303 18. Almatarneh, M.H., C.G. Flinn, and R.A. Poirier, *Mechanisms for the deamination reaction of*
304 *cytosine with H₂O/OH(-) and 2H₂O/OH(-): a computational study*. J Chem Inf Model, 2008.
305 **48**(4): p. 831-43.
- 306 19. Cadet, J. and J.R. Wagner, *DNA base damage by reactive oxygen species, oxidizing agents,*
307 *and UV radiation*. Cold Spring Harb Perspect Biol, 2013. **5**(2).
- 308 20. Kreutzer, D.A. and J.M. Essigmann, *Oxidized, deaminated cytosines are a source of C --> T*
309 *transitions in vivo*. Proc Natl Acad Sci U S A, 1998. **95**(7): p. 3578-82.
- 310 21. Krokan, H.E., F. Drablos, and G. Slupphaug, *Uracil in DNA--occurrence, consequences and*
311 *repair*. Oncogene, 2002. **21**(58): p. 8935-48.
- 312 22. Aravind, L. and E.V. Koonin, *The alpha/beta fold uracil DNA glycosylases: a common origin*
313 *with diverse fates*. Genome Biol, 2000. **1**(4): p. RESEARCH0007.
- 314 23. Savitsky, P., et al., *High-throughput production of human proteins for crystallization: the*
315 *SGC experience*. J Struct Biol, 2010. **172**(1): p. 3-13.
- 316 24. Datsenko, K.A. and B.L. Wanner, *One-step inactivation of chromosomal genes in Escherichia*
317 *coli K-12 using PCR products*. Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6640-6645.

318

319

320 **Figure Legends**

321 **Figure 1: The *E. coli* Lhr-CTD is a uracil DNA glycosylase requiring a catalytic aspartic acid**

322 **(A)** AlphaFold 2 structural model of *E. coli* Lhr that is based on strong homology with the cryo-EM
323 structure of Lhr helicase core and Lhr-CTD from *M. smegmatis*, respectively PDB: 5V9X and
324 PDB:7LHL. The *E. coli* Lhr-core helicase (amino acids 1-897) contains RecA domains, a beta-sheet
325 bundle (β) and a winged helix domain (WH) as indicated. Lhr-CTD (amino acids 898-1538) comprises
326 folds with structural homology to SecB chaperones and AlkZ glycosylases, as indicted.

327 **(B)** Coomassie stained SDS-PAGE acrylamide gels showing purified Lhr and Lhr-CTD, with molecular
328 mass ladder (M) values in kDa.

329 **(C)** Products from mixing Lhr-CTD (50, 100, 200, 400 and 800 nM) with 5' Cy5-ssDNA (12.5 nM)
330 containing a d-Uracil base 18 located nucleotides from the fluorescent moiety as indicated (lanes 1-

331 12), seen in a 15% denaturing acrylamide TBE gel. Addition of NaOH (lanes 8-12) causes β/δ
332 elimination at the site of the abasic DNA product, resulting in DNA backbone cleavage. This confirms
333 glycosylase protein activity. Marker (M) is made from known lengths of 5' Cy5 ssDNA.

334 (D) As for (C) in reactions containing unmodified 5' Cy5-ssDNA (12.5 nM).

335 (E) Phyre2 structural model of *E. coli* Lhr-CTD with predicted active site residues as labelled,
336 including Lhr-CTD residue Asp-1536 that we mutated in this work.

337 (F) Products from mixing Lhr-CTD and Lhr-CTD^{D1536A} proteins with 12.5 mM d-uracil containing 5'
338 Cy5-ssDNA substrate, viewed in a 18% acrylamide denaturing TBE gel. Product formation is shown
339 every 5 minutes for 30 minutes, observing no glycosylase activity from Lhr-CTD^{D1536A}.

340

341 **Figure 2: Lhr^{D1536A} is inactive as a glycosylase but binds to DNA**

342 (A) EMSA assays showing Lhr (12.5, 25, 50, 100 and 200nM) complexes bound to DNA (12.5 nM)
343 that are stable migrating through a 5% acrylamide TBE gel, compared with Lhr-CTD at the same
344 concentrations.

345 (B) Products of Lhr glycosylase activity seen in an 18% acrylamide denaturing TBE gel were absent
346 when reactions contained Lhr^{D1536A}. Proteins were used at 25, 50, 100 and 200 nM, with 12.5 nM
347 of d-uracil containing 5' Cy5-ssDNA substrate.

348 (C) EMSA showing that Lhr^{D1536A} and Lhr (12.5, 25, 50, 100 and 200 nM) form stable complex with
349 DNA in a 5% acrylamide TBE gel

350

351 **Figure 3: Lhr is inactive against 8-oxoguanine, and its uracil DNA glycosylase activity on duplex**
352 **DNA functions independently from Lhr helicase activity**

353 (A) Time-dependent uracil DNA glycosylase activity of Lhr (50 nM) compared with Lhr^{D1536A}. The
354 data shows means of glycosylase activity (n=3, with bars for standard error) alongside a
355 representative gel used for quantification.

356 (B) Comparison of Lhr (50 nM) glycosylase activity on ss-, ds- and forked d-uracil containing DNA
357 substrates (12.5 nM) as a function of time, with samples taken at time points indicated — plots
358 are means of two independent experiments showing standard error bars.

359 (C) Time-course assays (10, 20, 30 minutes) showing products from Lhr and Lhr-CTD (each 80 nM)
360 mixed with the preferred flayed duplex uracil-DNA, seen in an 18% acrylamide denaturing TBE gel.

361 Known length DNA strands are shown (M) and the positive control reaction (+ve) is product from 5
362 units of *E. coli* uracil DNA glycosylase.

363 (D) As for (C) except d-uracil DNA was replaced with otherwise identical 8-oxo-d-Guanine DNA,
364 and the control reaction (+ve) shows product formed by 5 units of formamidopyrimidine DNA
365 glycosylase (Fpg) protein.

366 (E) Lhr (80 nM) uracil-DNA glycosylase activity seen as products in 18% acrylamide denaturing TBE
367 gels (lanes 1-4), after 30-minute reactions in either EDTA, manganese or calcium, each replacing
368 magnesium as indicated, compared with unmodified DNA (lanes 5-8).

369 (F) DNA unwinding by Lhr and Lhr^{D1536A} proteins (12.5, 25, 50, 100 and 200 nM) on 12.5 nM of 5'
370 Cy5 labelled flayed duplex DNA, seen in 10% acrylamide TBE gel.

371

372 **Figure 4: *E. coli* cells lacking Lhr are sensitive to oxidative stress**

373 (A) Viability spot tests showing moderately increased sensitivity of Δ lhr cells to AZT (7.5 μ g/mL)
374 compared with wild type (wt) cells, and (B) represented in viability curves when Δ lhr and wild type
375 cells were grown independently in media containing AZT at 2.5, 5, 7.5 or 10 μ g/mL. The plots show
376 grow relative to wild type cells grown in media lacking AZT.

377 (B) Growth of Δ lhr and wild type cells monitored in 96 well plates in media containing 12.5 mM
378 H₂O₂, and (C) viability spot tests comparing Δ lhr and wild type cells grown in H₂O₂ added to media
379 at 1.5625, 3.75, 6.25 and 12.5 nM.

380

Figure 1

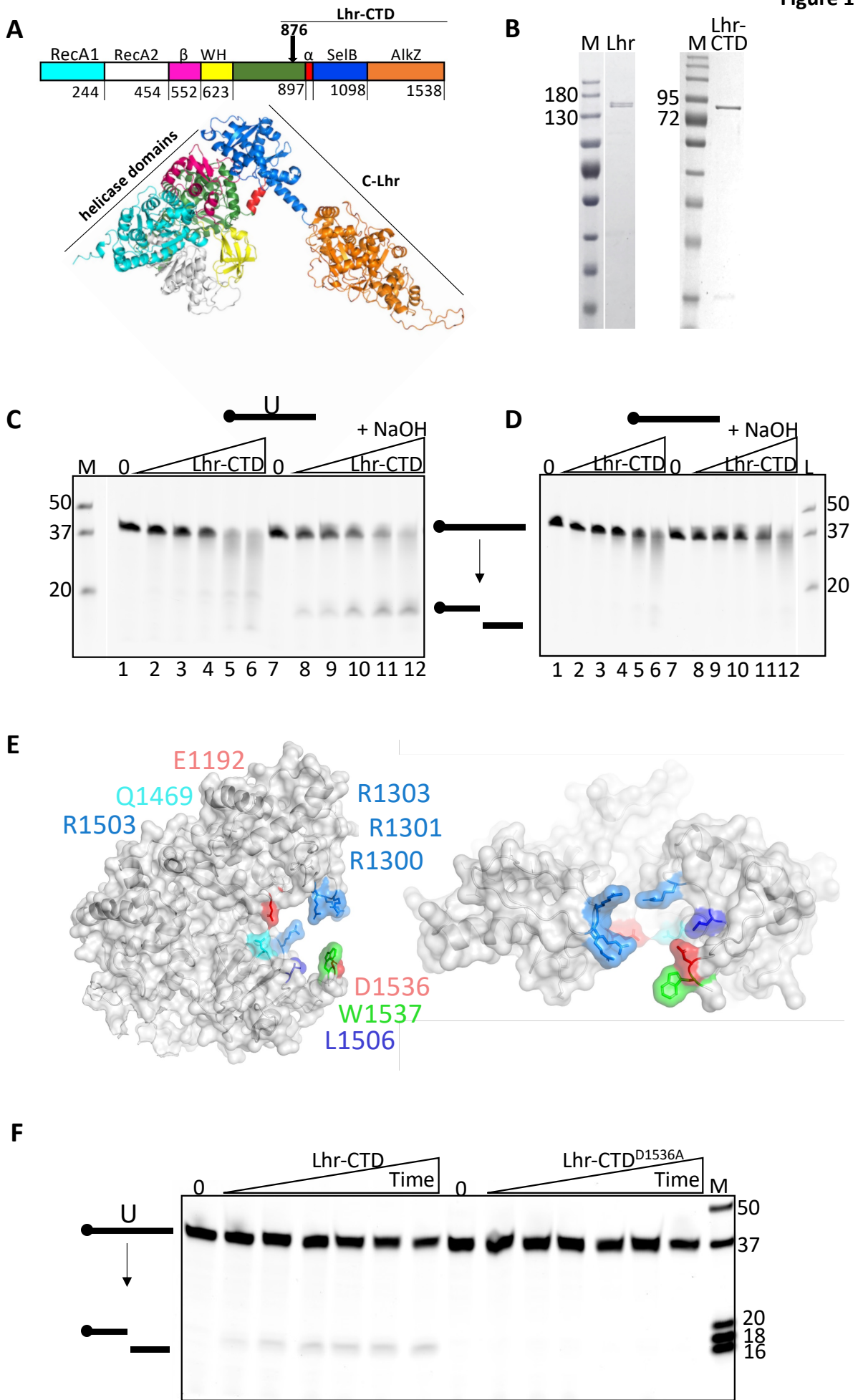
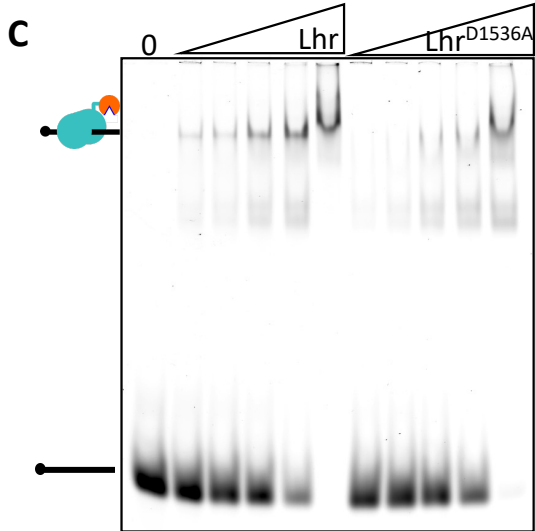
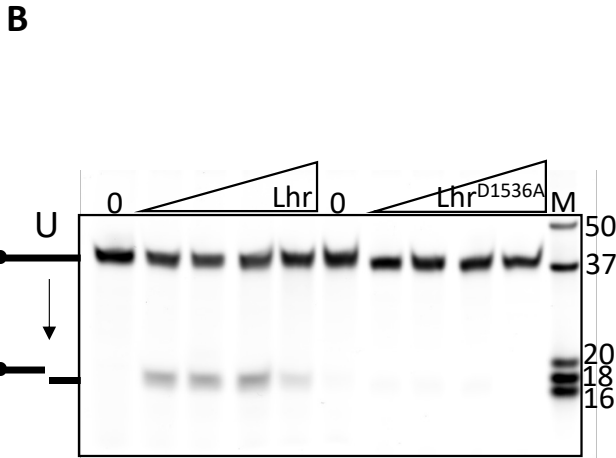
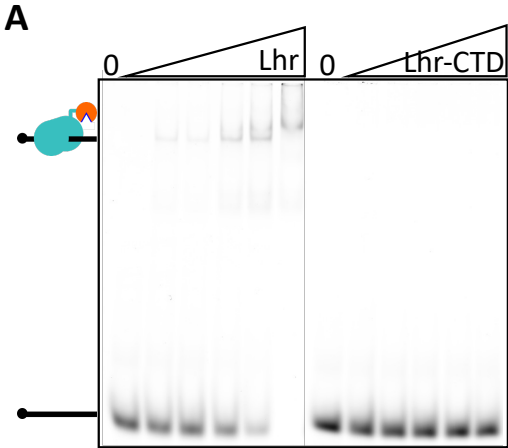
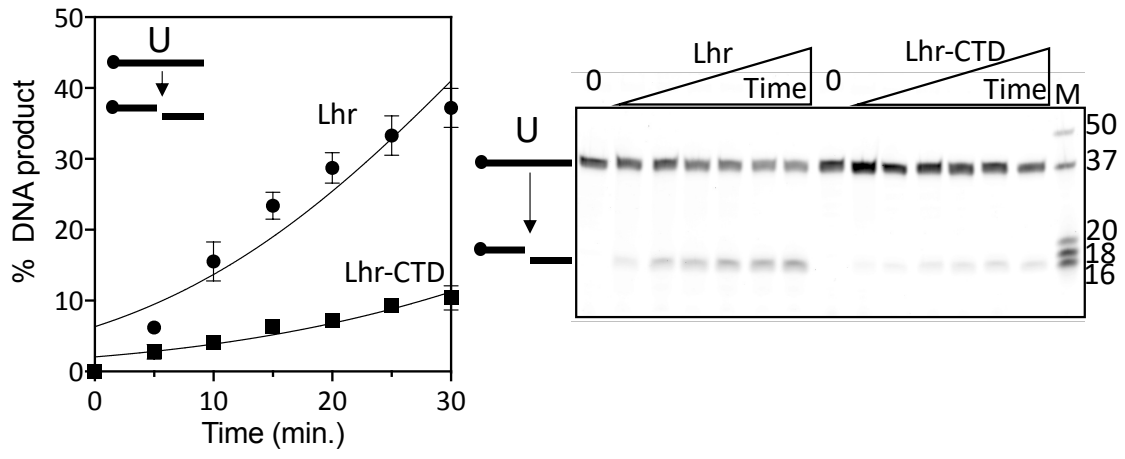


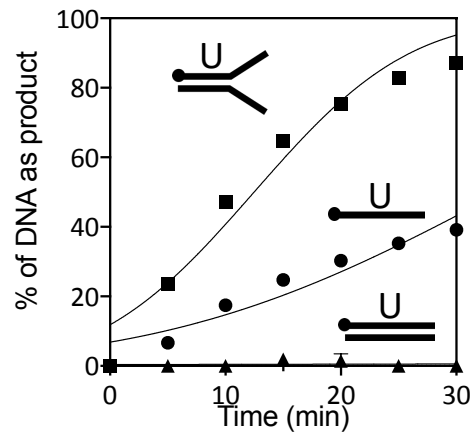
Figure 2



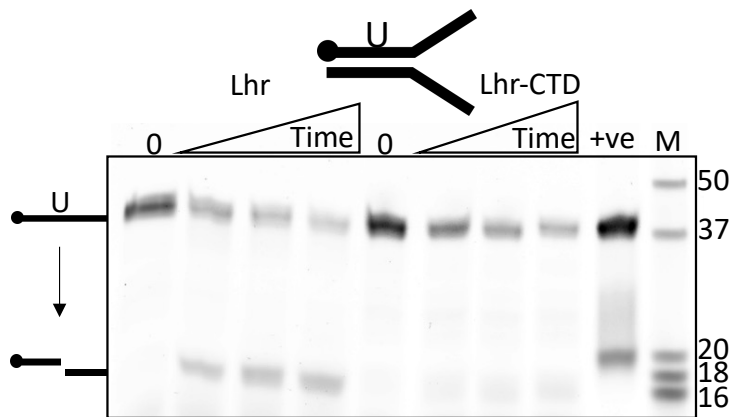
A



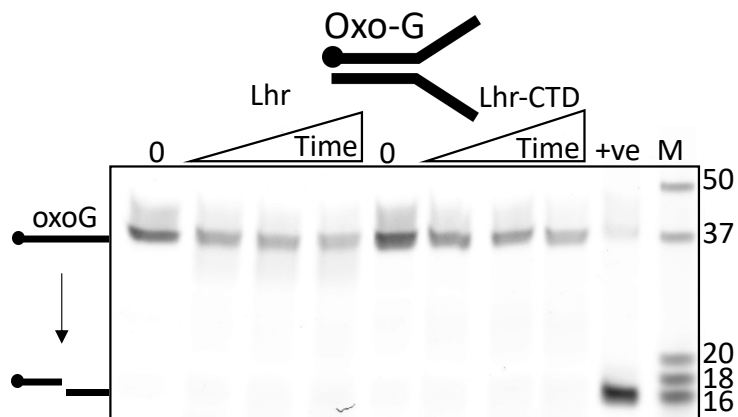
B



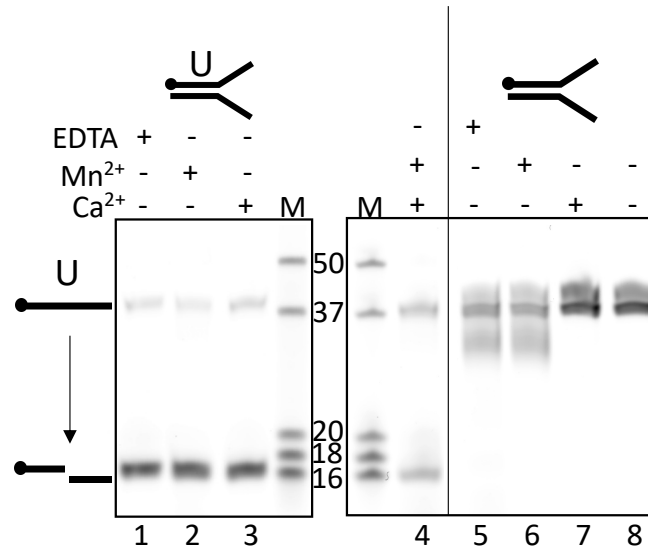
C



D



E



F

