Academic Unit of Molecular Oncology

Department of Oncology and Metabolism

University of Sheffield

Characterization of the human Pif1 helicase as a potential cancer-therapy target

Saba Dehghani-Tafti

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

September 2018

Acknowledgments

ABSTRACT

Pif1 is a multifunctional 5′ → 3′ superfamily 1 (SF1) helicase that is conserved in eukaryotes and also some prokaryotes. In addition to unwinding activity, the protein is a G-quadruplex (G4) DNA binding protein and strand annealing enzyme. Eukaryotic Pif1 proteins have roles in mitochondrial and nuclear genome stability and S-phase completion that are best characterised in *S. cerevisiae*. Pif1 knock-out mice are normal and recent studies show that the functions of human Pif1 (hPif1) only become critical for survival of some tumour cells during oncogene-induced replication stress, indicating that the enzyme is a potential cancer therapy target. However, hPif1 is poorly characterised at the structural and biochemical level and its function(s) critical for tumour cell survival are unknown. Here, a structure-function study and a discovery campaign to identify small molecule inhibitors of hPif1 was initiated.

A recombinant *E. coli* expression system was developed for hPif1 yielding milligrams to tens-of-milligrams of highly purified intact enzyme or truncation mutants encompassing helicase core (helicase domain, hPif1-HD). hPif1 crystal structures for the apo (1.4Å) as well as ATP hydrolysis ground (1.1Å) and transition state (3.9Å), representing structural events along the chemical reaction coordinate, were obtained. Comparisons with the structures of yeast and bacterial Pif1 reveal a conserved ssDNA binding channel in hPif1 that was demonstrated to be critical for single-stranded DNA binding during unwinding, but not the binding of G quadruplex DNA. Mutational analysis suggests that while the ssDNA-binding channel is important for helicase activity, it is not used in DNA annealing. Structural differences, in particular in the DNA strand separation wedge region, highlight significant evolutionary divergence of the human Pif1 protein from bacterial and yeast orthologues.

Libraries of drug-like molecules and chemical fragments were screened *in vitro* for inhibitors of hPif1 helicase activity and, using a protein thermal shift assay, direct binding. This screening identified chemical entities, some of which demonstrated preferential inhibition of hPif1 helicase activity compared to the SF1 helicase hUPF1, with the potential to progress in development to effective inhibitors. The established hPif1 crystallisation conditions were used to exploit the XChem fragment-based screening service at the Diamond Light Source national facility. A 233.3 Da chemical entity, EN300-02473, was identified in a 1.73Å crystal structure bound in a pocket between domains 2A and 2B of hPif1, close to the ssDNA binding site. EN300-02473 and two structurally related analogues inhibited hPif1 helicase activity *in vitro* with IC_{50} values from 80-110 μM and significant selectivity relative to hUPF1. Together, the findings reported here will serve as a platform to dissect and target hPif1 functions in intact cells.

ABBREVIATION

APS	Ammonium persulphate
AS	Ammonium sulphate
ATP	Adenosine triphosphate
ATR	Ataxia telangiectasia mutated- and Rad3-related
bp	base pair
BSA	Bovine Serum Albumin
dsDNA	Double-stranded deoxyribonucleic acid
DTT	1,4-Dithiothreitol
DMEM	Dulbecco's modified Eagle medium
DMS	Dimethyl sulphate
DMSO	Dimethyl Sulphoxide
EDTA	Ethylene diamine tetra acetic-acid
EMSA	Electrophoretic mobility shift assay
EtBr	Ethidium bromide
FA	Fanconi anaemia
FACS	Fluorescence activated cell sorting
FCS	Fetal calf serum
FRET	Förster Resonance Energy Transfer
GST	Glutathione-S-transferase
HD	Helicase domain
HEPES	4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid
His	Histidine
HRR	Homologous recombination repair
IC_{50}	Inhibition concentration at 50%
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl β-D-1 thiogalactopyranoside

kDa	Kilo Dalton
LB	Lysogeny broth
MES	2-(N-Morpholino)ethanesulfonic acid
MMC	Mitomycin C
MME	Monomethyl ether
MMR	Mismatch repair
mwco	molecular weight cut-off
NA	Nucleic acid
nM	Nanomolar
NER	Nucleotide excision repair
NHEJ	Non-homologues end joining
NMR	Nuclear magnetic resonance
OD_{600}	Optical density at 600 nm
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
Pfu	*Pyrococcus furiosus*
PI	Propidium Iodide
PMP	Poly(ethyleneimine)P, Polymin P
PMSF	Phenyl methyl sulfonyl fluoride
PTS	Protein thermal shift
PV	Papillomavirus
RNase	Ribonuclease
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SF	Super-family
ssDNA	Single-stranded DNA

TAE	Tris/Acetate/EDTA
TBE	Tris/Borate/EDTA
TCEP	Tris(2-Carboxyethyl)Phosphine
TE	Tris/EDTA
TEMED	N,N,N',N'-tetramethylethane-1,2-diamine
Tm	Melting temperature
Tris	Tris(hydroxymethyl)aminomethane
UPF1	Up-frame shift 1
UV	Ultra violet

CONTENTS

1 Introduction

1.1 Helicases

Helicases are a sub-group of nucleic acid (NA) translocase enzymes that unwind polynucleotide DNA or RNA secondary structures or remove bound proteins from NA substrates using NTP hydrolysis as an energy source (Bruening et al., 2014), Figure 1. NA unwinding is mostly considered in the context of duplex separation, but as will be discussed later other NA secondary structural forms are processed. Helicases are found in all kingdoms of life, from eukaryotes to prokaryotes and viruses and have been recognised as important factors in all DNA and RNA metabolic processes, including DNA/ RNA replication, transcription, translation and RNA splicing. They also function in a variety of specialised cellular pathways for genome stability maintenance such as DNA recombination and repair.
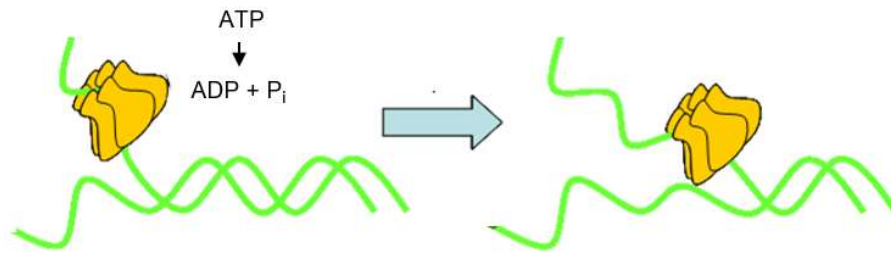
Figure 1. Schematic figure illustrating the typical action of a helicase (yellow) unwinding duplex NA (shown in green).

The importance of helicases to cell growth and survival is predominantly highlighted in some serious human diseases such as cancer and neurodegenerative disorders where they have been found to be defective (Uchiumi et al., 2015). These cellular abnormalities have been linked to impaired helicase function in a particular

cellular pathway, caused by either depletion or mutation/s in a gene encoding the helicase. Good example of this are the Werner's and Bloom's syndromes that are associated with a predisposition to cancer and/or features of premature ageing. They are caused by mutations in the *WRN* and *BLM* genes respectively, that each encode for a member of the RecQ family of DNA helicase (Karow et al., 2000). The vital contribution of helicases in maintaining genome stability has made them attractive targets for anti-cancer drug development, as discussed below. Helicases have also been considered as anti-viral targets for treating serious viral diseases (Brai et al., 2016).

The human genome encodes for an estimated 95 helicases or putative helicases, 31 DNA helicases and 64 RNA helicases (Umate et al., 2011). The need for this large number of helicases likely exists because of the variety of different DNA/ RNA secondary structures that can form in the many different NA metabolic pathways that may require resolution. This need also applies in all other organisms from complex eukaryotes to simple prokaryotes. For example, at least 14 different DNA helicases have been identified in simple single-cell organisms such as *Escherichia coli (E. coli)* (Tuteja and Tuteja, 2004).

1.1.1 The mechanistic categorisation of helicases

In order to explain and categorise helicases and their action, they have been considered in five main frameworks:

(1) Whether the helicase action is active or passive. This general classification of helicases indicates that there are two forms of unwinding activity. Helicases mostly unwind NA secondary structures in an active way, in which they are

3

directly involved in destabilizing secondary structure through NTP-induced conformational changes. In this type of unwinding action, the helicase is proposed to have at least two NA binding sites; dsNA and ssNA or two separate ssNA binding sites. The second possible type of unwinding is passive. There is no direct evidence for a purely passive unwinding activity of helicases. However, in this model it has been proposed that helicase activity is facilitated by its binding to ssNA that becomes available through momentary fraying of the dsNA junction caused by spontaneous thermal fluctuations. In other words, in this model ssNA that is dissociated momentarily at the NA duplex junction is trapped by the helicase hence re-pairing is inhibited without any consumption of ATP (Lohman and Bjornson, 1996, Manosas et al., 2010). Theoretical models suggest that in a passive helicase only a single ssNA binding site is required, however helicases with multiple DNA binding sites could still unwind dsNA using a passive mechanism in the absence of ATP.

(2) The helicase rate. The rate of helicase unwinding is described as the number of bases/ base-pairs (bp) translocated/ unwound per second. This can vary from a few bp/s to hundreds of bp/s among different helicases. Only for a few helicases have some quantitative data on the rates of duplex NA unwinding (*in vitro*) been reported. Other than the nature of the helicase itself, these rates are highly dependent on the solution conditions under which the experiment is taking place, such as pH, temperature, ATP concentration and especially salt concentrations. A good example of a detailed study of DNA unwinding rates is one performed with the RecBCD helicase-nuclease, an enzyme known to initiate the repair of double-stranded DNA breaks by homologous recombination in *E. coli*. Under the optimum salt concentration of the solution this helicase can unwind duplex DNA at rates of 180-470 bps per second (bp/s) at 25°C, which increases up to approximately 900 bp/s per second at 37°C (Roman and

Kowalczykowski, 1989). Duplex DNA unwinding rates, for example in the hexameric T7 helicase, can also be significantly influenced by guanine-cytosine base-pair composition (Donmez and Patel, 2008)

(3) Helicase processivity. Helicase processivity is the number of catalytic cycles that the enzyme undergoes before it dissociates from the product (ssDNA/RNA). In other words, helicase processivity is the numbers of bases it can translocate on or base pairs unwound in a single binding event with substrate (Lohman and Bjornson, 1996) . Usually helicases that are involved in DNA/RNA replication (hexameric helicases, see below) are highly processive, whereas those involved in DNA repair pathways do not have such processive ability and may unwind only a few tens of base pairs.

(4) Helicase directionality or polarity. Polarity is the direction which the helicase moves along on the NA substrate $3' \rightarrow 5'$ (type A) or $5' \rightarrow 3'$ (type B), which can be either ds- or ss-NA. The ssNA binding site of a helicase usually has a unique orientation with respect to the polarity of the sugar-phosphate backbone, dictating the translocation direction along the ssDNA/RNA (Lohman and Bjornson, 1996). The situation is less clear for dsNA translocases. There are also a group of bi-polar helicases including *E. coli* RecBCD, *Staphylococcus aureus* PcrA, and bacteriophage T5 helicase D2 that unwind NA secondary structures without any directional preferences (Dillingham et al., 2003, Anand and Khan, 2004, Wong et al., 2013).

(5) Step size. Step size is the number of base/ base-pairs a helicase translocates through or unwinds during each catalytic cycle. The active function of a helicase can be viewed as a number of repetitive steps that each contain an NTP binding and hydrolysis cycle linked to protein and NA conformational changes (Singleton et al., 2007, Lohman

et al., 2008). Individual steps during unwinding can be resolved using single molecule Fluorescence Resonance Energy Transfer (FRET) methods. Recent data with the hexameric T7 helicase indicate that steps size can be variable and that the helicase can store elastic energy like a spring before making a step (Syed et al., 2014).

1.1.2 The phylogenetic classification of helicases

All helicases characterized so far are classified into six main super-families; SF1-SF6 (Fig. 2). This is the most widely used classification of helicases and is based on structure and sequence similarities and also, to a certain level, functional similarities. Each SF may include helicases that have subsets of any of the five characteristics mentioned above and can contain helicases from both prokaryotes and eukaryotes. SF1 and SF2 helicases function as a monomers whereas the other four super-families (SF3-SF6) are known to form ring-like structures and function as a hexameric or double-hexameric units (Fig. 2). The classification of SF3, SF4, SF5 and SF6 families based on sequence analysis also defines a direct mechanistic relationship between the helicase members in these families. However, the helicases in SF1 and SF2 are more functionally and sequence diverse and include RNA or DNA helicases with either $5' \rightarrow 3'$ or $3' \rightarrow 5'$ (or bi-polar) polarity (Berger, 2008). Within the classification as a whole, helicase subtypes A and B have $3' \rightarrow 5'$ and $5' \rightarrow 3'$ polarity respectively. Also another subdivision of α and β is proposed for helicases that distinguishes enzymes that translocate on ssNA and dsNA respectively (Singleton et al., 2007) .
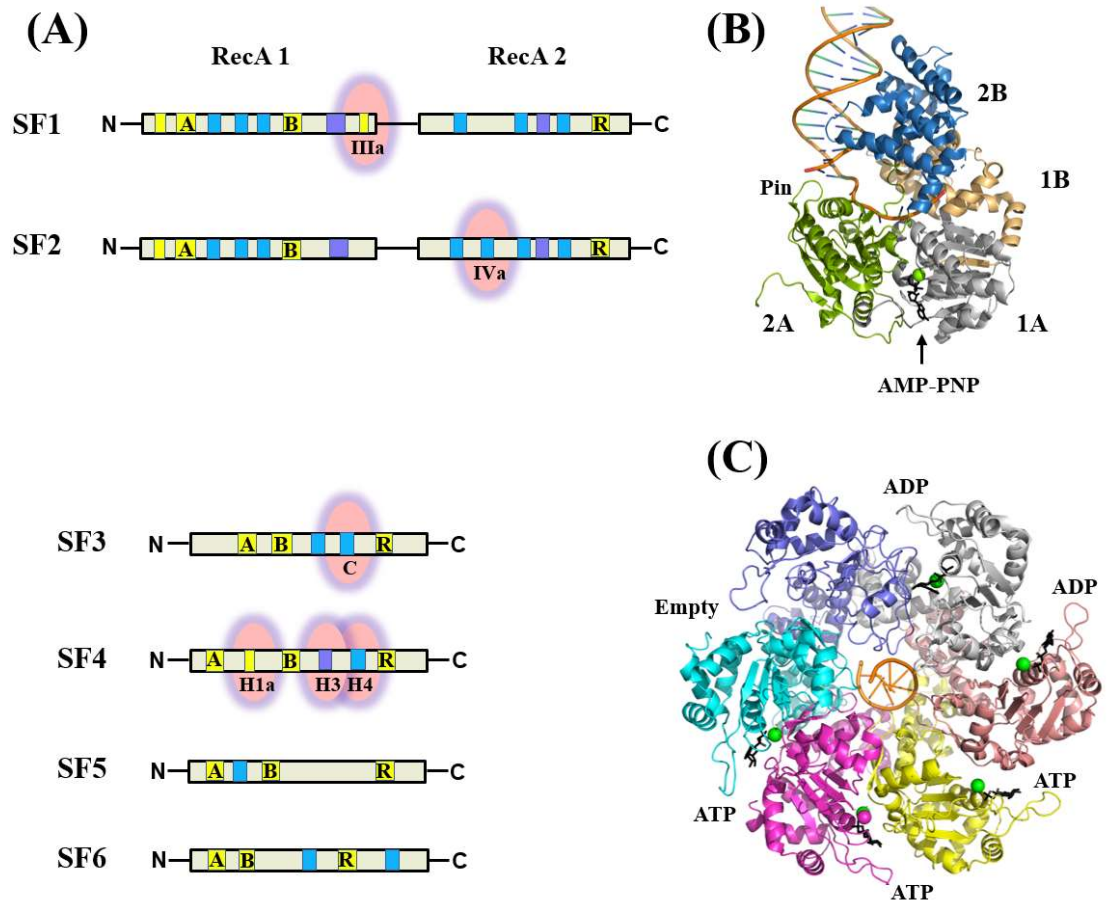
Figure 2. Sequence conservation and structure of the helicase super-families. **(A)** Sequence similarities in the helicase core of each super-family. Yellow boxes show positions of the conserved motifs involved in NTP binding and hydrolysis (A, Walker A region; B, Walker B and R, arginine finger motif). Segments shown in purple are associated with translocation, and blue are the sites that interact with the nucleic acid substrate. The domains highlighted in each superfamily are the conserved domain/s in that helicase family. **(B)** Structure of the monomeric SF1 helicase UvrD bound to a branched DNA structure and the nucleotide analogue AMP-PNP (pdb: 3FLU). The 1A and 2A domains are the RecA-like domains found in all SF1 and SF2 helicases. Nucleotide binding and hydrolysis events take place between the RecA domains and coordinate the chemomechanical events of translocation and unwinding. ssDNA runs between domains 1A and 2A, while other domains make additional interactions. **(C)** The hexameric helicase E1 (2GXA) illustrates the general organisation of this structural class. In the structure a short segment of ssDNA is bound in the central tunnel, while nucleotide binding and hydrolysis events between subunit interfaces power translocation. Adapted from (Jackson et al., 2014).

1.1.2.1 Helicase superfamilies 1 and 2

SF1 and SF2 are the largest and best understood classes of helicases, containing enzymes from both prokaryotes and eukaryotes. In spite of having considerable sequence variability these helicases generally all contain several related domains and a single NTP binding site at the interface of two highly conserved RecA-like folds (RecA1 and RecA2) as shown in Figure 2. Although the majority of helicases in these two super-families are known to function as monomers *e.g.* hepatitis C helicase NS3 , the bacteriophage T4 Dda helicase, *E. coli* RecQ helicase (Nanduri et al., 2002, Xu et al., 2003, Levin et al., 2004), there are a number which act additively or cooperatively as dimers or oligomers resulting in significant enhancement in their activity (Patel and Donmez, 2006). A good example of a dimerizing helicase is the UvrD helicase from *E. coli.* It has been shown that the monomeric form of UvrD is not fully competent at unwinding even a short 18 bp duplex DNA substrate. However, in the dimeric UvrD complex it is proposed that one subunit is bound to the ssDNA/dsDNA junction, while the second subunit is bound to the 3′ ssDNA tail, possibly as a back-up motor. It appears that a specific interaction between two UvrD monomers is required for efficient DNA unwinding *in vivo* and *in vitro* (Maluf et al., 2003).

The helicases of SF1 and SF2 can be divided into three mechanistic categories: (1) The ones that unwind duplex NA substrates with a preferred polarity (RecQ, Ski2-like, DEAH/RHA, NS3/NPH-II and all SF1 helicases), (2) Those that locally destabilize nucleic acid duplexes without any specific direction of unwinding (DEAD-box helicase) and (3) Those that can translocate on duplex substrates without strand displacement (RecG, Rig-I like, Swi/Snf2 and Type 1 restriction enzyme T1R) (Fairman-Williams et

al., 2010). Helicases like RecG are also known as branch-migration enzymes and have important roles in DNA recombination and repair.

1.1.2.2 Helicase superfamilies 3-6

The helicases in these super-families (SF3-SF6) are structurally very similar, typically having NTP binding sites shared between two adjacent monomeric units, but as a group they are not as extensively characterized as SF1 and SF2. These hexameric helicases are considered to be highly processive (as described above compared to the SF1 and SF2 helicases). In the hexameric unit, a strong topological binding arrangement encircling the NA substrate allows the entire unit to continue its action for a longer period of time. High processivity is required for unwinding long chromosomes and this is why most of the hexameric helicases are directly involved in genome replication. Examples include the eukaryotic DNA replication helicase MCM, bacteriophage T7 helicase, SV40 LTag and the BPV E1 helicase (Seo et al., 1993, Labib et al., 2000, Kim et al., 2002). Recently low- and high-resolution structures determined by electron microscopy (EM) have confirm the hexameric nature of the E1 and MCM protein complexes bound to replication fork-like substrates (Chaban et al., 2015, Georgescu et al., 2017). Alternatively, the SF1 and SF2 helicases that work as monomeric or sometimes additively as dimeric helicases, such as the RecQ and FANCJ helicases are mostly associated with recombination and repair processes (Hickson, 2003, Xu et al., 2003, Wu et al., 2009, London et al., 2008). It should be noted however that high processivity is not uniquely a property of hexameric helicases. The human SF1 helicase UPF1 has recently been shown to have remarkable processivity, 1000s of bp/ catalytic cycle (Fiorini et al., 2015).

The SF3 helicases are also known as AAA+ helicases (ATP*ase* associated with various cellular activities) because of their homology to the wider group of AAA+ proteins. The group predominantly holds helicases from small DNA and RNA viruses. They form hexamers or double hexamers and have a $3' \rightarrow 5'$ polarity of action. Some of the best insights into the structure and function of this super family comes from the structure of the papilloma virus E1 helicase in complex with ssDNA and ADP and with no cofactors bound (Enemark and Joshua-Tor, 2006, Sanders et al., 2007). These data suggest a mechanism for DNA translocation whereby ATP binding and hydrolysis events are linked to positional changes in ssDNA binding hairpins that track ssDNA in a spiral arrangement in a central tunnel of the assembly.

The SF4 helicase family is also known as the DnaB-like helicases and is exemplified by the *E. coli* DnaB replication fork helicase. All characterized SF4 helicases have a $5' \rightarrow 3'$ polarity of activity and like SF1 and SF2 are also composed of a RecA-like fold. Members of helicase SF4 were first identified in bacteria and bacteriophages and are all replication helicases. In bacteria, the helicase's activity is associated with a primase (the DnaG protein) composed from an independent polypeptide. However, in some bacteriophage systems, the homologous protein contains both catalytic activities (helicase and primase) within a single polypeptide chain. An example of this is a gp4E protein in bacteriophage T7 which is one of the most extensively studied SF4 helicases. This protein is composed of primase and helicase domains that are joined together by a linker polypeptide (Toth et al., 2003).

Helicases SF5, also known as the Rho-like helicases, are closely related to SF4 and have a $5' \rightarrow 3'$ polarity. In spite of being very similar to the SF4 helicases, on the basis of their sequence differences they were categorized in a separate super-family.

Rho is involved termination of transcription in bacteria. It binds to a specific sequence on the newly synthesised RNA and by unwinding the DNA/RNA duplex it dissociates the transcription machinery (Skordalakes and Berger, 2003, Lyubimov et al., 2011). The crystal structure of Rho bound to a short segment of ssRNA is solved. Interestingly, like BPV E1, six bases of RNA are bound in the interior of the assembly, but the protein architecture governs the opposite translocation polarity (Thomsen and Berger, 2009).

SF6 helicases like SF3 are part of the AAA+ group of proteins and are $3' \rightarrow 5'$ helicases. This super-family contains a number of eukaryotic and archaea replication factors including the MCM helicases (MCM2-7 complex), human RuvBL1 and prokaryotic RuvB helicases (Singleton et al., 2007, Berger, 2008). Like the bacterial and viral SF3 replicative helicases, MCM2-7 is required for both initiation and elongation in genome replication, during the late G1 and S phase of the cell cycle (Tanaka and Diffley, 2002). The role of yeast MCM2-7 has recently been substantiated by structural observation. The cryo-EM structure of MCM double hexamer bound to double-stranded origin DNA suggests a mechanism for establishment of a replication fork complex (Noguchi et al., 2017). In eukaryotes, the MCM helicase is composed of six heterologous subunits each encoded by a separate gene MCM2, 3, 4, 5, 6 and 7. The name MCM1 (Mini Chromosome Maintenance) had already been given to a transcription factor which is why the labelling of the subunits starts from number 2 (Treisman and Ammerer, 1992). Studies have shown that any type of defect in any of the subunits or their depletion can cause a cell-cycle block at G1 phase in yeast (Labib et al., 2000). In G1 phase of the cell cycle all six subunits of MCM2-7 are loaded on origins of replication in an inactive form by a factor called Cdt1, to form the pre-replication complex (pre-RC) that contains Cdc6 and the hetero-hexameric origin

recognition complex Orc1-6. Ultimately, the purpose of the formation of a pre-RC is to assemble the MCM2-7 onto the DNA replication fork. This is an active process and requires ATP hydrolysis by both Orc1-6 and Cdc6; once loaded the rest of the pre-RC factors are no longer needed and are dispensable for subsequent DNA synthesis (Donovan et al., 1997, Wyrick et al., 2001, Randell et al., 2006).

1.1.3   Conservation of the helicase ATP*ase* catalytic core

As a growing number of crystal structures of helicases from different super-families became available it became clear that there are three catalytic motifs that are absolutely conserved in all helicase families (shown in Figure 2a) that are required for NTP binding and hydrolysis: (1) The Walker A motif or P-loop which is in the NTP binding site. (2) The Walker B motif which contains at least one acidic residue and comprises an Mg^{2+} binding site critical for NTP hydrolysis site (Hall and Matson, 1999, Jackson et al., 2014) and (3) The arginine finger/s that is known to make important contacts with ATP and is required for the ATP hydrolysis step. Although the exact function of this motif is not fully understood it has been shown that any mutation in this specific site significantly reduces the helicase activity. Many studies have highlighted the importance of this motif in different helicases. For example, a mutagenesis studies performed on the *E. coli* DbpA RNA helicase showed that two arginine residues in the ATP binding site, R331 and R334, make vital contacts with the ATP. When mutated to alanine the ATP*ase* and helicase activity of the enzyme is completely abolished even, though the helicase is able to bind RNA at wild-type levels (Elles and Uhlenbeck, 2008).

1.2 Helicases in DNA repair, genome stability and cancer

Helicases have major roles in maintaining genome stability. They directly resolve B and non-B DNA/RNA forms and participate in DNA replication and repair pathways by co-functioning with other proteins in the replication/ repair machinery.

1.2.1 Non-B DNA/RNA structures and their resolution by helicases

As mentioned previously, helicases are enzymes that unwind DNA or RNA secondary structures. Other than unwinding the simple B-form of DNA which is the well-known right-handed Watson and Crick linear double helix (Watson and Crick, 1953), helicases are able to unwind non-B form NA structures. Under certain conditions, DNA (and RNA) can adopt non-B conformations such as looped-out slipped (Hairpin) structures, triplex DNA (H-DNA), G-quadruplex (G4) DNA, cruciform/ Holliday junctions and left-handed Z-DNA, shown in Figure 3 (Bacolla et al., 2006b). Other specific structures can form in polypurine or polypyrimidine homopolymeric tracts including the I-motif, a C-quadruplex structure (Fleming et al., 2017) as well as single-stranded helical structures.

Mounting evidence now suggests that at least a subset of these non-B DNA structures can form *in vivo* (see below), at least transiently, usually in a specific polynucleotide sequences during DNA metabolic processes such as replication, transcription, repair, or recombination (mostly outside of the coding regions). There are several studies supporting the existence of some non-B structures as part of gene regulatory elements, *i.e.* influencing the transcription of a particular gene or groups of genes (Huppert and Balasubramanian, 2007, Oh et al., 2002, Wong et al., 2007, Khuu et al., 2007), and therefore they may have genuine physiological functions. However, if

these structures remain un-resolved they can stall or collapse replication forks resulting in single- and double-stranded breaks in the DNA leading to cell death (if detected by the cell cycle checkpoints). Alternatively, they can alter DNA repair mechanisms leading to chromosome re-arrangements, mutations/ deletions in the genome. Some very serious genetic disorders (see below) have been linked to the occurrence of un-resolved non-B DNA structures, in particular G-quadruplexes (Inagaki et al., 2009, Wang and Vasquez, 2004, Wang et al., 2006, Bacolla and Wells, 2009). Loss of genomic integrity is a common hallmark of cancer (Hanahan and Weinberg, 2000). Recent large-scale cancer genome profiling studies have led to the conclusion that the ability of helicases to resolve these non-B DNA structures and also participate in their re-annealing to dsDNA (in some cases) is vital for sustaining a cell's genomic integrity (De and Michor, 2011). The details of the principal categories of genotoxic non-B form DNA structures are described below:

(1) The DNA Hairpin. Nearly 30% of the human genome is comprised of repetitive DNA sequences spread throughout the chromosomes (Moyzis et al., 1989, Stallings et al., 1990, Padeken et al., 2015). When these repeats are <10 nucleotides long they are referred to as microsatellites, and when they contain 10–100 bases are called mini-satellites. A number of genetic human diseases have been associated with either a decrease or expansion of these satellites (Caskey et al., 1992, Krontiris, 1995, Mirkin, 2007, Lopez Castel et al., 2010). These structures, also known as slipped-hairpin structures, are usually formed when direct repeats base-pair in a misaligned manner with the complementary strand and from a loop-like structure extruding from the double helix (Fig. 3). The bases in the looped-out strand may form duplexes by inter-strand stacking interactions that stabilise the hairpin (Sinden et al., 2007, Edwards et al., 2009).
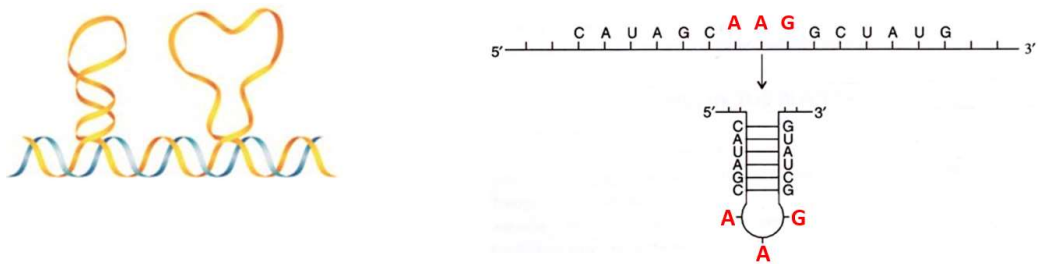
Figure 3. Schematic figure of slipped-hairpin structures (**left**). Example of a sequence of hairpin forming site in an ssRNA (**right**). Adapted from (Zhao et al., 2010).

(2) Cruciform DNA structures: These structures usually form in inverted repeat sequences where one side of an inverted repeat is complementary to the sequence on the other side of the same strand, *e.g.*, GACTGC….GCAGTC (Fig. 4). The formation of these structures requires energy which some studies claim may come from negative supercoiling (Gellert et al., 1983, Oussatcheva et al., 2004, Zhao et al., 2010, Nag and Petes, 1991). The two inverted repeats pair with each other on the same DNA strand and form a hairpin-loop, leaving an un-paired sequence at the symmetric centre of the loop. Overall, the cruciform structure is composed of two hairpin-loops that are positioned opposite each other at a four-way junction, which is structurally similar to Holliday junctions that form in DNA recombination processes (Sinden and Pettijohn, 1984, Watson et al., 2004, Panayotatos and Wells, 1981).
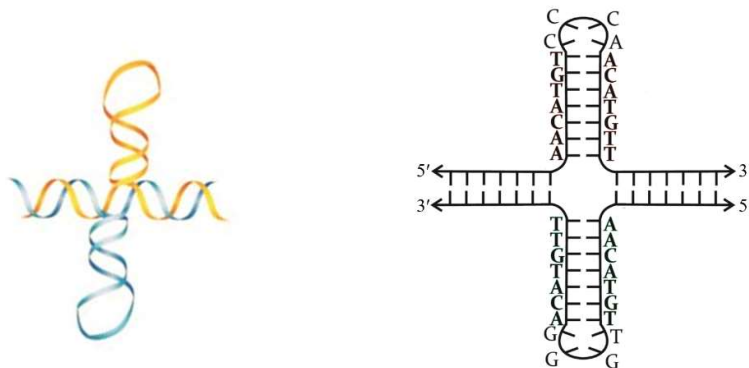
Figure 4. Schematic figure of a cruciform structure (**left**). Example sequence of cruciform forming site (**right**). Adapted from (Zhao et al., 2010) and (Brázda et al., 2011).

(3) Z-DNA: (CG/GC)$_n$ and (CA/GT)$_n$ sequences can cause the double helix of DNA to wind into an unusual form of a left-handed zigzag structure (Z-DNA) that causes morphological changes in the major and minor grooves relative to normal B-form DNA (Fig. 5). In these sequences the purines (A, G) adopt an inverted orientation compared to the B-form, causing organizational changes in the sugar–phosphate backbone that lead to changes in the entire conformation of the double helix and base extrusion at the B-Z helical junctions (Singleton et al., 1982, Ha et al., 2005, Kim et al., 2009). These structures form in the genome of both eukaryotes and prokaryotes in certain purine-rich sequences and have been visualized and analysed in a few detailed studies. With a frequency of 1 in 3000 bp in the human genome, they are known to be the cause of large scale deletions (Hamada and Kakunaga, 1982, Wang et al., 2006, Schroth et al., 1992).
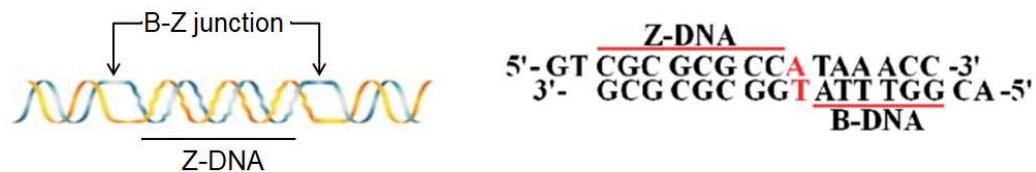
Figure 5. Schematic figure of Z-DNA structures, showing the B-DNA and Z-DNA junction (**left**), modified from (Zhao et al., 2010). Example of a Z-DNA forming sequence. The extruded base is shown in red at the Z and B junctions (**right**), Reproduced from (Kim et al., 2009).

(4) H-DNA: Triplex-forming sequences exist at relatively high levels in mammalian genomes, at about 1 in ~50,000 bps (Schroth and Ho, 1995, Bacolla et al., 2006a). Triplex DNA structures usually form in homopurine/ homopyrimidine sequences that have mirror symmetry and a DNA single-strand can bind in the major groove of the underlying duplex DNA helix to form a three-stranded helix (Fig. 6). Triplex DNA is classified into two types depending on the positioning and sequence composition of the invading single strand. One is when a pyrimidine-rich single-strand binds to the complementary purine-rich strand of the duplex DNA, in which case the orientation of the single-strand is parallel to the duplex helix (Y*R:Y). The second type is when a purine-rich single-strand binds to the complementary pyrimidine-rich strand of the duplex DNA in which case the orientation of the single strand is anti-parallel to the duplex helix (R*R:Y). The second type of triplex is more common and usually forms under physiological conditions whereas the (Y*R:Y) type forms at acidic pH (Htun and Dahlberg, 1988, Jain et al., 2008, Wang and Vasquez, 2004).
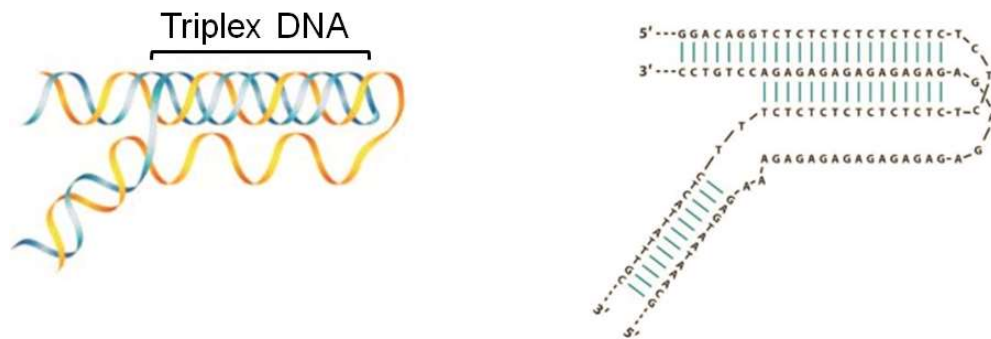
Figure 6. Schematic figure of H-DNA (**left**). Example of the triplex forming sequence (Y*R:Y type), (**right**). Modified from (Zhao et al., 2010).

(5) G4-DNA: These structures are the most important non-B secondary structures that have been comprehensively investigated in *in vivo* and *in vitro* studies (Lipps and Rhodes, 2009, Biffi et al., 2013). They form due to the unique ability of guanosine to hydrogen bond with itself (Hoogsteen bonding) forming tetrads within a ssDNA sequence (uni-molecular G4 DNA) or between other ssDNA strands (polymolecular G4 DNA). Such structures are stabilized by a metal-ion at the core of the structure *e.g.* K^+, Na^+ and Ca^{2+} (Fig. 7).

G4 DNA forms in the G-rich sequences of DNA (*e.g.* the human telomere sequence, d[AGGG(TTAGGG)$_3$]). Three types of this structure have been observed and investigated by NMR and X-ray crystallography studies; parallel, anti-parallel and a hybrid form which is a mixture of the two (Fig. 7). It has been proposed that the type of metal-ions in the environment affects the folding of the G-quadruplex, for example under Na^+ dominant conditions the G-rich sequences are more likely to form the parallel form, whereas in K^+ dominant conditions (*i.e.* intracellular) they are likely to form the anti-parallel form (Phan and Patel, 2003, Singh et al., 2016, Laughlan et al., 1994, Parkinson et al., 2002, Zhao et al., 2014).
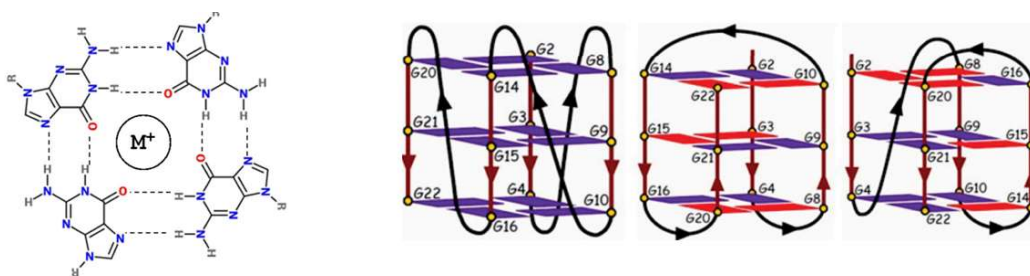
Figure 7. Schematic figure of G-quadruplex DNA structures. Left, a G-tetrad structure showing the hydrogen bonds (Hoogsteen) between four guanosines and the stabilizing metal-ion in the core. **Right**, three types of G-quadruplex folding; parallel, anti-parallel, and the hybrid form respectively from left to right. Reproduced from (Frees et al., 2014, Singh et al., 2016).

About 37000 tracts in the human genome, mostly (>60%) outside of the coding regions, have been predicted to form G4-DNA structures, which approximates to one G-quadruplex in every 10 kb (Paeschke et al., 2011, Huppert and Balasubramanian, 2005, Todd et al., 2005). G4-DNA structures are involved in several fundamental biological processes such as aging and disease, and most notably cancer. Some studies suggest that G4-DNA structures are associated with 85% of all cancers (Lipps and Rhodes, 2009, Lansdorp, 2009, Rezler et al., 2002). G4-forming sequences are found most commonly in telomeric DNA sequences (Williamson et al., 1989), poly d(G) runs (Panyutin et al., 1989) and in some of the gene expression regulatory sequences such as promoter regions (Siddiqui-Jain et al., 2002). Also, large numbers of genes contain G4-forming sites in the 3′ UTRs (end of the coding region), suggesting a role in facilitating the termination of transcription, possibly by promoting RNA polymerase complex dissociation (Huppert and Balasubramanian, 2007, Huppert et al., 2008).

There are number of DNA repair pathways and specialised helicases that can recognize these non-B structures and process them. For example, human helicases such as DHX9, ChlR1 and UPF1 have been shown to efficiently unwind triplex DNA (Jain et al., 2013, Guo et al., 2015, Dehghani-Tafti and Sanders, 2017). WRN and BLM RecQ helicase (Huber et al., 2002, Popuri et al., 2008), FANCJ (London et al., 2008, Wu and Spies, 2016, Wu et al., 2008) and the Pif1 helicase (Paeschke et al., 2013, Paeschke et al., 2011) are known to resolve G-quadruplexes. Recently, a crystal structure of a mammalian DHX3 helicase bound to a G-quadruplex DNA structure from a telomeric sequence has led to the proposal of a model for the mechanism of the recognition and resolution of this secondary DNA structure by the helicase (Chen et al., 2018).

1.2.2 The DNA –damage response and cancer treatment

Simplistically, cancer is caused by un-repaired mutations, deletions and duplications occurring in genes that are responsible for either encoding the factors that are involved in suppressing cell growth (tumour suppressors) or factors that induce cell growth (oncogenes). Chemotherapy and radiotherapy are designed to target DNA integrity by inducing DNA-damage that ultimately leads to apoptosis. Although these treatments are mainly designed to target the rapidly dividing cancer-cells, they unavoidably also affect normal cells. One consequence of this is that these treatments can potentiate the emergence of adapting secondary cancer. In cancer therapy there is a constant struggle to generate solutions to minimize the affective dose of these therapies, by identifying and targeting, with specific drugs, factors that are vital for cancer cell survival. Since genome stability pathways are essential for cancer cell survival, these have been keenly addressed as potential targets.

1.2.2.1 DNA-damage and repair

DNA replication is sensitive to damaged DNA. Damage is typically caused by chemical agents such as oxygen free radicals, base-reactive agents, DNA cross-link inducers and physical agents such as UV light and X-rays or spontaneous replication errors (mismatched bases, base insertion or deletions). Un-repaired errors and damage, if in tumour suppressor genes or oncogenes, can lead to cancer. There are several cellular checkpoints that function at different stages of the cell cycle to prevent cell cycle progression when cells carry unrepaired DNA damage: the G1/S and intra-S checkpoints stop incorrect DNA replication and G2/M checkpoint prevents cells that carry any DNA damage from entering mitosis. Research over the past few decades to understand the cell's DNA damage response mechanism has resulted in the proposal

that there are at least five main pathways that function in the DNA damage response; (1) direct repair (DR), (2) base excision repair (BER), (3) double-stranded recombination dependent repair, which includes both homologous recombination repair (HRR) and non-homologous end-joining repair (NHEJ), (4) nucleotide excision repair (NER) and (5) mismatch repair (MMR)(Bernstein et al., 2002) (Fig. 8).
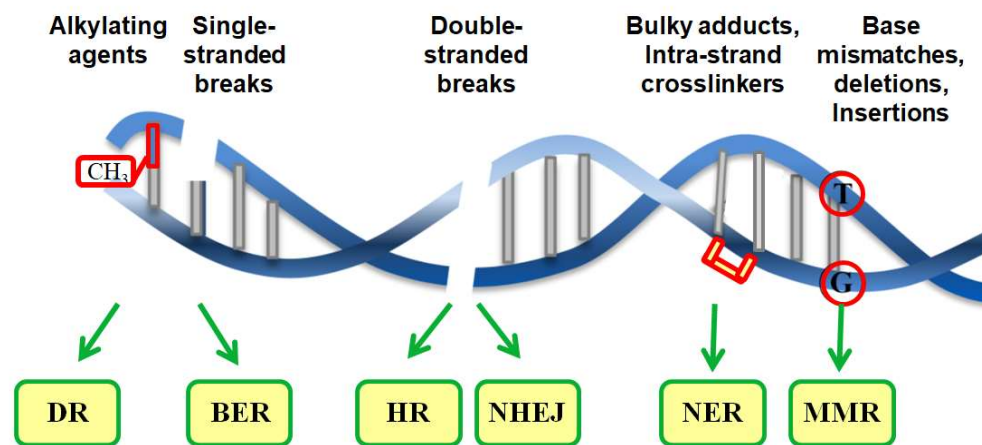
Figure 8. The major DNA repair pathways showing each pathway's specific type of DNA lesions that they are specialized to repair. Adapted from (Cervelli et al., 2012).

(1) The direct repair (**DR**) pathway repairs covalently modified bases by direct alteration without any base removal (Yi and He, 2013). Often, there is only a single protein or enzyme in the pathway. For example, O6-alkylguanine-DNA-alkyltransferase (AGT/MGMT) that relocates the methyl/alkyl groups from the O6 position of guanine to an active site cysteine residue in the enzyme (Hegi et al., 2004).

(2) The base excision repair (**BER**) pathway is activated when there is a damaged base caused by environmental mutagens or anti-cancer agents such as alkylating agents

and ionising radiation. The damaged base is removed by DNA glycosylase and the apurinic/ apyrimidinic endonuclease (APE1) which generates a single strand end. The damaged base is replaced and re-ligated by binding of poly ADP-ribose polymerase (PARP) followed by recruitment of DNA-polymerase β and DNA-ligase III to the damage site (Chan et al., 2006, Sharbeen et al., 2015).

(3) The double-stranded break repair (**DSBR**) pathway: Double-stranded breaks are the most cytotoxic kind of DNA lesions. These breaks can be caused by ionising radiation, anti-cancer agents, free radicals or when a replication fork encounters a single-stranded break. Double-stranded break repair comprises two main pathways; homologous recombination repair (HR), which is an error-free mechanism and an error-prone repair pathway, non-homologous end joining repair (NHEJ). The cell's choice of repair pathway depends on the stage of the cell-cycle and nature of the DSB (Symington and Gautier, 2011, Chapman et al., 2012). However, an important determinant for the choice is dependent on the 5´ and 3´ resection ends which triggers HR repair and blocks NHEJ pathways and the vice versa (Pierce et al., 2001). The HR repair pathway uses a sister chromatid as a sequence template in S and G2 phase of the cell-cycle, therefore this repair is almost error-free. HR repair also can function in G1 and uses the homologues chromosomes as the sequence donor. The main proteins in the HR pathway are BRCA1, BRCA2 and RAD51, which together promote strand invasion and homologous pairing (Dong and Fasullo, 2003, Aylon and Kupiec, 2004).

Although the NHEJ repair pathway is not an error-free mechanism, since it simply re-ligates DSB ends and can proceed with loss of DNA fragments, it remains the main DNA double-strand break repair pathway in G0/G1 phase. This repair mechanism involves factors such as Ku70/Ku80 to align the DNA break ends which is followed by

the recruitment and activation of the DNA dependent protein kinase catalytic subunit (DNA-PKcs) that processes the DNA ends and facilitates the ligation step for DNA-ligase IV and its accessory factors (Hefferin and Tomkinson, 2005).

(4) The nucleotide excision repair (**NER**) pathway is the most complex DNA repair pathway involving more than twenty different proteins. NER repairs bulky helix distorting DNA lesions such as pyrimidine dimers caused by UV irradiation. This repair system is subdivided into two main pathways: the global genome repair (GGR) and transcription coupled repair (TCR) pathways. TCR rapidly detects and removes any lesions in DNA that inhibit the progression of RNA Pol II during transcription, whereas GGR inspects the whole genome in a more-timely process. Xeroderma Pigmentosum and Cockayne syndrome patients are examples of a deficiency in NER and a corresponding propensity to develop skin cancers (Takebayashi et al., 2001, Erba et al., 2004, Kamileri et al., 2012).

(5) The mismatch repair pathway (**MMR**) system efficiently recognises and corrects (in most cases) base-base mismatches (A-C, T-G) and any base deletion/unpaired-insertions in a newly synthetized DNA strand. It has been demonstrated that an impaired function of this repair system has a direct link to pro-malignant progression (Jiricny, 2006).

Many chemotherapeutics are compounds that induce DNA inter-strand cross-links which covalently join both complementary DNA strands. These can be repaired by the NER pathway but they also block replication fork progression and transcription and ultimately causes DSBs that are often repaired by HR (Vare et al., 2012). Several factors in the repair pathways described above have been introduced as cancer treatment targets, as illustrated in Figure 9.
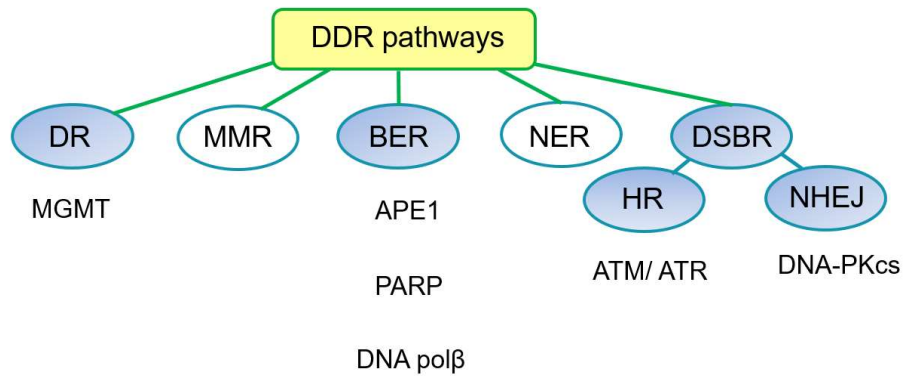
Figure 9. Diagram showing the DNA repair pathways that are currently being targeted for enhancing cancer treatments (highlighted in blue). The specific factors in each pathway that are being drug-targeted is listed below them. Modified from (Damia and D'Incalci, 2007).

1.2.3 Helicases involved in DNA repair pathways

Helicases are important factor in DNA repair pathways. They resolve non-B DNA structures (described above) and provide ssDNA for and drive recombination and replication re-start. The best characterized DNA helicases that are known to be essential in genome stability are the SF2 RecQ helicases (RECQL1, WRN, BLM, RECQL4 and RECQL5). These helicases are highly conserved and their deficiencies are associated with genetic disease that predispose to cancer; WRN helicase in Werner syndrome, BLM helicase in Bloom syndrome and the RECQL4 helicase is linked to Rothmund-Thomson syndrome (RTS), Baller-Gerold syndrome (BGS) and RAPADILINO syndrome (Siitonen et al., 2003, Kitao et al., 1999). Individuals carrying mutations in the *WRN* helicase gene show many characteristics of premature aging and usually develop a wide range of cancer such as soft-tissue sarcomas, osteo-sarcomas, lymphoid and haematological neoplasms, malignant melanomas and thyroid

epithelial tumours (Lauper et al., 2013). Patients affected by Bloom syndrome, RST and RAPADILINO also frequently develop a variety of cancer types earlier in their life compare to the normal population (Siitonen et al., 2009).

Although there are no genetic disorders directly linked to RECQL1 and RECQL5, laboratory studies have shown that these helicases play roles in genome integrity linked to cancer. RECQL1 siRNA–depleted human cells have increased chromosomal instability and are more sensitive to ionizing radiation which is in accord with an involvement in DSB repair (Sharma et al., 2007, Sharma and Brosh, 2007). RECQL5-defiecient mice have a high potential to develop cancer such as lymphomas or solid tumours. Further studies suggested that RECQL5 has a role in DNA repair through an involvement in HRR and BER pathways (Hu et al., 2007, Tadokoro et al., 2012).

Some of the most important helicases that are known to have a contribution in genome stability and cancer are the iron-sulphur cluster family helicases, FANCJ, XPD, XPB and RTEL1, that are in SF2 (Brosh Jr, 2013). Also included in the list of helicases involved in cancer is the Pif1 helicase, the topic of this thesis, which will be discussed further in detail below. It should also be noted that as well as defective helicases (loss of function) being associated with cancer predisposition, many helicases are also sharply up-regulated in human tumours (Kawabe et al., 2000, Abdelhaleem, 2004, Fang et al., 2013, Lin et al., 2015), and in response to chemotherapy (Parvathaneni et al., 2017). This overview highlights the importance of DNA helicases as key targets for improving cancer therapies as they are essential for maintaining genome integrity. Many of the above mentioned helicases have been proposed as drug targets for enhancing cancer treatments *e.g.* the RecQ helicases (Mao et al., 2010, Arai et al., 2011, Mendoza-Maldonado et al., 2011, Futami and Furuichi, 2014) and FANCJ helicase (Wu

et al., 2008, Gong et al., 2010). The context and rationale in which they can become targets is discussed below.

1.2.4   Synthetic lethality (SL)

As noted previously, DNA replication, recombination and repair pathways are suggested to be good targets for enhancing cancer therapy strategies that are based on radiation or chemotherapy in a combination therapy. Defects in DNA repair pathways allow cancer cells to accrue genomic deviations (the "mutator phenotype") that add to their aggressive nature. However, tumour cells still need to maintain a certain level of genome stability to escape cell-cycle checkpoints and apoptosis. In Figure 10 this concept is illustrated, showing that cells rely on at least one repair pathway to escape apoptosis. Cancer cells that are genetically deficient in one pathway (pathway A) are vitally dependent on the existence of the second related repair pathway (pathway B) for cell survival. Therefore, if by any synthetic means (e.g. siRNA knock-out or drug induced inhibition) pathway B is disabled, cancer cells that are entirely deficient in pathway A are drawn towards apoptosis. On the other hand, the synthetic deactivation of pathway B in normal cells is tolerated since they still have the alternative repair pathway (pathway A) and will survive. The selectivity of this kind of treatment (synthetic lethality, SL) is a powerful and positive aspect which the general chemotherapies and radiotherapy mostly lack.
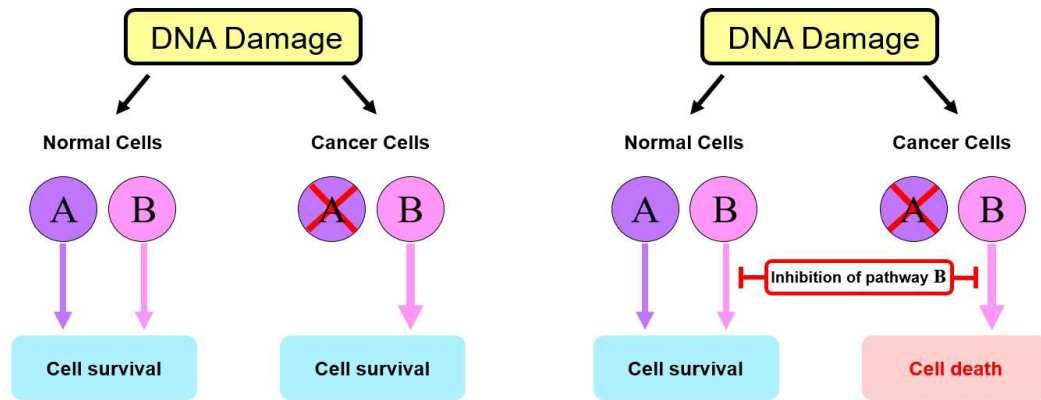
Figure 10. The general concept of synthetic lethality (SL). SL as applied to the DNA-damage response and cancer therapy.

1.2.4.1 PARP inhibitors: A prime example of synthetic lethality

Poly (ADP-ribose) polymerase (PARP1) is a key factor in the repair of single-strand breaks by the BER pathway (Banasik and Ueda, 1994, Sugimura and Miwa, 1994). PARP1 is a nuclear protein with a DNA binding domain which promptly detects DNA damage sites (single-strand breaks) and initiates DNA repair mechanisms. PARP inhibitors efficiently kill cancer cells defective in the BRCA1 or BRCA2 genes according to the theory of synthetic lethality. PARP1 siRNA knock-out in normal cells compared to BRCA1/2 deficient cancer cells demonstrates that these cancer cells are extremely sensitive towards PARP1 knock-out, whereas the PARP1 siRNA knock-out shows no cytotoxicity in normal cells. It is proposed that PARP1 inhibitors cause an increase in DNA single-strand breaks (SSBs), which during replication are transformed to toxic DNA double-strand breaks (DSBs). As mentioned previously, BRCA1 and BRCA2 are key factors in the HR repair pathway and if cells are BRCA deficient the HR repair pathway is blocked and therefor these cells are highly susceptible to toxic

dsDNA breaks and die by apoptosis. The cancer cell lines are otherwise dependent on the existence of BER pathway to escape cell-cycle checkpoint arrest and apoptosis (Bryant et al., 2005, Damia and D'Incalci, 2007), as illustrated in Figure 10.

Recently, targeting APE1 (another BER factor, see section 1.2.2.1) with inhibitors has also been shown to have a similar outcome as PARP1 inhibitors by inducing synthetic lethality (Sultana et al., 2012). Anti-APE1 drugs have now entered clinical trials (Shah et al., 2017).

1.3   The Pif1 helicase

The Pif1 helicases are DNA and RNA helicases that unwind in a directionally specific manner ($5' \rightarrow 3'$) and are classified in helicase SF1, so are monomeric enzymes (Bochman et al., 2010). The *PIF1* gene was first discovered as a factor involved in yeast *Saccharomyces cerevisiae* (*Sc*) mitochondrial DNA replication and stability (Foury and Kolodynski, 1983, Lahaye et al., 1991). Later, as more genetic data became available for other organisms, it became evident that this gene was the founding member of a highly conserved group of non-processive $5' \rightarrow 3'$ DNA helicase found in almost all eukaryotes and some prokaryotes (Bochman et al., 2010, Bochman et al., 2011). Nuclear *Sc*Pif1 has roles in Okazaki fragment maturation (Budd et al., 2006, Pike et al., 2009), telomere length regulation (Zhou et al., 2000), replication through loci that normally impede the replication fork, *e.g.* the rRNA Replication Fork Barrier (Ivessa et al., 2000, Tran et al., 2017), DNA strand annealing (Ramanagoudr-Bhojappa et al., 2014) and the resolution of G4 DNA structures (Ribeyre et al., 2009, Paeschke et al., 2011).

1.3.1 Human Pif1 activities linked to genome stability

Although some research has been done on characterising human Pif1's (hPif1) properties and functions, most of the *in vitro* and *in vivo* studies have focused on *Sc*Pif1. Accordingly, the functions of hPif1 remain largely unclear. The human genome encodes one *PIF1* gene (chr15, q22.31) which produces two known protein isoforms by alternative splicing. The major transcript product, hPif1a, is a nuclear protein with a molecular weight of approximately 68 kDa and the second transcript product, hPif1b, is a 75 kDa protein which is mainly localized in mitochondria (Futami et al., 2007, Mateyak and Zakian, 2006). Here I discuss the 68 kDa nuclear hPif1a, hereafter referred to as simply hPif1, and its known functions in genome stability maintenance.

As mentioned above, most of the available information on the human Pif1 helicase is based on *Sc*Pif1. However there are few cell-based studies that show hPif1s' importance in maintaining genome stability under oncogenic stress, demonstrating similarities with *Sc*Pif1. The important functions are characterized in three main categories:

1) Telomerase inhibition: Telomeres are nucleoprotein structures that are located at chromosome-ends. In humans, each chromosome end carries 5 to 15 kb of guanine rich TTAGGG/CCCTAA repeated nucleotide sequence. The main role of telomeres is to protect the chromosomal ends from both fusion with other chromosome-ends and from nuclease digestion. Telomere length is usually maintained by a reverse transcriptase enzyme (telomerase) that uses a small integral RNA fragment as a template to extend the 3′ end of telomeres. There are also low-level alternative lengthening of telomere (ALT) pathways that are less well understood (Bryan et al., 1997, Dilley and Greenberg, 2015). However, the expression and activity of telomerase

is minimal in most somatic cells which is consistent with shortening of telomere ends by each cell division that eventually drags them towards apoptosis. Many cancer cell lines show over-expression of this enzyme which helps them escape apoptosis due to telomere loss (Kim, 1997, McEachern et al., 2000). As this suggests, telomerase activity is essential for cancer cell survival and other studies have shown that the inhibition of telomerase significantly reduces cancer cell growth (Hahn et al., 1999).

Studies have shown that *in vitro Sc*Pif1 reduces the processivity of yeast telomerase (Boule et al., 2005), potentially explaining why yeast cells lacking Pif1 have longer than expected telomeres while its overexpression results in their shortening (Schulz and Zakian, 1994, Zhou et al., 2000). Similarly, it has been shown that hPif1 reduces the processivity and overall activity of human telomerase *in vitro* by dissociating telomerase RNA from telomeric DNA (Zhang et al., 2006).

2) Resolving G4 DNA structures: Resolving G4 structures is important for replication fork progression, as described above. *Sc*Pif1 helicase activity is involved in processing the G-quadruplex-forming sequence in the CEB1 locus *in vivo* to allow efficient replication (Ribeyre et al., 2009), and also bind G4 DNA *in vitro* (Paeschke et al., 2011). Later, it was shown that hPif1 is also able to bind and unwind tetramolecular G4-DNA *in vitro* (Sanders, 2010) and that hPif1 co-localises with G4 DNA forming sequences in cells (Rodriguez et al., 2012). Clearly, telomeres are also sites of G4 DNA formation where Pif1 may act.

3) Replication fork progression: *In vitro* experiments show that hPif1 is able to recognize and unwind DNA structures resembling a stalled replication fork. hPif1 displaces the leading replication strand from a fork-like substrate and it needs both arms of the DNA fork to re-start the unwinding. The result of this activity could be to

provide a free 3′ ssDNA ends that could potentially be used to promote replication re-start in combination with its strand annealing activity (George et al., 2009). These *in vitro* data are supported by a recent study showing that hPif1 depletion slows the rate of replication fork progression in normal cells and hence prolongs S-phase. The effect is significantly elevated in RAS oncogenic transformed cells (Gagou et al., 2014).

1.3.2 Targeting human helicases for cancer therapy

Two human helicases from the RecQ helicase family have been proposed as drug targets for cancer therapy WRN and BLM. These helicases are known to be involved in the DNA replication stress response (Brosh Jr, 2013). Both BLM and WRN have key roles in the HRR pathway. It has been claimed that the WRN inhibitor NSC617145 induces synthetic lethality. However, these studies required initial cellular defects in two repair pathways, the Fanconi Anemia pathway (FA), that functions in replication to repair a variety of lesions including inter-strand cross-links, and NHEJ as well as induction of damage with the DNA inter-strand crosslinking agent Mitomycin C (MMC) (Aggarwal et al., 2013). In conclusion, the effectiveness of this inhibitor requires defects in two major DNA repair pathways, so in total three repair pathways must be neutralised for cell death. This study has confounding problems. First, there have been no cancers reported to be deficient in both of these pathways (FA and NHEJ) and such cell lines have been created artificially using cells that are deficient in NHEJ and transfected with FANCD2 siRNA to knock-down the FA repair pathway. These cells showed a significant reduction in cell proliferation (even without being induced with MMC or any other DNA damaging agents) and then by adding MMC and WRN inhibitor (NSC617145) only a modest further decrease in cell proliferation was

obtained. This situation could be considered highly engineered or contrived and potential drugs therefor relatively ineffective. However, recent studies have shown that both NSC617145 and NSC19630 (another WRN inhibitor) can induce apoptosis in HTLV-1-transformed adult T-Cell leukaemia (ATL) (Moles et al., 2016) where the expression of WRN is highly up-regulated in this type of cancer and required for survival of ATL cells (Sallmyr et al., 2008). Therefore, the WRN inhibitors may find effective clinical use in this context.

BLM is necessary to suppress sister chromatid exchanges (SCE) in the HRR pathway. In cells deficient in this helicase major genome instability, elevated SCE levels and ultimately cancer are observed. This suggests that although it may be possible to use the BLM inhibitor (ML216) as an anti-cancer drug, there could be a major risk in developing secondary mutations in normal cells. In addition, ML216 does not seem to be highly specific to BLM as it also inhibits DNA unwinding *in vitro* of the closely related WRN helicase (IC_{50} of 2.7 μM compared to 1.8 μM for BLM) (Rosenthal et al., 2013, Datta and Brosh, 2018). Furthermore, the poor solubility and permeability of this compound series could prevent reliable *in vitro* data from being generated. Although it has been reported that no specific cell-cycle checkpoint activation have been observed in cells treated with BLM inhibitors (Nguyen et al., 2013) it is early days for the ML216 which needs to be further investigated.

1.3.3 hPif1 helicase as a potential cancer therapy target

It is likely from the arguments made above that both of the RecQ inhibitors may have significant problems for use as chemotherapy agents. One potential advantage of targeting Pif1 in cancer therapy is that studies have demonstrated that Pif1 knock-out

mice show no abnormal phenotypes and they continue to survive (Snow et al., 2007). On the flip side, hPif1's activity has not been linked to any specific repair pathways and no genetic disease has been linked to the defects or depletion of the human Pif1 helicase. However, studies have shown that the survival rate of some of p53-deficient and p53-proficient cancer cell-lines are reduced when hPif1 is knocked down (siRNA), and this effect is augmented in the presence of DNA-damaging agents (thymidine, hydroxyurea, or gemcitabine, an agent used in the clinic). In contrast, control normal Pif1 knock-down cells were not significantly affected with or without DNA-damaging inducers (Gagou et al., 2011). These findings clearly reveal the dependence of these cancer cell-lines on Pif1's presence, which is likely due to its contribution to replication fork integrity, cell cycle progression and ultimately genome stability.

There is a report of breast cancer predisposition linked to a specific hPif1 mutation (Chisholm et al., 2012) . However, as stated above, although so far, no genetic disease with abnormalities linked to hPif1 have been identified, there are activities linked to the hPif1 helicase which are functionally important for genomic integrity. This suggests that even though it has important contributions to genome stability, its activities may overlap with other nuclear helicases such as RecQ or FANCJ. This could be the reason that the hPif1-depleted normal cells survive while some cancer cell-lines with underlying DNA repair deficiencies are sensitive to its depletion when triggered by DNA-damage inducers.

In summary, the information above on hPif1 activity suggests that this helicase is a potential target for synthetically induced death (SL) in cancer treatment that will possibly have minimal undesirable side effects. The main challenge in developing an effective drug against hPif1 is primarily to have a complete insight into its function and

contribution to genome stability. This requires a full characterization of its biochemical properties and functions in genome stability in intact cells. The second challenge is to find an efficient and specific hPif1 inhibitor that will be stable in blood and efficiently taken up by cells (of course the drug resistant mechanisms adapted by tumour-cells always remain as a problem). Initial targeting of hPif1 can be achieved by small-molecule inhibitor library screening or structure-based drug design. The latter requires a high-resolution crystal structure of the protein in order to specifically design a molecule to bind to active sites critical for function. The next challenges, if the targeting approach is successful, are use of *in vivo* models to test activity and to find a genetic marker for its use in patients to identify appropriate candidates for treatment.

1.4   Hypothesis, aims and objectives

In accordance with the information summarised above, the overarching hypothesis under test is that the hPif1 helicase is a key factor in maintaining genome stability in a sub-set of cancer cells and that chemical inhibition of the helicase can induce synthetic lethality, mimicking the effects of siRNA depletion.

The ultimate aim of the research initiated here was to identify specific chemical inhibitors of hPif1 for functional testing in cell-based assays. As well as providing a starting point for subsequent drug development such tool compounds would help further characterize hPif1 functions related to the maintenance of genome stability.

The specific objectives were:

1. To obtain high-resolution crystal structures of the protein, with and without nucleotides and DNA substrates bound, to provide important information on the catalytic mechanism and indicate site for drug targeting.

2. To identify amino acids critical for binding ssDNA, dsDNA and G4 DNA and test models of function by site-directed mutagenesis and the biochemical analysis of purified variant forms. It was hoped that this analysis would reveal separation of function (*i.e.* G4 DNA binding, helicase and strand annealing activities) that could ultimately help understand functions in intact cells.

3. To identify specific inhibitors of hPif1's *in vitro* helicase activity and begin to understand their mode of action.

4. Obtain crystal structure of hPif1 with an inhibitory or non-inhibitory molecule bound. This information would then be used to design a more specific and potent drug-like molecule or series of molecules that can be tested *in vitro* and *in vivo* to characterise their inhibitory function. This approach is known to be particularly useful for fragment-based drug design (FBDD).

2 Materials and methods

2.1   DNA manipulation

2.1.1   Generation of expression constructs

The *E. coli* DNA expression constructs used were created by inserting the protein coding sequence of hPif1 ORFs into pET11c or pET15b based plasmid vectors (Novagen; T7 expression system, shown below). *E. coli* DH5α was employed for cloning. The purified plasmids were then used to transform the expression-proficient *E. coli* strains based on BL21(DE3), as indicated. Three constructs expressing different hPif1 proteins were finally employed in the study that are hereafter referred to as follows: (i) FL-hPif1, full-length ORF, residues 1-641 (69.9 kDa), (ii) hPif1$_{206-641}$, residues 206 to the C-terminal end of the protein (47.3 kDa) and (iii) hPif1-HD, helicase core, residues 206-620 (45.1 kDa). The wild-type constructs were originally generated by PCR cloning (see below) using the following primers (restriction enzyme sites employed in bold, *Nde*I, *Bgl*II and *Bsa*I):

PifNde: 5′-ATCGTCTAGA**CATATG**CTGTCTGTATTGAAGCAAACAG-3′

Pif206: 5′-ATGCTCTAGA**CATATG**CAGCTGTCTGAAG-3′

Pif620: 5′-ACGC**AGATCT**TCACAGGCTACGACCACGACG-3′

PifSTOP: 5′-CTG**GGTCTC**GGATCCTCCTCACAGAATCGGATCCATATTTTCTT-3′
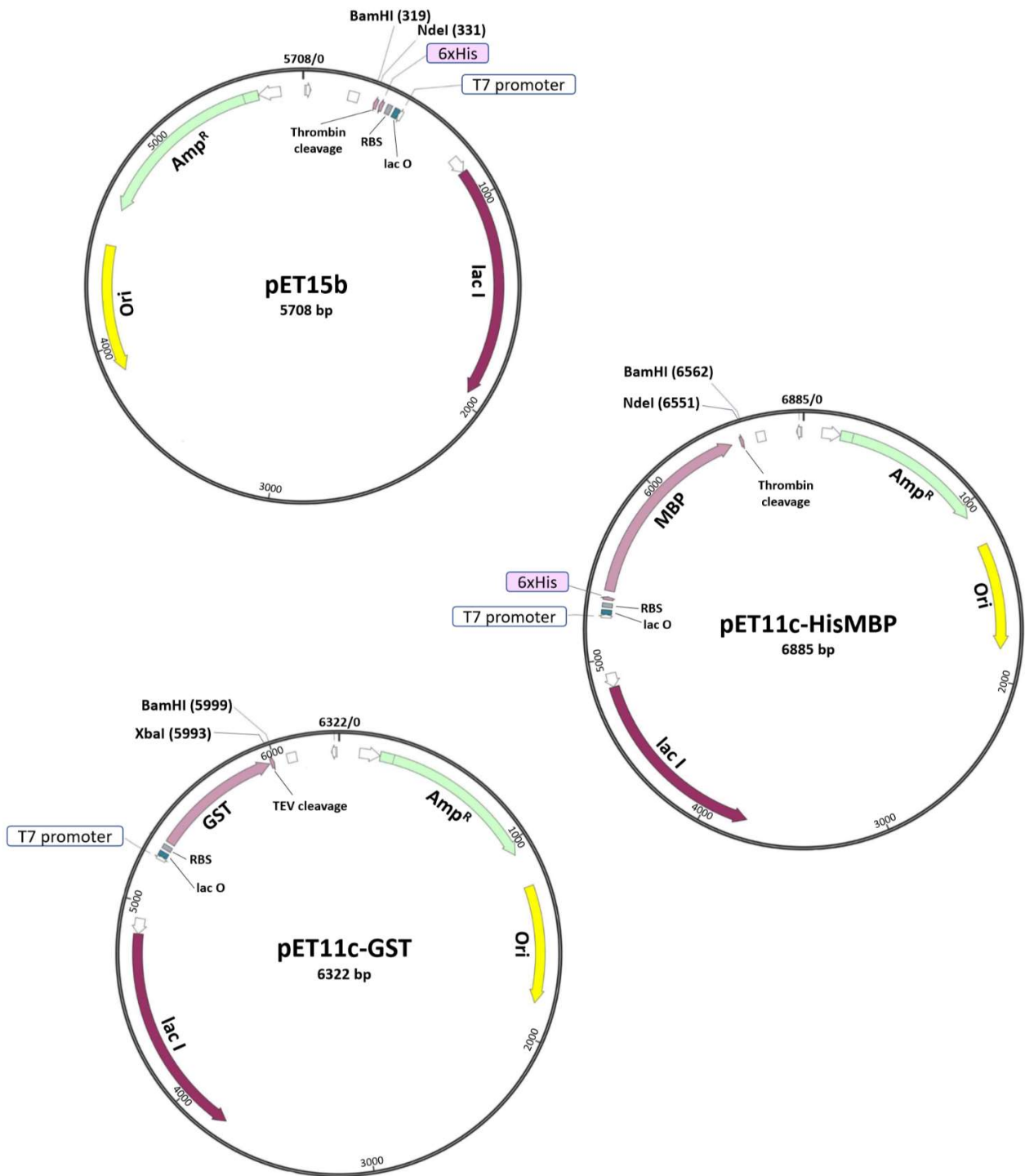
Figure 11. Plasmids used for cloning different hPif1 constructs. The hPif1 protein-coding DNA sequences were inserted between the *BamH*I and *Nde*I digestion sites of pET15b or pETHisMal, or a methylation sensitive *Xba*I site and *BamH*I in pETGST (generated by Dr. C.M. Sanders, unpublished). Vector maps were created using SnapGene viewer.

2.1.2 Agarose gel electrophoresis

DNA fragments and plasmids were analysed by running on standard 1% (w/v) agarose gels containing 0.5 µg/ml ethidium bromide for staining double-stranded DNA. DNA samples were loaded on the gel with DNA-loading dye (20% (w/v) Ficoll 400, 0.1 mM EDTA (pH 8.0), 1% (w/v) SDS, 0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol) and run at a constant voltage of 140 V in 1 × TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA).

2.1.3 Polymerase chain reaction (PCR)

PCR was performed in a 100 µl reaction containing: 1 × Pfu buffer, 1 ng/µl DNA template (hPif1 DNA sequence in a pET expression vector), 1 µM of each primer, 200 µM each dNTP, 2 mM $MgCl_2$ and 0.05 U/µl Pfu DNA polymerase (Promega). The PCR program was set for 25 cycles with an annealing temperature of 55°C for 60 seconds and extension times at 72°C for 60-120 s depending on the primer pairs.

2.1.4 DNA digestion

1-10 µg of DNA (plasmid vectors, as described above or PCR product) was digested with 10-100 Units of restriction enzyme in the appropriate digestion buffer according to manufacturer's instructions (New England Biolabs, NEB). The reactions were incubated at optimal temperatures for 1-2 hours. The digested fragments were gel-purified using QIAquick gel-extraction kit (Qiagen) according to manufacturer's instruction. Purified DNA fragments were quantified against 1 µl of λ DNA-BstEII marker (500 ng) on a 1% (w/v) agarose gel as described above.

2.1.5 DNA Ligation

The target DNA fragments were inserted into the appropriate vector using approximately 2-3 ng of digested and gel-purified target DNA fragment (insert) with 10 ng of digested and gel-purified plasmid vector combined in a volume of 1:1. After mixing and centrifugation the mix was heated at 68°C for 3 minutes and snap-cooled on ice for a minute. It was then mixed and centrifuged down and the ligation reaction was carried out by adding an equal volume of 2 × ligase buffer, 7.8% (w/v) PEG 8000, approximately 80 Unit of T4 DNA ligase enzyme (NEB). The reaction was incubated at room temperature for 10-30 minutes and was placed on ice for 5 minutes before transformation of DH5α.

2.1.6 Competent cell transformation

Chemically competent cells were generated by the method of Inoue et al. (Inoue et al., 1990). DH5α competent cells were transformed by adding the cells directly to the ligated DNA. After 20 minutes incubation on ice the cells were heat shocked for 30 seconds at 42°C and immediately cooled on ice for 2 minutes. The cells were then incubated in a 37°C shaking incubator for an hour in 400 μl of outgrowth media (1 × LB (10 g/L tryptone, 5 g/L yeast extract and 10 g/L NaCl)), 0.4% (w/v) glucose, 25 mM KCl, 10 mM $MgCl_2$) and were plated on LB agar plates (1 × LB, 1 × agar (15 g/L), 100 μg/ml ampicillin or carbenicillin). The plates were incubated at 37°C overnight.

2.1.7 Plasmid preparation

Single DH5α colonies (successfully transformed) were cultured in 100 ml of 1 ×
LB and 100 μg/ml ampicillin at 37°C overnight. The bacteria were pelleted by
centrifugation (3000 × g, 5 mins.). Plasmids were isolated from bacteria using the
standard alkaline lysis method (Birnboim and Doly, 1979) and purified further by LiCl
and PEG precipitation (RNA and DNA respectively) according to standard protocols
(Sambrook and Russell, 2001). DNA was analysed and quantified on agarose gels
(0.8% (w/v)) against a standard dilution series of plasmid (for quantification) and λ
DNA-BstEII marker as described above.

2.1.8 DNA sequencing

The sequence of the inserted DNA fragment was confirmed by sequencing at the
Genetics Core Facility, The University of Sheffield using a BigDye Terminator v3.1
Cycle Sequencing Kit and ABI PRISM™ 3730 DNA analyser (Applied Biosystems).

2.1.9 Creating the hPif1 mutant proteins

The mutant hPif1 ORFs were created by the overlapping primer extension method
(Ho et al., 1989) in three steps as shown in Figure 12. All the mutant proteins tested
were based on the His-tagged hPif1-HD (residues 206 to 620) expression construct.

The primary PCR (section 2.1.3) used a linearized pET11c plasmid containing the
full-length His-tagged hPif1 as a template. One PCR reaction (R1) contained a forward
primer carrying the mutation (P1′, Table 1) and a reverse primer (P1, Pif620, see above)
complementary to the top strand with a stop codon and restriction enzyme site

(generating a *BamH*I compatible cohesive end) down stream of codon 620. PCR reaction R2 containing the forward primer P2 (PifNde) with an *Nde*I restriction enzyme site upstream of hPif1 codon 206 and a reverse primer carrying the mutation (P2′). PCR products were resolved on agarose gels (1% (w/v), section 2.1.2) and purified with a QIAquick gel extraction kit (Qiagen) according to the manufacturers' instructions.

For the secondary PCR reaction, the gel-purified products of the primary PCR (which contains an overlapping sequence at the mutation site, ~15 base pairs) were used as a template. The primers P1 and P2 were used as forward and reverse primers to generate the mutated sequence in the hPif1-HD ORF. This piece of DNA was purified using the QIAquick reaction clean-up protocol, digested with *Nde*I and *BamH*I, agarose gel-purified and cloned into the original pET15b plasmid digested with the same enzymes.
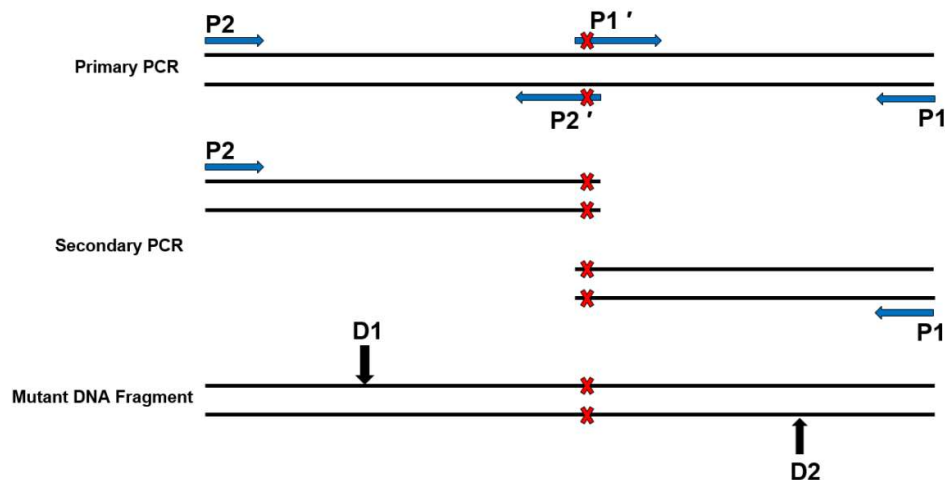
Figure 12. Schematic illustration of the overlapping primer extension procedure. Red Cross indicates the mutation site, primers labelled as (P) and digestion enzyme labelled as (D).

Primer	Sequence
R290A_T	5′-**GCA**CCTGGTGTGCGTCAGGGT-3′
R290A_B	5′-CTGACGCACACCAGG**TGC**CTGGGCCAGTGCAA-3′
P291A-T	5′-**GCA**GGTGTGCGTCAGGGTTG-3′
P291A-B	5′-ACCCTGACGCACACC**TGC**CCGCTGGGCCAGTGCAAC-3′
R294_T	5′-**GCA**CAGGGTTGGCTGAATTGT-3′
R294_B	5′-ACAATTCAGCCAACCCTG**TGC**CACACCAGGCCGCT-3′
H269A_T	5′-**GCA**GCATTTGCCGGTATTGGT-3′
H269A_B	5′-CAATACCGGCAAATGC**TGC**CAGGGTGGTGCC-3′
E307Q-T	5′-**CAA**ATTAGCATGGTTGAAGC-3′
E307Q-B	5′-AACCATGCTAAT**TTC**ATGAATCACCAGTCGCTG-3′
N495A_T	5′-**GCA**GGTGCACGTGGTGTTGTT-3′
N496A_B	5′-AACAACACCACGTGCACC**TGC**AACCAGACCACGGC-3′
H555A-T	5′-**GCA**AAAAGCCAGGGTATGAC-3′
H555A-B	5′-CATACCCTGGCTTTT**TGC**AATGCTCATTGCCCA-3′
K556A-T	5′-**GCA**AGCCAGGGTATGACACT-3′
K556A-B	5′-TGTCATACCCTGGCT**TGC**ATGAATGCTCATTGC-3′
F573A-T	5′-**GCA**GCAAGCGGTCAGGCCTA-3′
F573A-B	5′-GGCCTGACCGCTTGC**TGC**AACACGACCCAGAGAA-3′
K485E-T	5′-**GAA**AATCTGAGCGTTAGCCG-3′
K485E-B	5′-CTAACGCTCAGATT**TTC**AACCAGCATAACCT-3′
N486A-T	5′-**GCA**CTGAGCGTTAGCCGTGG-3′
N486A-B	5′-CGGCTAACGCTCAG**TGC**TTTAACCAGCATAA-3′
K414A_T	5′-**GCA**GTTGGTCGTGATGGTATT-3′
K414A_B	5′-AATACCATCACGACCAAC**TGC**ATGGCTTGCGGTGG-3′

Table 1. List of forward (T) and reverse (B) primers used to create the hPif1-HD mutants. Mutated codons are in bold.

2.2 Protein production and analysis

2.2.1 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

The molecular weight and purity of protein samples were analysed by running them on an SDS-PAGE gel. A standard gel consisted of 10% (w/v) acrylamide: bis-acrylamide 29:1, 375 mM Tris-Cl pH 8.8, 0.1% (w/v) SDS with a 5% (w/v) stacking gel (5% acrylamide:bis-acrylamide 29:1, 125 mM Tris-Cl pH 6.8, 0.1% (w/v) SDS) and were set with 1% (w/v) APS and 0.1% (v/v) TEMED. Protein samples were dissolved in 1 × Laemmli sample buffer (62.5 mM Tris-Cl pH 6.8, 10% (v/v) glycerol, 0.1% (w/v) bromophenol blue, 2% (w/v) SDS, 2.5% (v/v) 2-Mercaptoethanol) and were boiled (5 mins) before loading on the gel. The gels were run at a constant voltage of 200 V in SDS running buffer (25 mM Tris-base, 0.1% (w/v) SDS, 0.2 M glycine) and stained with Coomassie-blue solution (1.2% Coomassie brilliant blue, 40% (v/v) methanol, 10% (v/v) acetic acid) and de-stained with de-staining solution (7.5% methanol, 7.5% acetic acid).

2.2.2 Protein concentration measurement by Bradford assay

Protein concentrations were determined by Bradford assay (Bradford, 1976) using the BioRad kit according to the manufacturer's instructions using bovine serum albumin (BSA) as a standard. All measurements were obtained in triplicate.

2.2.3 Protein concentration measurement by Abs$_{280}$

Molar extinction coefficients (280 nm) were calculated using an online tool (https://web.expasy.org/protparam/) with cysteine residues reduced. The protein concentration was estimated by Bradford assay and the protein diluted in 100 µl volumes to give readings between 0.05 and 0.5 absorbance units after addition of 700 µl of 8 M guanidinium hydrochloride (800 µl total sample volume). Samples were heated to 55°C for 5 minutes and allowed to cool to room temperature before mixing and spinning briefly. Abs$_{280}$ readings were obtained in a quartz cuvette cleaned with chromosulphuric acid (92% cc. H_2SO_4, 2% K_2CrO_7, 6% milliQ H_2O). Protein concentration was determined using the Beer-Lambert law.

2.2.4 Bacterial transformation and large-scale culture

1 ng of purified plasmid was used to transform 100 µl of BL21(DE3) (B F$^-$ *ompT gal dcm lon hsdS$_B$($r_B^-m_B^-$)* λ(DE3 [*lacI lacUV5-T7p07 ind1 sam7 nin5*]) [*malB$^+$*]$_{K-12}$(λ^S)), or the ArcticExpress^{TM} derivative (Agilent Technologies) harbouring a plasmid expressing the cold chaperone Cpn10/60 from *Oleispira antarctica* (gentamicin^r), using the transformation procedure described in section 2.1.6. Transformed cells were plated on expression agar plates (1 × agar (15 g/L); M9 salt, ZB broth, $MgSO_4$ and glucose (see below), 100 µg/ml carbenicillin and 100 µg/ml gentamicin as appropriate) and were grown overnight at 37°C. Colonies were grown in 20 ml of complete expression media and incubated at 37°C for 7-8 hours (day culture). The starter culture was then used to inoculate 30 litres of expression media (ZB broth, 10 g/L tryptone, 5 g/L yeast extract and 5 g/L NaCl, supplemented with 100 µg/ml carbenicillin, 0.4% (w/v) glucose and 1 × M9 salts (33.7 mM Na_2HPO_4, 22 mM KH_2PO_4, 9.35 mM NH_4Cl)). After

inoculation they were incubated in shaking incubators overnight at 22°C. At an OD_{600} 0.8 they were induced with 0.4 mM IPTG.

Cells were harvested by centrifugation at $3200 \times g$ for 15 minutes (Beckman Coulter J2-21, JA-14 rotor). The bacterial pellets were washed with buffer (20 mM Tris-Cl pH 7.5 and 200 mM NaCl), re-pelleted, snap-frozen in liquid nitrogen and stored at -80°C.

2.2.5 Purification of hPif1-HD

2.2.5.1 Expression of hPif1-HD

Expression of a His-tagged Pif1-HD fusion protein was performed in Arctic Express^{TM} BL21(DE3) cells for 3 days at 6°C. Wild-type protein for crystallography required 60-80 litres of culture (~200 g cells). Mutants for biochemistry were produced from 18 litres (~50 g cells).

2.2.5.2 Total protein extraction

All protein purification procedures were carried out at 4°C. After measuring the total mass of bacteria, the cells were resuspended in 1.5 ml per gram of cell mass of low-salt lysis buffer (50 mM Tris-Cl pH 7.5 (10% (v/v) glycerol, 5 mM DTT, 0.1 mM NaCl, 1 mM PMSF). The cells were lysed by adding 1/50 of volume of 50 mg/ml lysozyme for 30 minutes at 4°C. An equal volume of high-salt lysis buffer (50 mM Tris-Cl pH 7.5 (10% (v/v) glycerol, 5 mM DTT, 1 M NaCl, 1 mM PMSF) was added

before sonication. After sonication with brief pulses the lysate was cleared by centrifugation for 30 minutes at $25000 \times g$. 0.125 volume of 5 M NaCl was added and the high molecular weight DNA/RNA content of the lysate was removed by adding a 1/20 volume of Polymin P and incubation for 5 minutes followed by centrifugation at $25000 \times g$ (10 minutes). The supernatant was recovered and protein was precipitated by adding 0.226 g/ml ammonium-sulphate (40% cut, 20 minutes' incubation) and centrifugation at $25000 \times g$ force for 20 minutes. The precipitated protein pellet was dissolved in 1 ml per gram cell mass of bacteria of His-Trap IMAC buffer (50 mM Tris-Cl pH 7.5 (0.5 M NaCl, 2.5 mM DTT, 10% (v/v) glycerol, 20 mM imidazole).

2.2.5.3 His-trap HP IMAC column

His-tagged protein was purified on a His-Trap^{TM} HP nickel column (GE Healthcare) at the ratio of 1 ml of resin per 8 g starting mass of bacteria. Column equilibration was carried out by running 5-10 column volumes of high-imidazole buffer (50 mM Tris-Cl pH 7.5, 0.5 M NaCl, 200 mM imidazole, 2.5 mM DTT, 10% (v/v) glycerol) followed by 5-10 vols. of low-imidazole buffer (50 mM Tris-Cl pH 7.5 (0.5 M NaCl, 20 mM imidazole, 2.5 mM DTT, 10% (v/v) glycerol). The protein sample was loaded onto the column, washed with 5 column volumes of low-imidazole and eluted with an imidazole gradient (15 column vols.) to 200 mM.

2.2.5.4 Thrombin cleavage

Peak fractions of hPif1-HD were identified by SDS-PAGE and pooled. The protein concentration was measured by Bradford assay and 1 units of thrombin per

milligram of protein was added. The cleavage reaction was dialysed against 50 volumes of dialysis buffer (50 mM Tris-Cl pH 7.5, 0.3 M NaCl, 2.5 mM DTT, 10% (v/v) glycerol) for at least 1 hr. The buffer was changed (50 vols. fresh) and the protein digested overnight.

2.2.5.5 Second His-trap column

The digested protein was purified on a second His-trap column (~ 1ml resin per 40 grams of cell starting mass) as described above, with 2 mM low imidazole and 80 mM high imidazole buffers with a two column volume wash step following the load. Peak fractions were identified by SDS-PAGE.

2.2.5.6 Size exclusion column (SEC) chromatography

A single buffer was used for equilibration and protein elution, which contained: 20 mM Tris-Cl pH 7.5, 300 mM NaCl, 1 mM EDTA, 5% (v/v) glycerol, 2.5 mM DTT and 1mM PMSF. A Superdex 75 (16/70, 130 ml bed volume) was pre-equilibrated with 2-3 column volumes of buffer and protein eluted at a flow rate of 0.8 ml/minute. 2 ml fraction volumes were collected. Peak fractions were identified by SDS-PAGE.

2.2.5.7 Source S cation exchange

A high-resolution Source S (cation exchange) column (GE Healthcare) was used to separate the protein molecules according to charge. The column was equilibrated

with 5-10 column volumes of high-salt buffer (20 mM sodium-phosphate pH 7.2, 400 mM NaCl, 0.05 mM EDTA, 5% (v/v) glycerol, 2.5 mM DTT, 0.1 mM PMSF) followed by low-salt buffer (20 mM sodium-phosphate pH 7.2, 100 mM NaCl, 0.05 mM EDTA, 5% (v/v) glycerol, 2.5 mM DTT, 0.1 mM PMSF). The protein solution was diluted with 2.5 volumes of no-salt buffer (20 mM sodium-phosphate pH 7.2, 0.05 mM EDTA, 5% (v/v) glycerol, 2.5 mM DTT, 0.1 mM PMSF), loaded onto the column and washed with 5 column volumes of low-salt buffer before elution by inducing a salt gradient (100 mM NaCl to 400 mM NaCl). Peak fractions were analysed by SDS-PAGE and dialysed as described in section 2.2.5.4 except the glycerol concentration was reduced to 2% (v/v). The protein was concentrated by dialysis (3000 MWCO Spectra/Por membrane) against dried PEG 35000, before snap-freezing in liquid nitrogen and storage at -80°C.

2.2.6 Purification of hPif1$_{206-641}$

2.2.6.1 Expression of hPif1$_{206-641}$

Expression of a His-Maltose binding protein (His-MBP) tagged fusion protein was performed in Arctic Express^{TM} BL21(DE3) cells for 2 days at 6°C wild-type protein for crystallography required 60-80 litres of culture (~200 g cells).

2.2.6.2 Protein extraction, IMAC chromatography and Thrombin cleavage

The hPif1$_{206-641}$ fusion protein was extracted and purified by IMAC as described for hPif1-HD (Section 2.2.5.2 and 2.2.5.3). The pooled fractions identified by SDS-

PAGE were digested with 0.5 units of thrombin per mg of protein and dialysed simultaneously against 4 litres of dialysis buffer (50 mM Tris-Cl pH 7.5, 0.3 M NaCl, 2.5 mM DTT, 4% (v/v) glycerol).

2.2.6.3 Size exclusion column (SEC) chromatography

The protein sample was concentrated using Vivaspin 20 (5000 MWCO) devices (4000 × *g*). A Superdex 200 column (GE Healthcare 16/100, 200 ml bed volume) was pre-equilibrated with 2-3 column volumes of buffer and developed as described in Section 2.2.5.6. Peak fractions were identified by SDS-PAGE.

2.2.6.4 Second His-trap HP IMAC column

The buffer used was 50 mM Tris-Cl pH 7.5, 0.5 M NaCl, 2.5 mM DTT and 5% (v/v) glycerol. The column (~ 1 ml resin per 40 grams of cells) was equilibrated and the sample (SEC peak fractions) applied immediately thereafter before developing a gradient to 50 mM imidazole over 15 column volumes. Peak fractions were dialysed overnight against 10 mM Tis-Cl pH 7.5, 300 mM NaCl, 5% (v/v) glycerol, 2.5 mM DTT, 0.1 mM EDTA and 0.1 mM PMSF.

2.2.6.5 Source S cation exchange

A high-resolution Source S column was equilibrated with 5-10 column volumes of high-salt buffer (20 mM sodium-phosphate pH 7.2, 300 mM NaCl, 0.05 mM EDTA, 5% (v/v) glycerol, 2.5 mM DTT, 0.1 mM PMSF) followed by low-salt buffer (20 mM sodiumphosphate pH 7.0, 100 mM NaCl, 0.05 mM EDTA, 5% (v/v) glycerol, 2.5 mM

DTT, 0.1 mM PMSF). The protein solution was diluted with 3 volumes of no-salt buffer (20 mM sodiumphosphate pH 7.2, 0.05 mM EDTA, 5% (v/v) glycerol, 2.5 mM DTT, 0.1 mM PMSF), loaded onto the column and washed with 5 column volumes of low-salt buffer. The protein was eluted by inducing a salt gradient (100 mM NaCl to 300 mM NaCl). Peak fractions were analysed by SDS-PAGE and dialysed as described in section 2.2.6.2 except the glycerol concentration which was reduced to 2% (v/v). The protein was concentrated by dialysis (3000 mwco Spectra/Por membrane) against dried PEG 35000, before snap-freezing in liquid nitrogen and storage at -80ºC.

2.2.7 Purification of FL-hPif1

2.2.7.1 Expression of FL-hPif1

Expression of a His$_6$-tagged fusion protein with an intervening TEV protease cleavage site was performed in Arctic Express^{TM} BL21(DE3) cells for 2 days at 9°C. The starting culture volume for purification was 60-80 litres of culture (~200 g cells).

2.2.7.2 Total protein extraction and SP-Sepharose chromatography

All purification steps were at 4ºC. Total protein was extracted from 200 g of cells as described in section 2.2.5.2 up to the point of ammonium sulphate precipitation. The precipitated protein pellet was dissolved in 1 ml per gram starting cell mass of bacteria of S-buffer (20 mM sodiumphosphate pH 7.2, 0.05 mM EDTA, 5% (v/v) glycerol, 2.5 mM DTT, 0.1 mM PMSF, 50 mM NaCl) and spun for 20 minutes at 25000 × g. The supernatant was diluted with a further 5 volumes of S-buffer/ 50 mM NaCl

before loading a 30 ml SP-Sepharose (GE Healthcare) ion exchange column equilibrated in the same buffer. The column was washed with 1 volume of low salt buffer and a gradient developed to 1 M NaCl over 10 column volumes. The fractions from 15 to 30 mSv conductivity were pooled and 1/100 volume of 2 M imidazole added.

2.2.7.3 IMAC chromatography, TEV protease cleavage, SEC, and Source S chromatography.

IMAC chromatography was performed as described in section 2.2.5.3. The pooled peak fractions were digested with a 1/100th mass of TEV protease (Sigma) and dialysed against 2 litres of 50 mM Tris-Cl pH 7.5, 0.3 M NaCl, 2.5 mM DTT, 10% (v/v) glycerol overnight at 4ºC The cleaved protein was concentrated using Vivaspin 20 (5000 mwco) devices (4000 × g) and SEC chromatography performed as described in Section 2.2.6.3 (Superdex 200 column). Peak fractions were identified by SDS-PAGE and a second IMAC column (1 ml) run as described in Section 2.2.5.5 (hPif1-HD), with a gradient from 2-100 mM imidazole. Peak fractions were further purified by Source S chromatography (1 ml column) as described in Section 2.2.5.7 (hPif1-HD).

2.3 X-ray crystallography and protein structure determination

Initial crystallisation conditions were identified at 4ºC from screens using 150 nL of precipitant solution and 150 nL of Pif1 proteins (10-15 mg/ml). Crystals of the hPif1-HD/ AMP-PNP complex were grown in hanging drops at 4ºC by mixing the

protein with a final concentration of 11 mg/ml in 10 mM Tris-Cl pH 8.0, 2% (w/v) glycerol, 5 mM $MgCl_2$, 5 mM AMP-PNP, 2.5 mM TCEP, 0.3 M NaCl with a reservoir solution in a 1:1 ratio. The well solution contained 0.3 M Na-acetate, 17-27% (w/v) PEG 2K MME in 0.1 M Tris-Cl pH 7.5. For phasing, crystals were soaked briefly in mother liquid supplemented with 0.8 M KBr. All diffraction data were collected at cryo-conditions using synchrotron radiation at Diamond Light Source (DLS, Oxford) and processed with XDS (Kabsch, 2010). Data sets at three different wavelengths for Br-containing crystals (supplementary Table 1) were collected for structure determination by the multiple-wavelength anomalous dispersion (MAD) using SHELX (Sheldrick, 2008). Crystals belonged to the $P2_12_12_1$ space group and contained two molecules in asymmetric unit.

Co-crystals of hPif1-HD-AMPPNP in a different space group, $C222_1$, diffracting to 1.13Å resolution and containing one molecule per asymmetric unit, were grown using 11 mg/ml protein solution in 0.1 M Tris-Cl pH 7.5, 1% (w/v) glycerol, 5 mM $MgCl_2$, 5 mM AMP-PNP, 2.5 mM TCEP, 0.3 M NaCl, 0.3 M Na-acetate, 10% (w/v) PEG 8K and 10% (w/v) PEG 1K. The Apo structure of $Pif1_{206-641}$ was obtained from crystals belonging to the $P2_12_12_1$ space group with one protein molecule per asymmetric unit grown using 15 mg/ml protein solution in 0.1 M MES pH, 0.5% (w/v) glycerol, 5 mM $MgCl_2$, 2.5 mM TCEP, 150 mM NaCl 0.2 M Li_2SO_4, and 25% (w/v) PEG 2K MME at 4°C. ADP•AlF_4^--hPif1-HD co-crystals in the $P3_121$ space group with two molecules per asymmetric unit were grown using 13 mg/ml protein solution in 0.1 M MES pH 6, 1% (w/v) glycerol, 5 mM $MgCl_2$, 5 mM ADP, 6 mM $AlCl_3$, 50 mM NaF, 150 mM NaCl, 2.5 mM TCEP, 0.2 M Ca-acetate, 8% (w/v) PEG 20K and 8% (w/v) PEG 550 MME. The coordinates of the refined Br derivative structure were used as a molecular replacement search model using the data obtained from the hPif1-HD-AMP-PNP, apo

hPif1$_{206-641}$ and hPif1-HD-ADP•AlF$_4^-$ crystals. An unambiguous solution was found using MOLREP (Vagin and Teplyakov, 1997). All atomic models were built with COOT (Emsley and Cowtan, 2004), and refined using restrained maximum likelihood approach implemented in REFMAC (Murshudov et al., 1997).

2.4 Enzymatic assays

2.4.1 Radio-labelling DNA oligonucleotides

The radio-labelling reactions were carried out in low-binding siliconized tubes containing 5 µM of the selected oligonucleotide (sequence shown below), [γ-^{32}P]ATP (1800 Ci/mmol/µl), 1 × polynucleotide kinase (PNK) buffer, 1 U/L T4 PNK enzyme (NEB). The reaction was incubated in a 37°C water-bath for an hour and was stopped by heating at 95°C. 0.5 µl of the labelled oligonucleotide (0.25 pmol) was added to 24.5 µl of stop buffer (1% SDS, 10 mM EDTA) as a reference for quantification (see below).

2.4.2 Generating partially single- and double-stranded test substrates

The substrate used for the helicase assays was a 20 bp duplex with a 55 base single-stranded poly-T tail generated from the following oligonucleotide: 5′-(dT)$_{55}$CGAATTCGAGCTCGGTACCC and 5′-dGGGTACCGAGCTCGAATTCG. For strand annealing assays the oligonucleotide with sequence 5′-(dT)$_{55}$CGAATTCGAGCTCGGTACCC was annealed to 5′-GGGTACCGAGCTCGAATTCG-(C)$_{30}$ to give a fork-like structure. The strand annealing reaction was done in annealing buffer (1 mM Tris-Cl pH 8.0, 0.1 M NaCl, 0.1

mM EDTA) with an equal amount of complementary DNA strands (2.5 μM final concentration), one or both strands radiolabelled. The reaction was heated in a beaker of boiling water (~ 400 ml) for 5 minutes and gradually cooled to ~ 20°C with stirring at 4°C. The substrates were resolved on 8% (19:1) polyacrylamide gels (1 × TBE running buffer; 89 mM Tris, 89 mM boric acid, 2 mM EDTA), visualised by autoradiography, the bands excised from the gel and soak eluted in 100-200 µl of elution buffer (1 mM Tris-Cl pH 8.0, 0.1 M sodium chloride, 0.1 mM EDTA).

Substrate concentration was determined with reference to the specific activity of the ^{32}P label oligonucleotide control (section 2.4.1). 1 µl of control sample and 1µl of the gel purified radio-labelled DNA substrates were sampled on a DEAD81 anion-exchanger chromatography paper in triplicate. After drying, the paper was washed four times with 0.5 mM sodium phosphate buffer pH 7.0, followed by washes with 70% and 100% ethanol. The paper was dried and exposed to a phospho-imaging screen and imaged for quantification using a Fujifilm FLA3000 phospho-imager. According to photo stimulated luminescence of each, the concentration of the substrates were calculated from the specific activity of the labelled oligonucleotides.

2.4.3 Generating a tetramolecular G4 DNA substrate

G4 DNA binding reactions were performed with a radiolabelled tetra-molecular G4 DNA substrate formed from the single-stranded precursor 5′-TTTTTTTTTTGGGGTTTTGGGG. The oligonucleotides were annealed at 250 µM in 0.1 M KCl, 1.9 M NaCl, 1 × TE buffer (10 mM Tris-Cl pH 8.0, 1 mM EDTA), 16-20 hours at 60°C. The substrate was labelled and purified as described above except that

the TBE gel purification buffers were supplemented with 5 mM KCl and the substrate elution buffer with 25 mM KCl.

2.4.4 Helicase assay

The helicase activity of hPif1 proteins was assessed by preforming the assay at its optimum helicase condition determined previously (George et al 2009). The reaction buffer was 20 mM HEPES-NaOH pH 7.5, 20 mM NaCl, 5 mM $MgCl_2$, 2 mM ATP, 1 mM DTT and 0.1% (v/v) NP-40. Reactions (0.1 nM substrate) were incubated at 20°C (hPif1 helicase domain constructs) or 37°C for full-length hPif1 for 30 minutes and terminated by the addition of 0.2 volumes of 120 mM EDTA, 0.6% (w/v) SDS, 60% (v/v) glycerol and 0.1% (w/v) bromophenol blue before gel polyacrylamide gel electrophoresis (8% acrylamide (19:1), 0.25 × TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA), 0.05% (w/v) SDS gel and running buffer), and exposure of dried gels to phospho-imaging plates.

2.4.5 Strand annealing assay

Strand annealing assays employed two partially complementary radiolabeled oligonucleotides 5′- T_{55}-CGAATTCGAGCTCGGTACCC and 5′- GGGTACCGAGCTCGAATTCG-$(C)_{30}$. The annealed substrate was gel-purified and heat denatured before addition to the reaction (0.1 nM final concentration). The reaction buffer was 20 mM HEPES-NaOH pH 7.5, 135 mM NaCl, 1 mM DTT and 0.1% (v/v) NP-40. Reactions were incubated for 30 minutes at 20°C and reaction products processed exactly as described for the helicase assays.

2.4.6 DNA binding assay (gel-shift, EMSA)

Single-stranded DNA binding reactions were performed with end-labelled substrates (0.4 nM). The binding buffer was 20 mM HEPES-NaOH pH 7.5, 135 mM NaCl, 5% (v/v) glycerol, 1 mM DTT and 1 mg/ml BSA. Reactions were incubated for 20 minutes at 20°C before resolving complexes on 5% (w/v) poly-acrylamide gels (29:1) using 0.25 × TBE running buffer. G4 DNA binding reactions and EMSA (0.1 nM G4 DNA) were performed with a radiolabelled tetramolecular G4 DNA substrate (section 2.4.3) in the buffer as described above containing 5 μM T$_{(35)}$ single stranded competitor DNA, except glycerol was omitted from the reactions and replaced with 2% (w/v) PEG 8000. Reaction products were visualised and quantified following exposure of dried gels to phospho-imaging screens.

2.4.7 ATP*ase* assay

The charcoal binding assay of Iggo and Lane (Iggo and Lane, 1989) was used to measure DNA-dependent ATP hydrolysis for 100 nM Pif1-HD/ 200 nM poly-T$_{(30)}$ as described previously. The reaction buffer used was 20 mM HEPES-NaOH pH 7.5, 75 mM NaCl, 5 mM MgCl$_2$, 5 mM ATP, 0.0125 mM [γ-^{32}P]ATP (6000 Ci/mmol), 1 mM DTT, 0.1 mg/ml BSA and 0.1% (v/v) NP-40. Phosphate release was determined after 10 minutes (20°C) when ~ 3% of the ATP was hydrolysed.

2.4.8 DMS DNA footprinting

Methylation of radiolabelled DNA (Maxam and Gilbert, 1977) substrates was performed in 200 µl of 50 mM cacodylate buffer, 1 mM EDTA, pH 7.0. 10 µl of 10% DMS (in ethanol) was added, the reaction vortexed for 10-20 seconds and incubated at room temperature for 5 minutes. 50 µl of 1.5 M sodium acetate pH 7.0, 1 M 2-mercaptoethanol and 20 µl and the DNA precipitated with 1.25 ml of chilled ethanol (15 minutes at $20000 \times g$, 4^oC). The pellet was re-dissolved in 200 µl of 0.3 M sodium acetate, 1 mM EDTA, pH 8.0 and re-precipitated with 2.5 volumes of ethanol. The pellet was washed in 80% and then 100% ethanol and vacuum dried. To footprint the products of helicase reactions, 200 µl reactions in helicase buffer were treated similarly.

Piperidine cleavage was performed on the pellets dissolved in 100 µl of 1 M piperidine and incubated at 90^oC for 30 minutes. Tubes were cooled and the contents recovered to the bottom by brief centrifugation. 1.3 ml of butanol was added, the tubes vortexed for 10 seconds before spinning at maximum speed in a bench top centrifuge. The pellet was re-dissolved in 100 µl of 1% SDS, transferred to a siliconised tube, and re-precipitated with butanol. The pellet was then re-dissolved in in 200 µl of 0.5 M ammonium acetate, 1 mM EDTA, 0.5% (w/v) SDS and DNA products precipitate with 500 µL of ethanol. The pellets were washed with 80 and 100% ethanol and dried before dissolving in 50% formamide loading buffer (98% formamide, 1 mM EDTA, 0.25% (w/v) bromophenol blue). Reaction products were run on a 12% acrylamide (19:1) denaturing PAGE gel (sequencing gel, $1 \times$ TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA), 8 M urea).

2.5   Protein thermal shift assay (differential scanning fluorimetry)
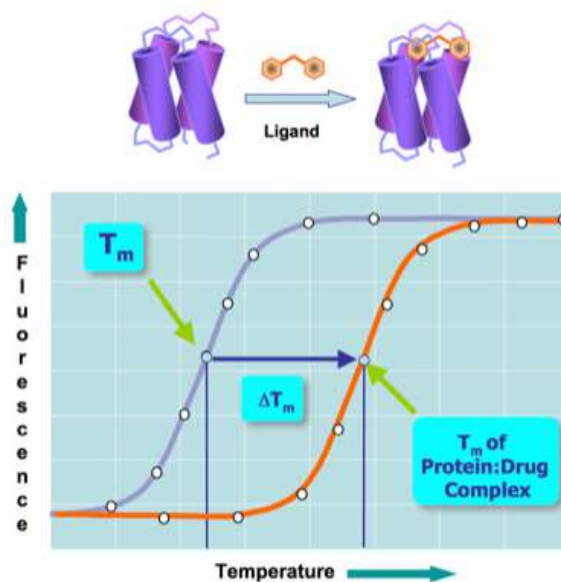
Figure 13. Principle of the protein thermal shift assay.

Reproduced from website: http://www.beta-sheet.org/page11/page17/page18/index.html

The MRCT fragment library (Medical Research Council Technology) of about 2200 fragment molecules was screened by protein thermal shift (PTS) assay for identifying physical protein-ligand interaction (Pantoliano et al., 2001). The protein thermal-shift dye kit (Applied Biosystems) was used according to manufacturer's instructions. An ABI real-time PCR machine at a ramp speed of 2°C per minute (2%) was used to generate a temperature gradient between 25°C to 99°C. Each reaction contained 3 µM protein (hPif1), 10 mM Tris-Cl pH7.5, 175 mM NaCl, 2% (v/v) glycerol, 5 mM DTT and a 1/1000 dilution of dye and 4% (v/v) DMSO. An addition of 500 µM AMP-PNP was used as a positive control (direct binder to the protein) in the

reaction. The data collected by the ABI machine was uploaded on an online software created by Paul Bond (University of York) (http://paulsbond.co.uk/jtsa/#/input) which generated the melting temperature of the samples and the ΔTm values of them in comparison to the reference sample (only protein + dye).

2.6 Testing chemical inhibitors in cell-based assays

2.6.1 Cell lines and culture

HCT116 cells (human colon cancer cell line) and the human embryonic kidney (HEK) 293 line (control cell line) were supplied from American Type Culture Collection. The cells were grown in DMEM medium containing 10% FCS in 6 well plates. After incubation for 24 hours at 37°C, the medium was replaced with DMEM that contained dialyzed FCS. The chemicals/compounds were then added to the medium. After 48 hours cells were harvested, followed by trypsinisation. The cells were centrifuged at $200 \times g$ force, washed with PBS and re-centrifuged. After removing PBS the cells were fixed in 70% ethanol and stored in -20°C prior to staining.

2.6.2 Propidium Iodide (PI) staining and fluorescence activated cell sorting (FACS)

PI is a fluorescence molecule that interacts with DNA and therefore it is used to determine cell-cycle phase by measuring the cellular DNA content. The ethanol of the fixed cells was removed by washing twice with PBS. The RNA content of the cells was degraded by adding 100 mg/ml RN*ase* A to the pellet and incubation for 20 minutes at room temperature. The cell pellets were then resuspended in PBS containing 5 mg/ml

PI (Sigma) and were stored in the dark at 4°C overnight before FACS analysis. FACS is a cell analysing technique that uses laser light to measure the amount of a specific dye such as PI in each cell. This cell-sorting technique was used to differentiate populations of cells. Further data analysis was performed using the FlowJo software (http://www. flowjo. com). The facilities were provided by department of oncology and metabolism at the University of Sheffield.

3 Results

3.1 Structural and biochemical characterization of hPif1

The majority of information about the structure and function of the Pif1 helicase has been obtained from two microorganisms: *Bacteroides sp.* (*Bs*Pif1) and *Saccharomyces cerevisiae* (*Sc*Pif1) (Zhou et al., 2016, Chen et al., 2016, Lu et al., 2018). These helicases share a level of similarity with human Pif1 helicase and have proven significantly more amenable to structural and functional analysis. However, the level of similarity, 21% identity in the case of *Bs*Pif1, prevents an accurate understanding of hPif1's mechanism of action based on these model systems.

To obtain the crystal structure of any protein the primary task is to purify it in large amounts (ideally 10 mg or above) with maximum purity. Concentrated protein will be used initially to screen for crystallization conditions appropriate for crystal formation. After identifying the initial crystallization conditions the focus would be to refine them to produce crystals with good X-ray diffraction leading to high-resolution structure determination. This study focused mainly on two hPif1 helicase domain protein constructs that were produced for both crystallography and biochemical experiments; hPif1-HD (45.5 kDa with the four residual residues from the tag) (Fig. 14) and a slightly larger protein construct of hPif1$_{206-641}$ (47.8 kDa, with the four residual residues from the tag; residue 641 is the terminal amino acid of the nuclear form). The larger protein is more soluble at low salt conditions (< 250 mM) and in the presence of DNA and most of the attempts of co-crystallization with DNA and inhibitor molecules were performed using this particular hPif1 construct. The smaller fragment, hPif1-HD, corresponds to helicase core and is analogous to the full-length *Bs*Pif1 protein. This construct was subjected to mutagenesis to produce variant forms of the protein to test models of DNA binding and unwinding. However, most of the biochemical experiment

involving uni-molecular G4 were carried out using the full-length protein (FL-hPif1, residues 1-641) exclusively.

3.1.1 Selection and construction of truncated hPif1 ORFs

Initially three constructs of the hPif1 helicase domain were designed for expression and purification for crystallography. These fragments were selected based on a computer predicted "Phyre^2" structural model (http://www.sbg.bio.ic.ac.uk/phyre2/, Appendix, Supplementary Figure 1 & 2), which was generated at the time without any structural information for related Pif1 proteins. As such, the model relied heavily on distant homologues such as *E. coli* RecD. The fragments were constructed by PCR using primers as stated in section 2.1.1. PCR products were digested and cloned into various plasmid vectors with purification tags (Table 2).

Prior to this study all hPif1 proteins in the Sanders lab (HD and full-length protein) were expressed as GST fusion proteins in BL21(DE3) cells at 22°C, yielding only moderate amounts of protein (2-3 mg from ~120 grams of cells for HD constructs, but much lower for the full-length hPif1). Expression levels and purity were significantly improved in this study using the pET15b plasmid (containing a poly histidine tag with a thrombin cleavage site for its removal or with an additional MBP fusion, HisMal) in combination with low temperature expression (the ArcticExpress™ system). The sequence of inserted DNA was confirmed by sequencing using flanking vector primers. Table 2 is a summary of protein expression time course results of all hPif1 construct used for attempts at protein purification.
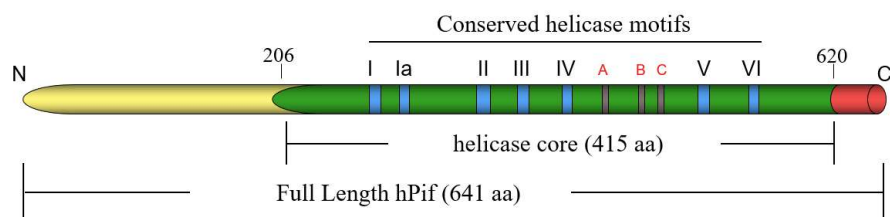
Figure 14. Human Pif1 protein construct showing the full-length and the helicase domain residues.

hPif1 construct (residues)	Expression system	Tag /cleavage site	Expression Time/ temp.	Comments
206-620	Arc	His /Thrombin	6°C /3 days	High expression and solubility
206-608	Arc	His /Thrombin	6°C /4 days	Poor purification
206-375	Arc	His /Thrombin	6°C /4 days	Poor purification
386-608	Arc	His /Thrombin	6°C /4 days	Poor purification
206-641	Arc	His /Thrombin	-	Expression very low
206-641	Arc	GST /Thrombin	6°C /4 days	Low expression. Problematic purification
206-641	Arc	His-MBP /Thrombin	6°C /2 days	High expression and solubility
171-641	Arc	His-MBP /Thrombin	6°C /2 days	High expression and solubility
171-641	Arc	GST /Thrombin	6°C /4 days	Low expression. Problematic purification
1-641	Arc	His /TEV	9°C /2-3 days	Moderate expression and solubility
1-641	Arc	GST /TEV	-	Low expression and solubility

Table 2.   List of all the hPif1 constructs prepared in this study and their protein expression time-course results.   Arc refers to as the ArcticExpress (DE3)™ system. Comments are given where appropriate.

3.1.2 Optimization of hPif1 protein expression

The ArcticExpress™ system was selected for expression of hPif1 protein due to problems encountered in expression and purification, including low level expression, insolubility and difficulties in separating the protein from the endogenous *E. coli* chaperon (GroEL/ES when using mesophilic BL21(DE3) cells). ArcticExpress™ (DE3) bacteria are able to grow and express protein at low temperatures due to the presence of the GroEL/ES related chaperone (Cpn10/60) from the psychrophilic bacteria *Oleispira antarctica*. Cpn10/60 can solve the insolubility problem and can be easier to separate from the target protein than its homologue GroEL/ES in the BL21(DE3) expression system. A time-course was initially performed with Pif1$_{206-620}$ at 6°C, 8°C and 10°C. A sample was collected at each of five time-points 6h, 24h, 36h, 48h and 72h. The result (not shown) indicated that the optimum temperature for soluble protein expression was at 6°C for a minimum of 48 hours incubation time. Similar trials were performed with other constructs (Table 2).

3.1.3 Purification of hPif1-HD (residues 206-620)

After optimization of protein expression and the various purification steps (not shown) the following strategy (below) was adopted for purification of hPif1-HD. The optimised chromatography of the various purification steps is shown in Figure 15 & 16. The procedure yielded approximately 0.2 mg per gram wet weight of cells.

Bacteria lysis, extract clarification and nucleic acid removal with Polymin P

⇩

Protein precipitation with 40 % (saturation) ammonium sulphate

⇩

First His-trap column (IMAC) and development of a 20-200 mM imidazole

gradient (Fig. 15)

⇩

Dialysis and removal of the His-tag from the protein using thrombin

⇩

Re-application of the protein to a His-trap column and gradient elution from 2-80

mM imidazole (Fig. 16A)

⇩

Size fractionation on a SEC column (Superdex 75 prep grade) (Fig. 16B)

⇩

Final polishing on a high-resolution Source S ion exchange column (Fig. 16C)

A

B

Figure 15. SDS-PAGE analysis of hPif1-HD extraction and primary IMAC column run. **(A)** UV absorption (mAU) trace. Peak elution of hPif1-HD was ~120 mM imidazole. **(B)** Purification controls and analysis of fractions corresponding to UV Abs_{280} peaks in the chromatogram shown in (A). Lane 1, un-induced (no IPTG) total protein extract sample. Lane 2, total cell lysate sample after induction. Lane 3, cleared lysate (total soluble protein extract). Lane 4, post Polymin P protein extract. Lane 5, sample applied to the column/re-dissolved ammonium sulphate precipitate (40% cut). Lane 6, flow-through sample of the column. Lane 7, low imidazole wash sample after loading the column. Subsequent lanes show the protein peak from the column.
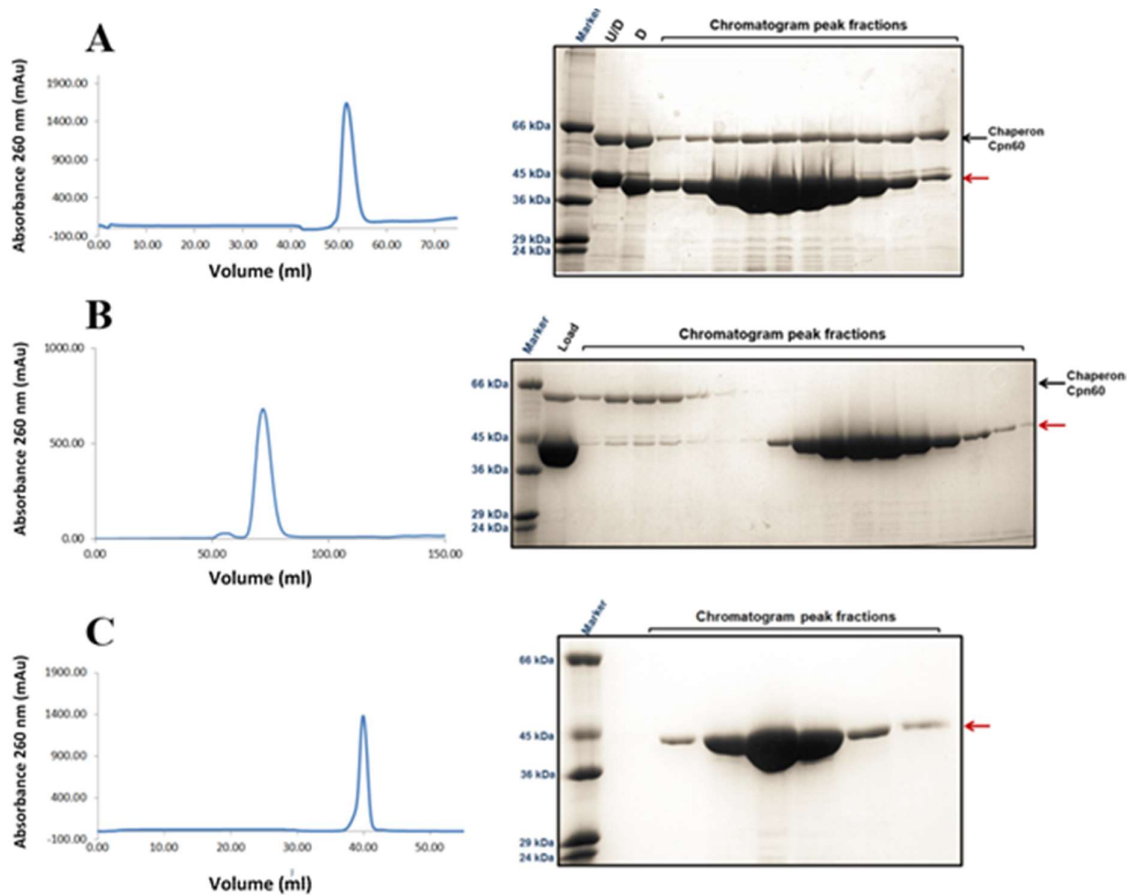
Figure 16. Chromatograms and SDS-PAGE analysis of hPif1-HD purification steps. (A) Second IMAC column, after digesting the histidine tag off the protein. Undigested (U/D) control sample and applied sample to the column -after digestion (D), and the fractions corresponding to the chromatogram peak are analysed on the SDS PAGE. **(B)** Gel-filtration chromatogram showing two distinctive peak of chaperon (and other contaminants) and hPif1-HD. "Load" refers to the pool of selected fractions from the previous IMAC column. **(C)** Source S column showing the final enhancement of protein purity.

3.1.4 Purification of hPif1$_{206-641}$

For unknown reasons hPif1$_{206-641}$ did not express effectively as a simple His-tagged fusion protein. A His-MBP tag was effective at generating soluble protein after expression at 6℃ and 2 days incubation, as stated in Table 2. From a starting mass of bacteria of about 200 g of cell approximately 30 mg of protein was obtained; the purification procedure is as summarized below:

Lysis, clarification of the extract and nucleic acid removal with Polymin P

⇩

50% (saturation) ammonium sulphate precipitation

⇩

Application to the first His-trap column (IMAC) and imidazole gradient elution (20-200 mM) (Fig. 17A)

⇩

Dialysing and digestion of the His-MBP tag from the protein using thrombin

⇩

Concentration of the protein using a Vivaspin 20 device (5000 kDa cut off)

⇩

Size fractionation on a SEC column to remove the chaperone (Superdex S200 prep grade) (Fig. 17B)

⇩

His-MBP tag removal by re-applying the protein to the His-trap column (Fig 17C)

⇩

Final polishing on a high-resolution Source S column ion exchange (Fig. 17D)

Figure 17. Purification of hPif1$_{206-641}$. (A) First IMAC column chromatogram showing the UV trace of the eluted protein (blue), the imidazole concentration gradient graph (red) **(B)** Gel-filtration column (Superdex S200, prep grade) showing two distinctive peaks. The second peak contains hPif1$_{206-641}$ and the first mainly chaperone. **(C)** Fractions collected from the second peak of the gel-filtration column were digested with thrombin and loaded onto an IMAC column to remove the His-MBP tag (second peak). **(D)** Chromatogram of the final Source S (conductivity graph shown in red) column and gel showing the purity of the protein.

3.1.5 Determination of the hPif1 helicase domain structure

Three principal structures (Table 3) were solved for the human Pif1 helicase domain as listed below (the Br structure was produced to solve the phase problem by Dr. V. Levdikov, University of York). The nucleotide and DNA free form is designated as apo and was crystallised with four sulphate ions bound. All crystallographic details are listed in Appendix, Supplementary Table 1.1.

Protein construct	Structure with	Resolution	PDB code
hPif1$_{206-620}$	AMP-PNP	1.13 Å	6HPH
hPif1$_{206-620}$	AMP-PNP (Br)	1.43 Å	6HPQ
hPif1$_{206-641}$	apo (SO$_4^-$)	1.44 Å	6HPT
hPif1$_{206-641}$	ADP•AlF$_4^-$	3.96 Å	6HPU

Table 3. List of principal structures determined for the hPif1 helicase domain.

High resolution X-ray structures were determined for the helicase domain of hPIF1 corresponding to an idle cycle of ATP hydrolysis: structures of complexes with the ground state analogue AMP-PNP in two different crystal forms (1.13Å, 1.43Å), an apo structure complexed with three sulphate moieties determined at 1.44Å resolution and also a low-resolution structure of a complex with the transition state analogue ADP•AlF4-, (see Supplementary Table 1 for details of the data collection and refinement statistics). While the hPIF1-ADP•AlF4- structure (Rfactor 17.9%; Rfree=25.3%) mirrors a previously published low resolution (3.6Å) crystal structure of the hPIF1 helicase domain (residues 200-641, pdb code: 5FHH, Rfactor 31.3%; Rfree=35.5%, ref. 26), it is based on using the high-resolution structure of the hPIF1-AMP-PNP complex during molecular replacement, resulting in a more accurate model

than the previously available structure (RMSCA=2.0Å; RMSoverall=2.5Å). Notably, average temperature factors indicate significant flexibility of the hPIF1-ADP•AlF4- structure relative to all others. Analysis shows (V. Levdikov, UoYork) that the events at the ATP binding site and in particular the ATP γ-phosphate are relayed to positional and conformational adjustments in defined protein segments. Such structural changes result in an increased flexibility of the hPIF1-ADP•AlF4- protein structure that could affect protein-DNA interactions.

Figure 18 shows an image of the crystals obtained for the apo (sulphate bound) form and the electron density map of the high-resolution 1.1Å structure of hPif1-HD bound to the ATP analogue (AMP-PNP) which mimic the ground state of the ATP hydrolysis cycle. The ATP binding site and the interactions/bonds between the AMP-PNP molecule and magnesium ion and the surrounding residues. In the ATP binding site, tryptophan 380 packs against the adenine ring, while the side-chain of Arg 381 contacts the other end of the ATP analogue, interacting with the γ-phosphate and being within 3.3Å of the nitrogen between the β and γ-phosphates. The Arg 381 residue may function as an arginine finger. The side chain of residue Gln 211 (not shown) also hydrogen bonds with the N6 (2.9Å) and N7 (3.0Å) atoms of the adenine ring and may therefore function as a canonical Q motif, also stabilising adenine nucleotide binding, that is conserved in SF1 RNA/ DNA helicases (Fairman-Williams et al., 2010). Four residues from the expression tag, GSRM, are present at the start of the hPif1 constructs. The Met-205 residue is present at an analogous position in *Bs*Pif1. Met 205 is within 3.5Å of adenine and it is possible these four residues stabilize the protein-nucleotide complex, enabling high-resolution structures determination for human Pif1 in these cases.
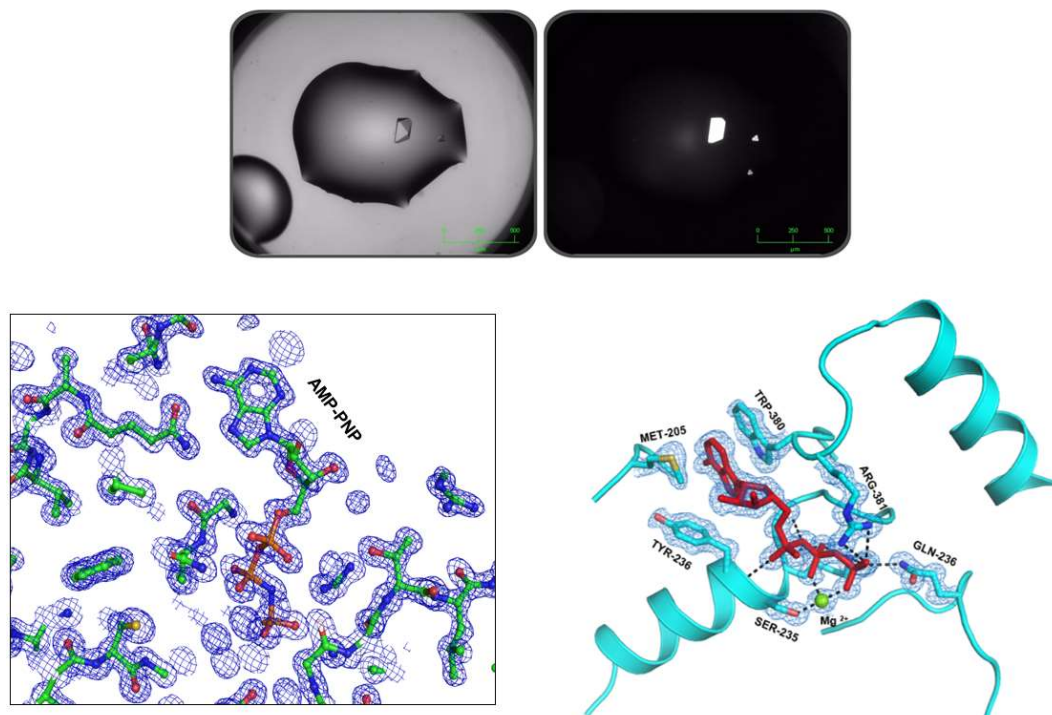
Figure 18. hPif1 helicase domain crystallisation. Protein crystals of human Pif1$_{206-641}$ without and with the UV trace (**top**). Modelled structure in the 1.13Å electron density map of the hPif1$_{206-620}$ bound to AMP-PNP. The structure was extracted from Coot software at $^\delta$1. The ATP binding site of hPif1 showing the AMP-PNP molecule and the magnesium ions' interactions/bonds with surrounding residues covered by the electron density of the site at $^\delta$1. AMP-PNP molecule in red and magnesium shown in lemon green (**bottom**).

Figure 19 shows the overall structure of the hPif1 helicase domain modelled using the 1.4Å apo structure. The domains are annotated and coloured according to the domain architecture characteristic of the RecD2 family of prototypic SF1 helicases. hPif1-HD contains multiple structural domains: two RecA-like domains 1A and 2A separated by the ATP binding cleft and a SH3 (SRC Homology 3) 2B domain. A small C-terminal domain (CTD), protruding from domain 2A contains an α-helix which packs against and is structurally part of domain 1A.
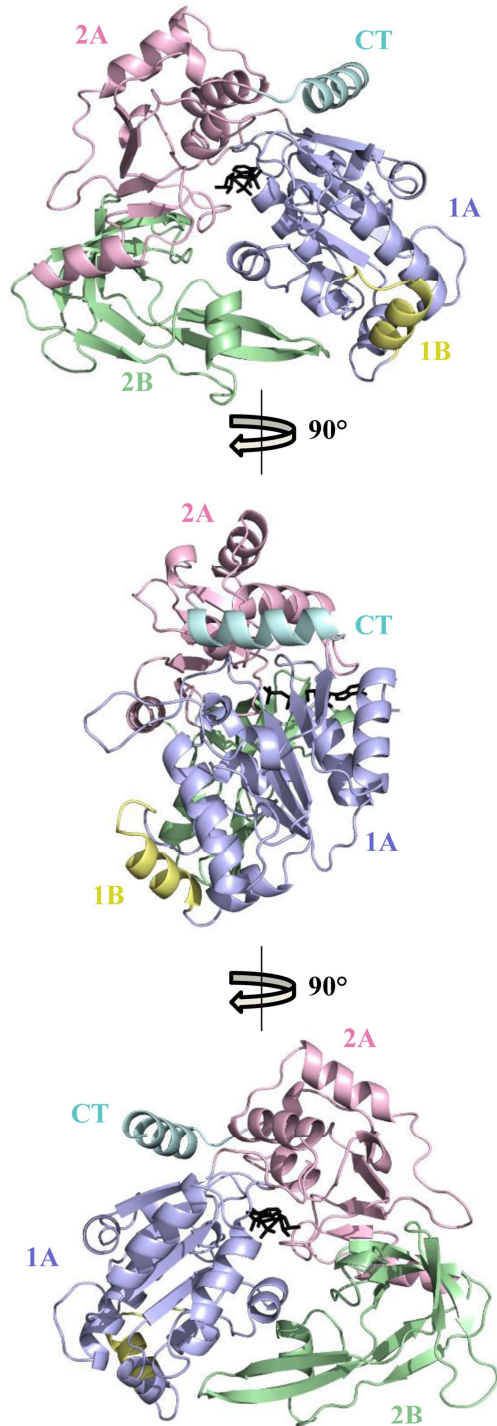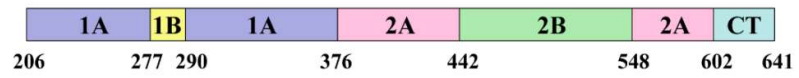
Figure 19. The hPif1 helicase domain structure residues 206-620. Map of the hPif1 helicase domain showing the residue boundaries of each domain (**top**). The overall structure of human Pif1$_{206-641}$ created using PyMol software showing three different angular views of the protein. The AMP-PNP molecule is shown in black (**bottom**).

The comparison of the structure of AMP-PNP bound and the apo form shows a significant movement of the residues around the ATP binding site induced by nucleotide binding. However, the overall conformation of the main chain of the protein shows no significant movement between apo and AMP-PNP bound (Fig. 20A). Also, no electron density for the C-terminal 21 residues of construct hPif1$_{206-641}$ (apo) are apparent. The comparison of the AMP-PNP structure to the ADP•AlF$_4^-$ structure which mimics the transition state of the ATP hydrolysis cycle (just after ATP hydrolysis) shows movements in some sites of the protein (significantly in domain 2B, 7.3Å dislocation) (Fig. 20B). These two sites are close to the ssDNA binding site (modelled based on the conserved DNA binding residues of Pif1 helicase, see below).
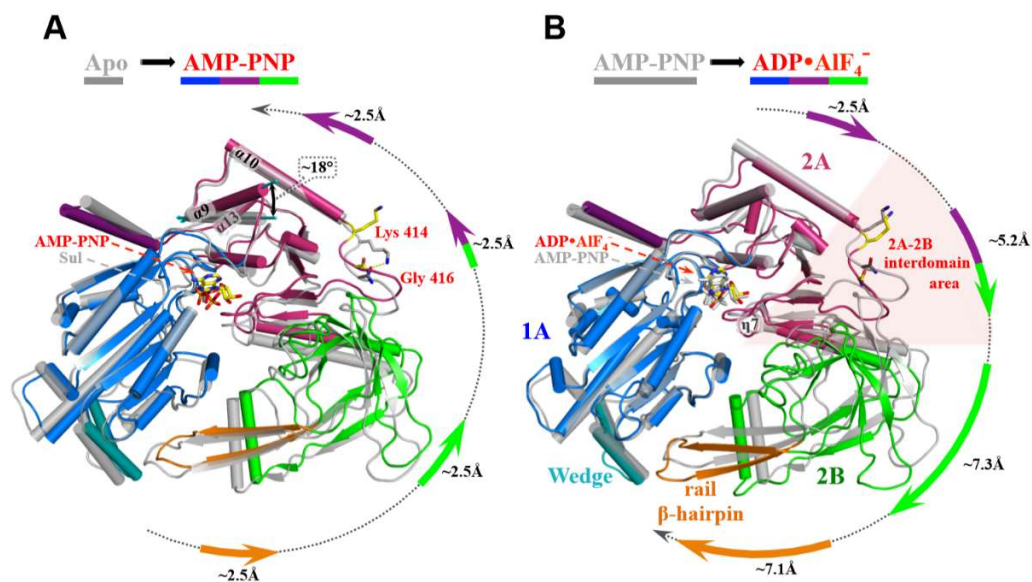
Figure 20. Overall structural rearrangements upon DNA binding and hydrolysis. Cartoon representation of the superposition on 1A domain of **(A)** Apo (gray) and AMP-PNP (domains colored) structures, **(B)** AMP-PNP (gray) and ADP•AlF$_4^-$ (domain colored) structures). Ligand positions and secondary structure elements α9, α10, α13 and η7 are shown. Lys414 and Gly416 carboxyl group interactions are represented by stick model. Shift and direction are represented by arrows in domains colors. **(A)** Angular displacement of α9 is indicated. **(B)** 2A-2B interdomain area depicted by pink shadow. Figure created by Dr. Vladimir Levdikov (University of York).

3.1.6 Functional analysis of variant hPif1 forms

After multiple attempts to co-crystalize the protein with ssDNA oligonucleotides (and other substrates, see Discussion) failed, the ssDNA bound conformation of hPif1 was modelled. The proteins' DNA binding sites were then analysed and variants created based on conserved residues in or at the entrance to the ssDNA binding channel (the "separation wedge") with reference to the available crystal structures of *Bs*Pif1-ssDNA (Table 4; note the lack of sequence conservation in the separation wedge).

Modelling of the hPif1-ADP•AlF$_4^-$-ssDNA was performed by Dr. Ben Bax (University of York/University of Cardiff). An initial model of a ssDNA bound conformation of human Pif1-HD was made by moving domains from the 1.13Å human complex with AMP-PNP onto the equivalent domains in the 2.0Å *Bacteroides* Pif1 complex with ssDNA (pdb code: 5FHD), and using the ssDNA from 5FHD (Figure 21 and Appendix, Supplementary Figure 5). The human Pif1-HD-ssDNA model was then energy minimized in Maestro (Schrodinger Release February 2017: Maestro, Schrodinger, LLC, New York, NY). A second model of the ssDNA bound conformation of human Pif1-HD was made when coordinates for the more closely related *Sc*Pif1 became available (the 2.03Å structure of *Sc*Pif1 in complex with GGGTTT and ADP•AlF$_4^-$ (Lu et al., 2018) PDB code 5O6B). The two models are very similar although there is a small shift in the DNA position and in the position of domains 2B (the SH3 domain – data not shown).

3.1.6.1 Construction of hPif1HD mutants and variant proteins

Figure 21 shows the hPif1-ADP•AlF$_4^-$-ssDNA complex model and the location of amino acids targeted for substitution (Table 4). Mutants of hPif1-HD were created by the overlapping PCR technique as described in section 2.1.9 and purified from 18 litres of culture following the hPif1-HD purification steps described in section 2.2.5. The purified proteins are also shown in Figure 21.

	hPif1 mutation	Domain	Motif/ functional site	*Bs*Pif1	*Sc*Pif1
1	R290A	1B	Separation wedge	S89	S325
2	P291A	1B	Separation wedge	E90	R326
3	R294A	1B	Separation wedge	V93	L329
4	H269A	1A	ssDNA binding, PO$_4$ interaction	H68	H393
5	E307Q	2B	Walker B	E106	E342
6	N495A	2B	ssDNA binding, base stacking, C motif	N296	N533
7	H555A	2A	ssDNA binding, ribose interacting	H361	H705
8	K556A	2A	ssDNA binding, PO$_4$ interaction	K362	K706
9	F573A	2A	ssDNA binding, base stacking	F379	F723
10	K485E	2B	B motif	V287	K525
11	N486A	2B	B motif	N288	N526
12	K414A	2A	SO$_4$ binding (apo hPif1 structure)	K221	P450

Table 4.   Residues mutated in hPif1 for biochemical analysis of purified hPif1-HD proteins.   The region or functional site within which the substitutions were made is indicated. Corresponding residues in *Sc*Pif1 and *Bs*Pif1 are given.
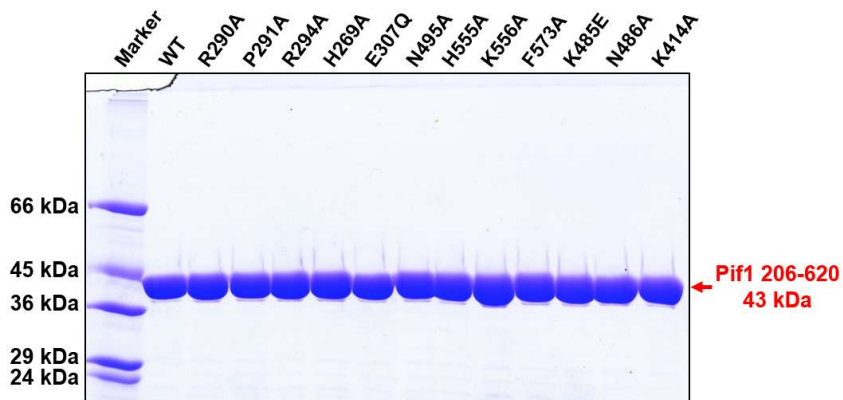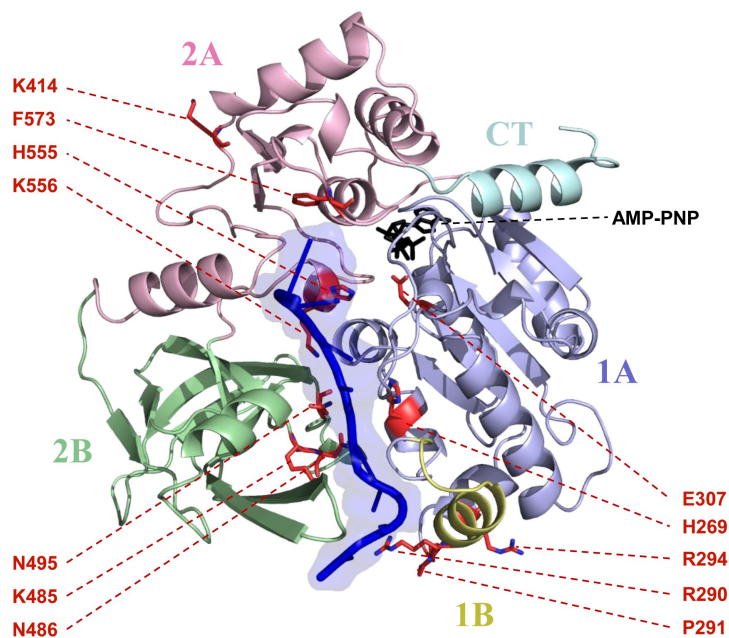
Figure 21. Modelled ssDNA in the hPif1 helicase domain structure showing the residues targeted for substitution (red) by site directed mutagenesis (**top**). SDS-PAGE analysis showing the purity of each of the hPif1-HD mutants, 20 µg of each (**bottom**).

3.1.6.2 Single amino acid substitutions target hPif1 ssDNA binding

Residues selected for functional analysis in human Pif1 were predicted to interact with DNA. The focus is mainly on the separation wedge (domain 2B) at the entrance to the ssDNA binding channel and the conserved residues within that region (Fig. 21 and Table 4). K485 and N486 of the previously uncharacterised B motif within domain 2B, one of three conserved segments (A, B and C) located between helicase SF1 motif IV and V and characteristic of the Pif1 family helicases were also targeted (see the alignment in Appendix, Supplementary Figure 4). These residues are in an exposed loop at the end of the B motif segment. In bacterial Pif1, domain 2B makes large movements in response to substrate binding, positioning the residues at the tip of the B motif in the ssDNA binding channel. Residue K414 was also selected for analysis. K414 coordinates a sulphate ion in the apo hPif1-HD structure and is conserved between mammalian and bacterial Pif1 (and not yeast).

In the absence of nucleotide cofactors hPif1-HD forms a discrete ssDNA complex in gel-shift assays (EMSA) with polydeoxynucleotide substrates greater than 30 bases (George et al., 2009). As shown in Figure 22, ssDNA binding of wild-type and variant hPif1-HD forms was probed using a radiolabelled poly $T_{(35)}$ substrate. In the separation wedge, binding of variant R290A was reduced nearly 3-fold while variants P291A and R294 bound at greater than ~80% of wild-type (data for 25 nM Pif1-HD, Fig. 22B). Variants E307Q (Walker B motif) and K414A (SO_4^{2-} binding in the apo hPif1-HD structure) also bound ssDNA at close to wild-type levels. All amino acid substitutions in the putative ssDNA binding channel and the B motif (residues K485 & N486) result in ~8-10 fold decrease in ssDNA binding, with variant K556A showing the greatest defect.
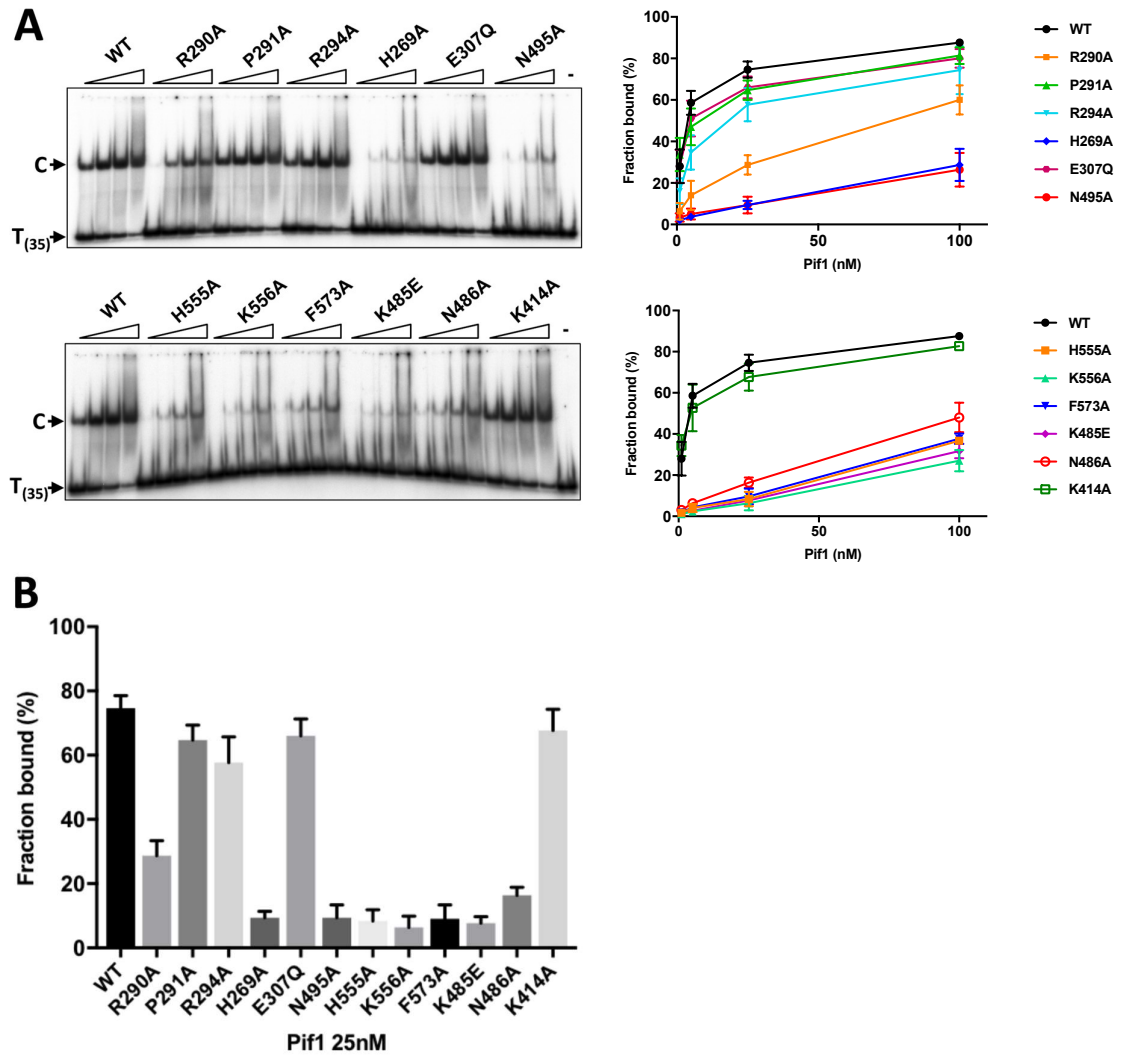
Figure 22. Binding of hPif1-HD and variant proteins to oligo dT$_{(35)}$. **(A)** Products of ssDNA binding reactions (0.4 nM substrate, protein titration series 1, 5, 25 and 100 nM). The primary protein-DNA complex 'C' is indicated as is the mobility of free DNA. The final lane in each panel shows the mobility of free substrate (no hPif1 protein). **(B)** Binding data derived from the 25 nM hPif1-HD concentration are shown in the bar graph. N=4 experimental repeats, mean and standard deviation defined by the error bars.

3.1.6.3 Enzymatic activity of the hPif1 variants

Wild-type and variant hPif1-HD proteins were tested in an ATP-dependent helicase assays for their ability to displace a 20 base radiolabelled oligonucleotide from a partially single- and double-stranded test substrate. Reaction products were separated by polyacrylamide gel electrophoresis (PAGE) and quantified by phosphorimaging. The protein concentration range chosen was at the upper-end of the range where maximum unwinding is observed for wild-type protein (Fig. 23A). The magnitude of the defects observed in the helicase activity (Fig. 23 A & B) of hPif1-HD variants R290A (25% of wild-type), P291A and R294A (> 80% of wild-type) directly paralleled the magnitude of the defects observed in ssDNA binding (Fig. 22). Substitutions E307Q (Walker B motif,), N486A (B motif) and all those in the putative ssDNA binding channel resulted in near complete abolition of unwinding activity. Variants K485E (B motif) and K414A retained unwinding activity at ~ 10 and ~ 60% wild-type levels respectively.

hPif1 is a DNA dependent ATP*ase*. ATP hydrolysis was determined for wild-type and variant forms under conditions where the majority of the substrate remains un-hydrolysed (Fig. 23C). Again, the relative extents of ATP hydrolysis measured for mutants R290A, P291A and R294A compared to wild-type reflected the defect observed in ssDNA binding and helicase activity described above. All amino acid substitutions in the ssDNA binding channel and B motif resulted in variant hPif1 forms with significantly impaired ATP hydrolysis. Again, there was a strong correlation between loss of ATP*ase* activity and the magnitude of the reduction in apparent ssDNA binding affinity, where variant K556A showed little or no ATP*ase* activity and N486A ~ 30% compared to wild-type.
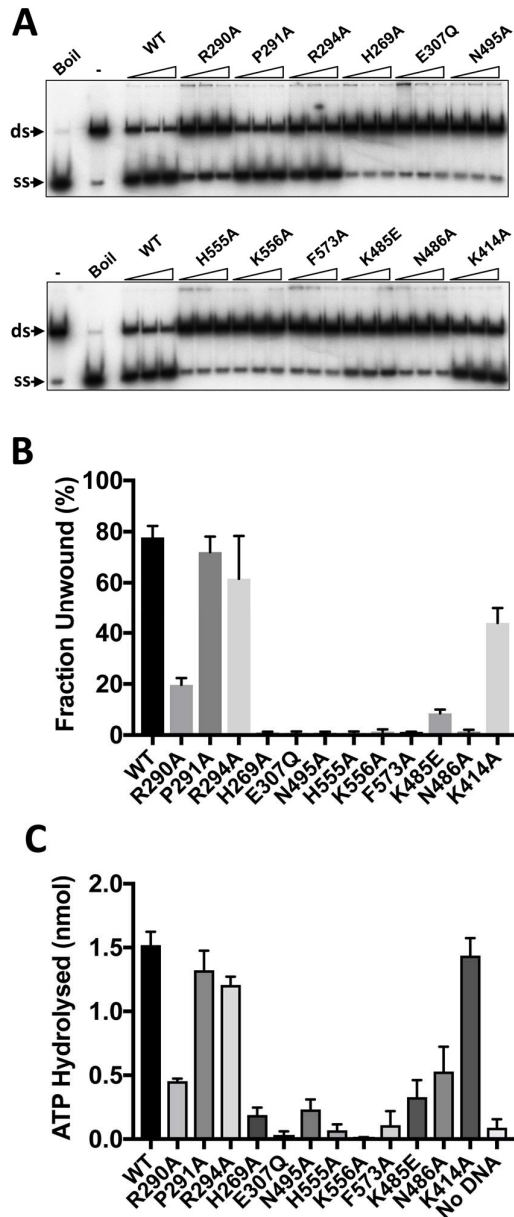
Figure 23. Enzymatic activities of hPif1-HD and variant protein forms. **(A)** Displacement of a ^{32}P radiolabeled 20 base ssDNA strand (ss) from a partially single- and double-stranded test substrate (ds). Reaction products (0.1 nM substrate, 3.75, 7.5 and 15 nM hPif1-HD) were resolved by PAGE for quantification following phosphorimaging. On the left of each panel the mobility of native substrate (ds, reaction with no hPif1-HD protein) and the labelled ssDNA product generated by heat denaturation (Boil/ (ss) product) are indicated. **(B)** Fraction unwound for reactions with 7.5 nM hPif1-HD. N=4 experimental repeats, mean and standard deviation delimited by the error bars are shown **(C)** DNA dependent ATP*ase* was determined at 200 nM Pif1-HD/T$_{(30)}$ ssDNA after 10 minutes' incubation ($\sim$ 3% of substrate hydrolyzed for wild-type hPif1-HD). N=3 experimental repeats, mean and standard deviation delimited by the error bars are shown.

3.1.6.4 ssDNA strand annealing activity of hPif1-HD proteins

The hPif1 helicase domain catalyses the re-annealing of complementary ssDNA strands in the absence of ATP and Mg^{2+} (George et al., 2009). To assess the ssDNA strand annealing activity of wild-type and variant hPif1-HD proteins a fork-like duplex substrate (20 base pairs) with ssDNA tails (5′ T55 and 3′ C30) and both strands end-labelled with ^{32}P was produced. The duplex was heat denatured before addition to the annealing reaction and reaction products were separated by polyacrylamide gel electrophoresis for visualisation and quantification. As shown in Figure 24, with the exception of mutants N495A and F573A all mutants catalysed strand re-annealing at or significantly above wild-type levels.
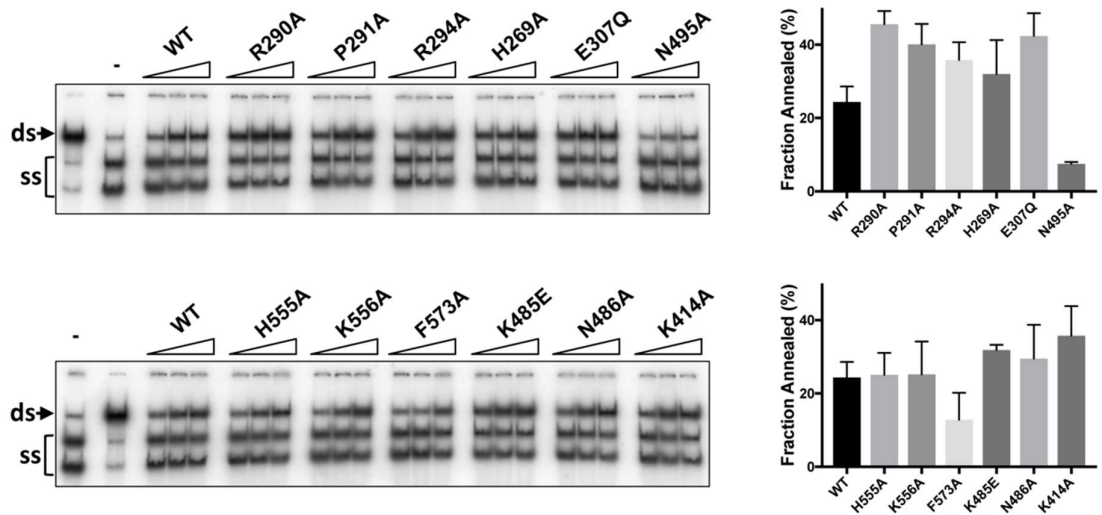
Figure 24. Strand annealing catalysed by hPif1-HD and variant proteins. Reactions (50, 100, 150 nM hPif1-HD, 0.1 nM substrate) employed two oligonucleotides with 20 bases of complementary DNA. The electrophoretic mobility of the partially single- and double-stranded product (native/ds) and the heat denatured substrate (- no hPif1-HD reaction) are indicate in the lanes to the left of each gel image. Quantified data (150 nM hPif1-HD), N=4 independent repeats, mean and standard deviation delimited by the error bars, are shown to the right of each panel.

3.1.6.5 Binding of hPif1-HD proteins to a synthetic tetramolecular G4 DNA substrate

A G4 DNA binding activity of hPif1 resides in helicase core (Sanders, 2010). This activity can be readily visualised by EMSA using a radiolabelled substrate. In the absence of nucleotide cofactors hPif1-HD demonstrates a substantially higher affinity for a synthetic tetramolecular G4 DNA substrate with 5′ ssDNA tails compared to its single-stranded DNA precursor (top panel of Fig. 25). The pattern of G4 DNA binding observed for wild-type shows a single dominant complex at low protein concentrations and then the formation of higher-order species at higher protein concentrations. To quantify G4 DNA binding extents all complexes were included in the bound fraction.

In significant contrast to the results observed for ssDNA binding (Fig. 22), all the variant hPif1-HD proteins tested bound the synthetic tetramolecular G4 DNA substrate at levels of 75% of wild-type or greater (Fig. 25B, data for the lowest concentration of protein tested, 1 nM hPif1-HD). Curiously, variant K414A consistently displayed an increase in G4 DNA binding activity relative to wild-type at the lower protein concentrations tested.
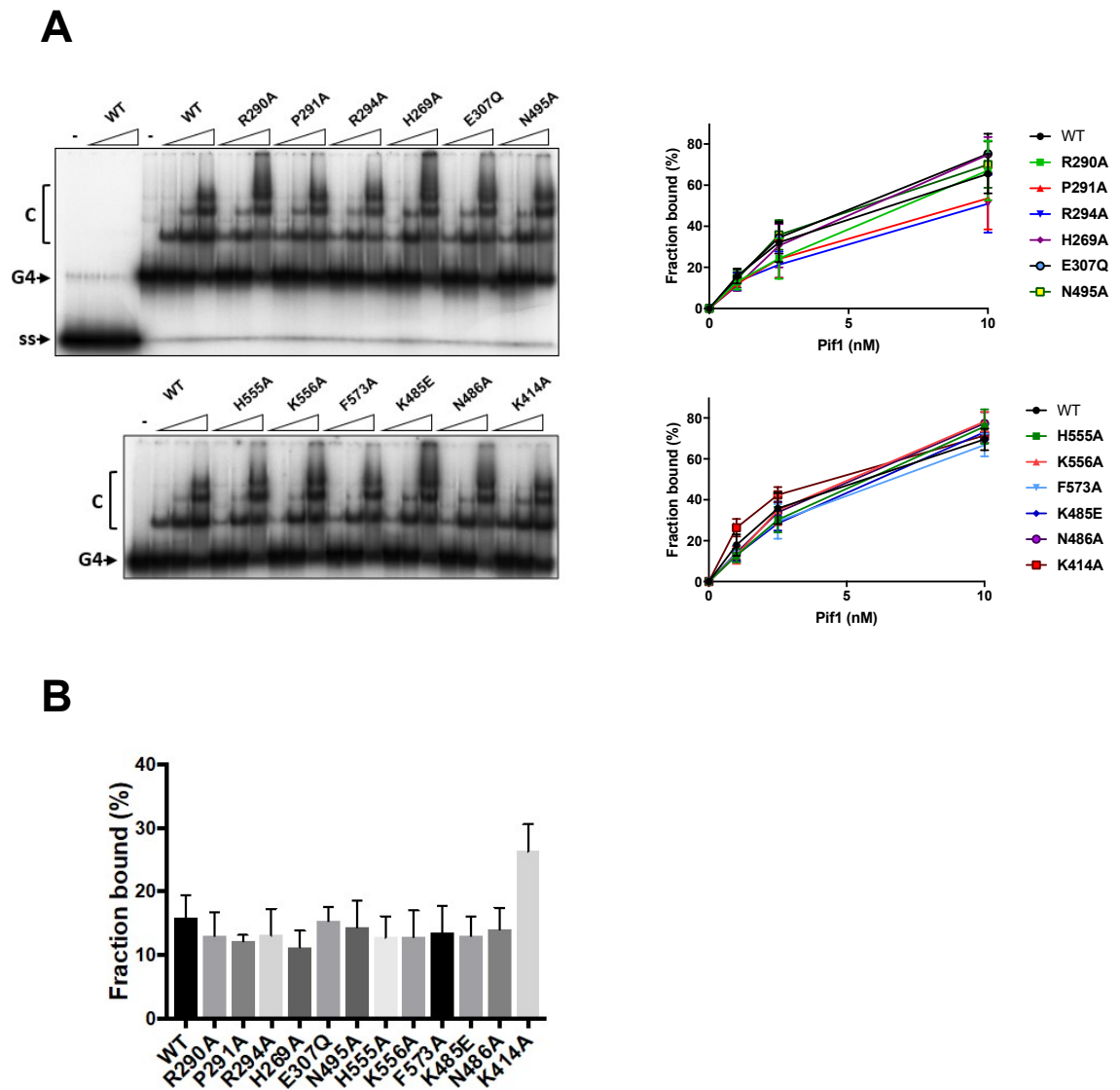
Figure 25. Tetramolecular G4 DNA binding by hPif1-HD and variant proteins. (A) Products of G4 DNA binding reactions performed in the presence of 5 µM unlabeled $dT_{(35)}$ competitor oligonucleotide (0.1 nM G4 DNA substrate; 1, 2.5, 10 nM hPif1HD) were resolved by PAGE and quantified following phosphorimaging. The first four lanes of the top panel demonstrate minimal binding of wild-type hPif1HD to the single-stranded precursor of the tetramolecular G4 substrate (ss). The graphs of quantified data for all titration points (left) (N=3). **(B)** Quantification of the binding data for 1 nM hPif1-HD, N=3 experimental repeats, mean and standard deviation delimited by the error bars.

3.2 DNA binding properties of hPif1

To understand the mechanism of action of a helicase it is necessary to understand its DNA binding and strand separation activity and how they are influenced by nucleotide cofactors. An investigation of the ssDNA binding activity was primarily done to support efforts to find a suitable substrate and conditions for hPif1-HD-ssDNA co-crystallisation. These experiments employed the hPif1$_{206-641}$ construct that had shown superior solubility and stability at high concentration and low salt conditions and hence was the construct of choice for protein-DNA co-crystallisation studies. EMSA experiments were done using ^{32}P-end labelled homo-polydeoxynucleotide ssDNA sequences dA$_{(35)}$, dT$_{(35)}$, and dC$_{(35)}$. Poly dG$_{(35)}$ could not be used in this experiment due to its propensity to form G-quadruplex secondary structures that would likely be heterogenous within long polymers. As the results show in Figure 26, hPif1-HD forms predominantly a single species with ssDNA in the absence of any nucleotide co-factor (5 mM Mg^{2+} is present in all the experiments), although there are indications of higher-order complexes and complex instability. hPif1-HD demonstrated the highest affinity for dC$_{(35)}$ and then dT$_{(35)}$ but no binding to poly dA$_{(35)}$ ssDNA.
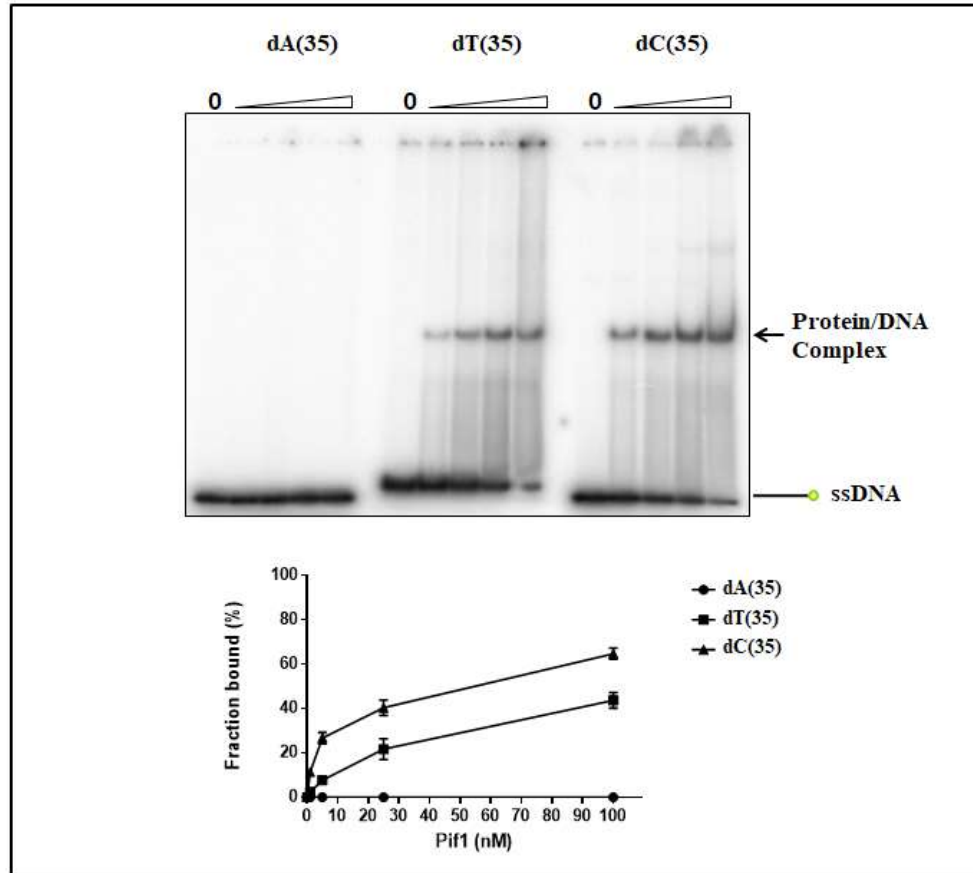
Figure 26. **hPif1$_{206-641}$ binding to DNA homopolymers.** The binding affinity of different homo-polynucleotides (0.1 nM substrate) was observed at increasing concentration of protein (1, 5, 25 and 100 nM hPif1$_{206-641}$). Lanes marked as 0 show the mobility of the substrate in the absence of protein. The graph shows the quantified results of the assays (N=3). Error bars indicate standard deviation.

In the presence of the nucleotide co-factor ADP•AlF$_4^-$, which mimics the transition state of the protein-nucleotide complex just after ATP hydrolysis, a substantial increase in binding affinity was observed (Fig. 27 and Fig. 28) and three discrete complexes were apparent, as shown with the poly dC$_{(35)}$ substrate in Figure 27. In addition, at the lower protein concentration (1-5 nM) the relative difference in affinity between the poly dC$_{(35)}$ and poly dT$_{(35)}$ appeared greater compared to the absence of cofactor (~3 fold, no cofactor, Fig. 26 and ~15 fold with ADP•AlF$_4^-$, Fig. 27).
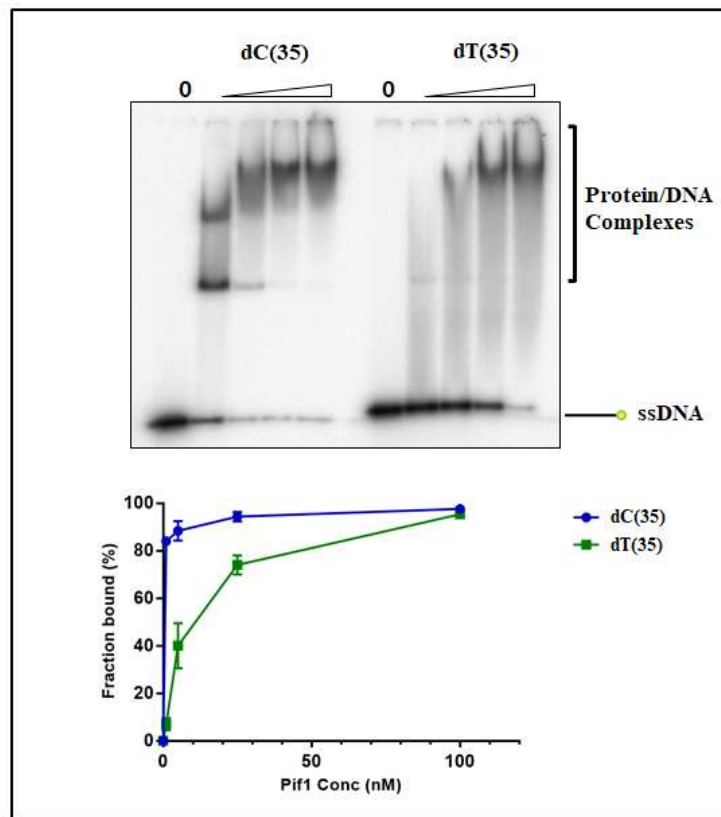
Figure 27. Comparing homo-polynucleotides dC$_{(35)}$ and dT$_{(35)}$ binding affinities in the presence of the co-factor ADP•AlF$_4^-$. 1, 5, 25 and 100 nM hPif1$_{206-641}$ and 0.4 nM ^{32}P-labelled ssDNA substrates. Lanes marked as 0 are showing the substrate in the absence of protein. Graph showing the quantified results of the assay (N=3). Error bars indicate standard deviation.

To further investigate the dependence of affinity on nucleotide co-factor type, binding was re-tested in the presence of co-factors ADP +Mg^{2+} (product) and AMP-PNP +Mg^{2+} (ground state) using the poly $dC_{(35)}$ substrate (Fig. 28). hPif1$_{206-641}$ clear shows the highest affinity for poly $dC_{(35)}$ in the presence of ADP•AlF$_4$, while ADP and AMP-PNP result in reduced affinity relative to the absence of nucleotide co-factor.
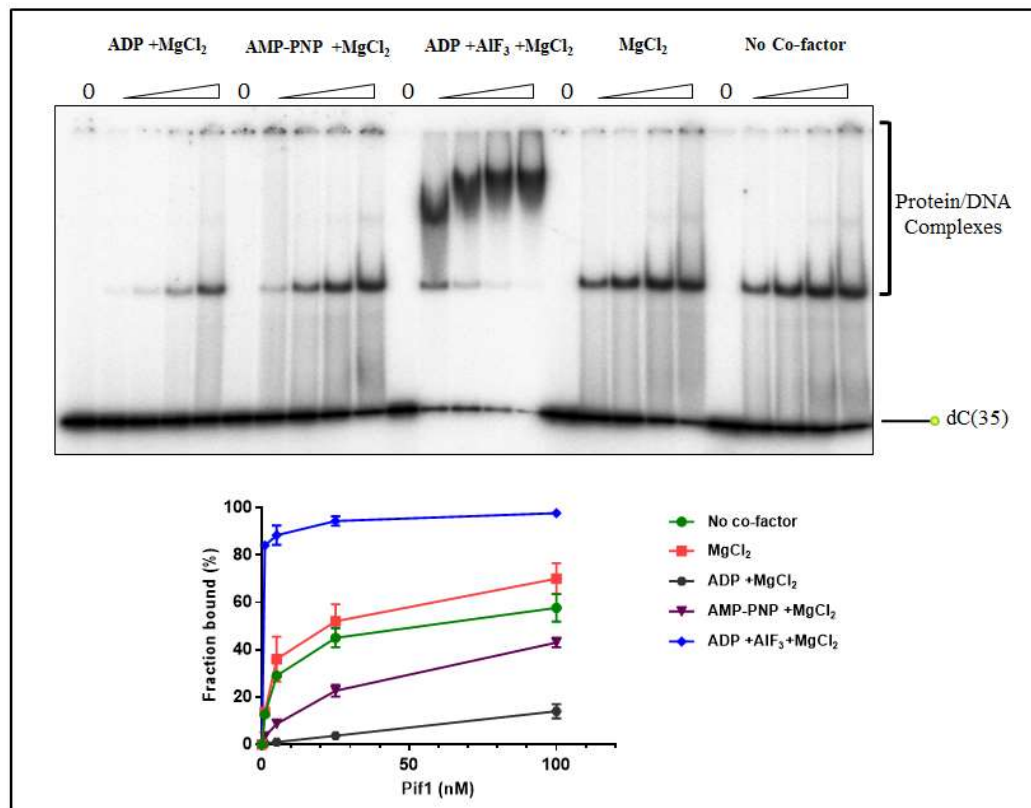
Figure 28. Comparing hPif1$_{206-641}$ binding affinities to ^{32}P end-labelled dC$_{(35)}$ in the presence of different co-factors. 1, 5, 25 and 100 nM hPif1$_{206-641}$ and 0.4 nM ^{32}P-labelled ssDNA substrates. Lanes marked as 0 are showing the ssDNA substrate in the absence of protein. Graph showing the quantified results of the assay (N=4). Error bars indicate standard deviation.

The appearance of multiple discrete complexes with 35 bases substrates in the presence of ADP•AlF$_4^-$ suggested multiple binding events. The ssDNA length dependence of hPif1$_{206-641}$-ssDNA binding is shown by gel-shift assay in Figure 29, which demonstrates increasing binding events as the length of the ssDNA increases. One discrete complex is observed with dC$_{(10)}$, a first and then a second with dC$_{(20)}$ and increasing protein concentration, and an indication of three shifted species with dC$_{(30)}$. Together with the data shown in Figure 27, these data suggest a nominal high-affinity ssDNA binding site size of ~12 bases in the presence of ADP•AlF$_4^-$.
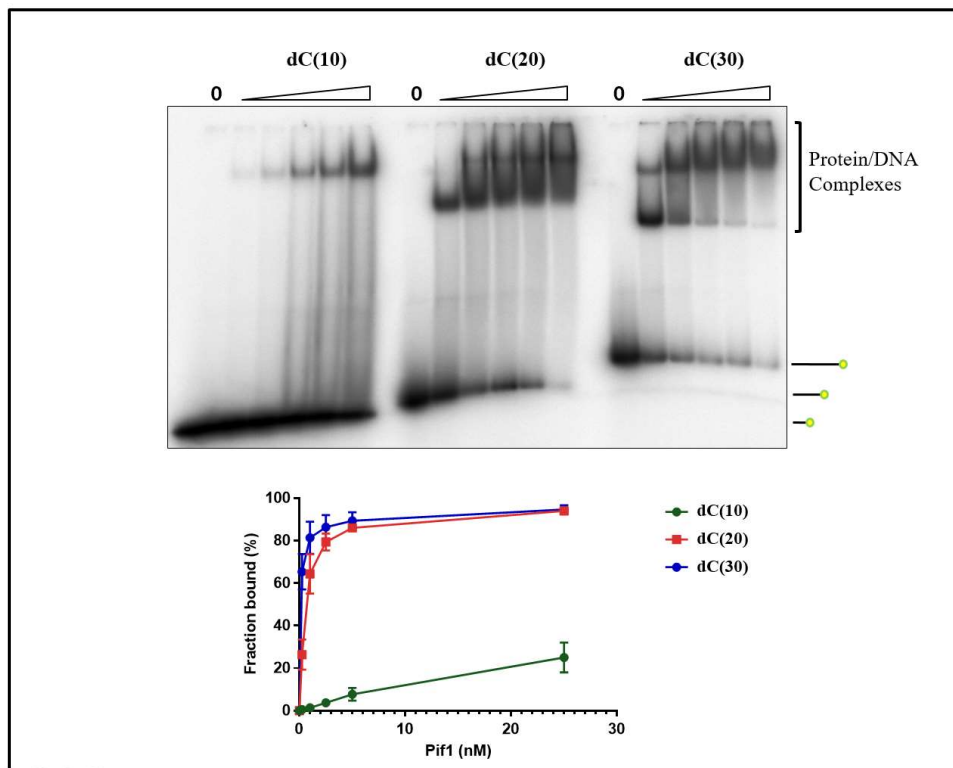
Figure 29. **Comparing binding affinities of hPif1$_{206-641}$ to different lengths of poly dC ssDNA in the presence of the co-factor ADP•AlF$_4^-$.** Lanes marked as 0 are showing the mobility of native substrates in the absence of protein. 1, 5, 25 and 100 nM hPif1$_{206-641}$ and 0.4 nM ^{32}P-labelled ssDNA substrates. Graph showing the quantified results of the assay (N=3). Error bars indicate standard deviation.

hPif1 is known to bind G4 secondary structures and it has been proposed that yeast ScPif1 has increased affinity for dG-rich ssDNA. An additional experiment comparing the binding affinity of short 12 base oligonucleotides of mixed dT and dG composition (3-4 dG bases, "G-rich") to the same length of poly dC was performed in the presence of ADP•AlF$_4^-$. Surprisingly, the data shown in Figure 30 demonstrate that poly dC$_{(12)}$ has substantially higher affinity than all other substrates tested. However, consistent with the results for ScPif1, hPif1shows a preference for binding mixed dG-dT polymers compared to homopolymeric dT$_{(12)}$.
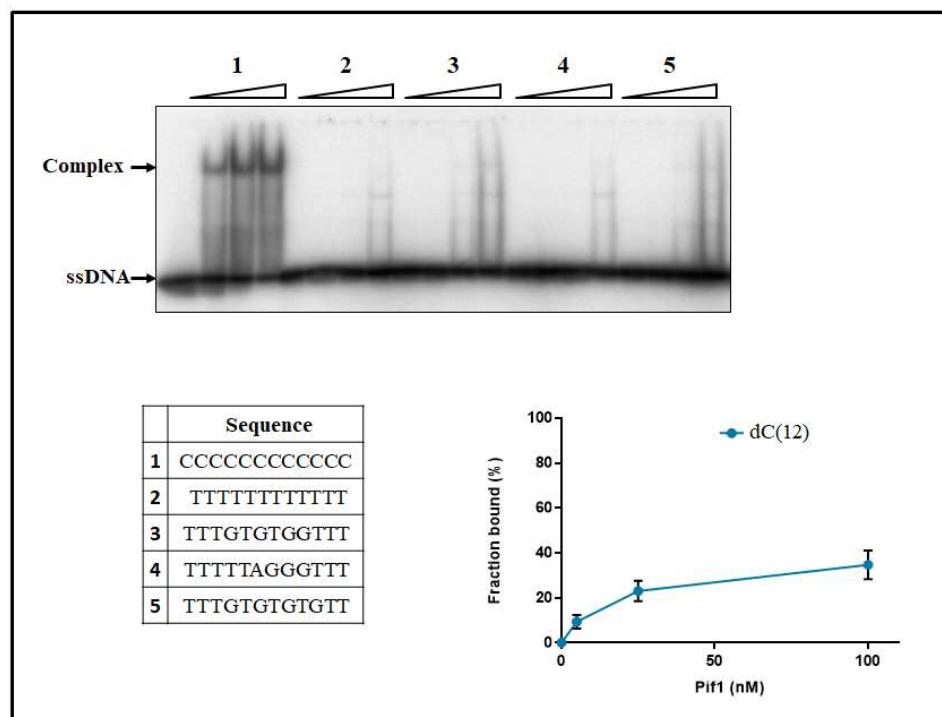
Figure 30. Comparing the DNA binding affinity of hPif1$_{206-641}$ to different polydeoxyribonucleotide sequences in the presence of ADP•AlF$_4^-$. The 12 base polydeoxyribonucleotide sequences shown were end labelled and tested for hPif1$_{206-641}$ DNA binding by EMSA (0.1 nM ^{32}P-end-labelled substrate, 5, 25 and 100 nM hPif1$_{206-641}$). Graph showing the quantified results of the assay (N=3). Error bars indicate standard deviation.

3.3 Resolution of uni-molecular G-quadruplex DNA by hPif1

The Pif1 helicase is known for its G4 binding and unwinding activity, *in vivo* and *in vitro* (Ribeyre et al., 2009, Sanders, 2010, Paeschke et al., 2011). Previously, because of the complications in using unimolecular G4 DNA substrates in electrophoresis-based assays, tetramolecular G4 substrates were used to visualize the unwinding activity. Alternatively, some studies have used more specialised and complex single molecule techniques to analyse this particular activity. Here, the unwinding of unimolecular G4 DNA by FL-hPif1 has been observed in ensemble experiments using EMSA and DNA footprinting employing DMS modification. In G4 DNA, the N7 position of guanine is protected from DMS methylation due to the formation of Hoogsteen base pairs. Protection from modification results in the absence of radioactive DNA band in DMS sequencing G-ladder displayed on a sequencing gel (denaturing PAGE).

A uni-molecular G4 DNA substrate forming 4 G tetrads was annealed as described in Section 2.4.2 for duplex DNA, in the presence of 100 mM NaCl. The substrate contained four dGGGG runs with intervening dTTTT segments, a $dT_{(55)}$ 5′ tail before the G4 forming sequence and a 20 based sequence of mixed base composition at the 3′ end (Fig. 31). A single stranded "mutant" control DNA sequence with a single G to A base change in each run of four otherwise consecutive dG residues was used to show the mobility difference between folded and unfolded G4 DNA in the electrophoresis. The sequence and structure of both substrates was confirmed by DMS-foot printing as shown in the sequencing gel image in Figure 31. This shows that the G residues are protected in the folded G4 DNA substrate but not in the boiled G4 single-stranded control, where the G residues demonstrate high reactivity to DMS and subsequent piperidine cleavage. As expected for the control ssDNA sequence, there is

no difference between the reactivity of the native substrate and the boiled control, since both are single-stranded and the G-residues susceptible to DMS modification.
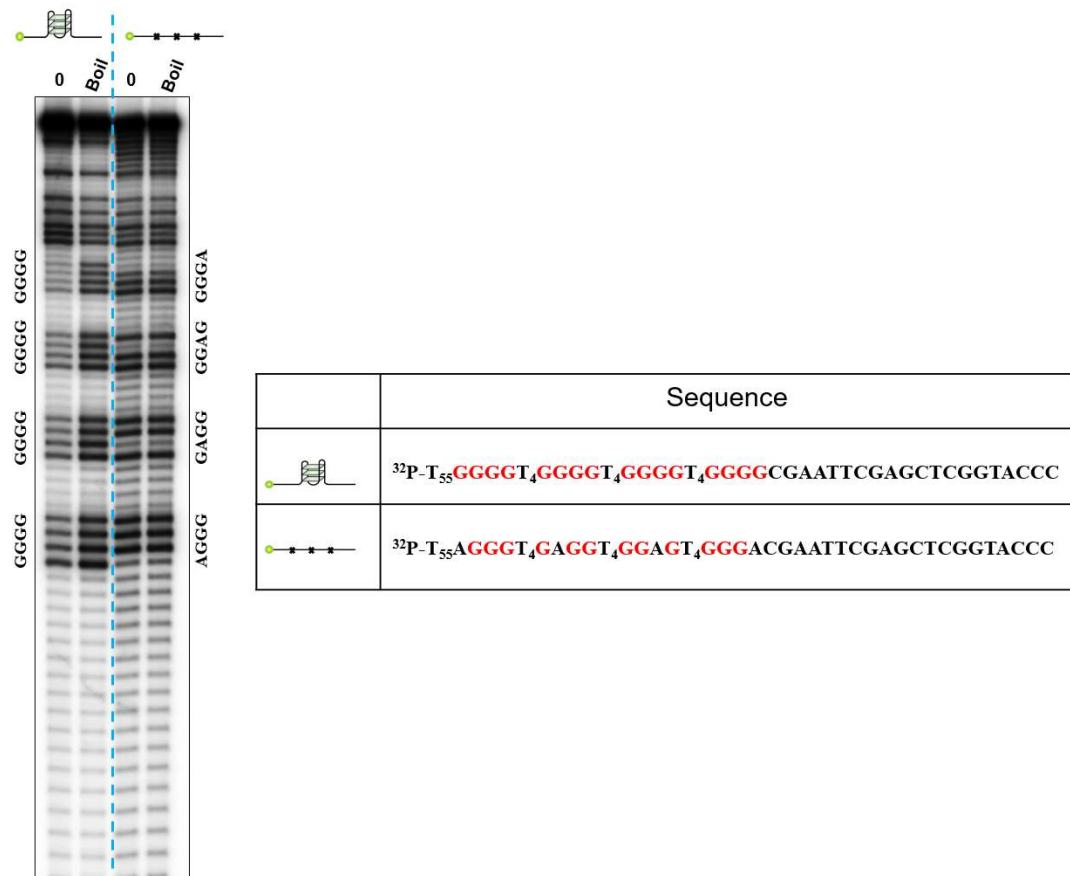
Figure 31. DMS footprinting of the unimolecular G4 DNA and mutant control substrate. The substrates used were end-labelled with [32]P. DNA footprinting experiment confirming the sequence and formation of the G-quadruplex structure. G residues are reactive in the dsDNA portion of both substrates and in the "mutant" control template, but not in the G4 DNA forming sequence of the test substrate.

The two substrates (G4 and mutant control) were annealed to a 20 base ^{32}P end - labelled complementary oligonucleotide at the 3′ end. The starting hypothesis was that when FL-hPif1 encounters the G4 DNA it will resolve it and then unwind the duplex, providing a readout for G4 DNA unwinding. As the results show in Figure 32, FL-hPif1 was able to displace almost 100% of the 20 base oligonucleotide from the control mutant template at the lowest protein concentration, but there was minimal dsDNA unwinding and strand displacement from the G4 DNA substrate (less than 10% displacement at the highest protein concentration). However, surprisingly, observation of the G4 resolving pattern on the electrophoreses gel in Figure 32 shows a small mobility shift of the substrate with increasing protein concentration, suggesting that the protein resolves the G4 DNA, but does not displace the oligonucleotide. Note, folded G4 DNA (O_2 in Fig. 32) would run faster than the resolved G4 DNA, and in this case the mobility of the putative unwound G4 DNA product co-migrates with the mutant control substrate, O_1 in Figure 32.
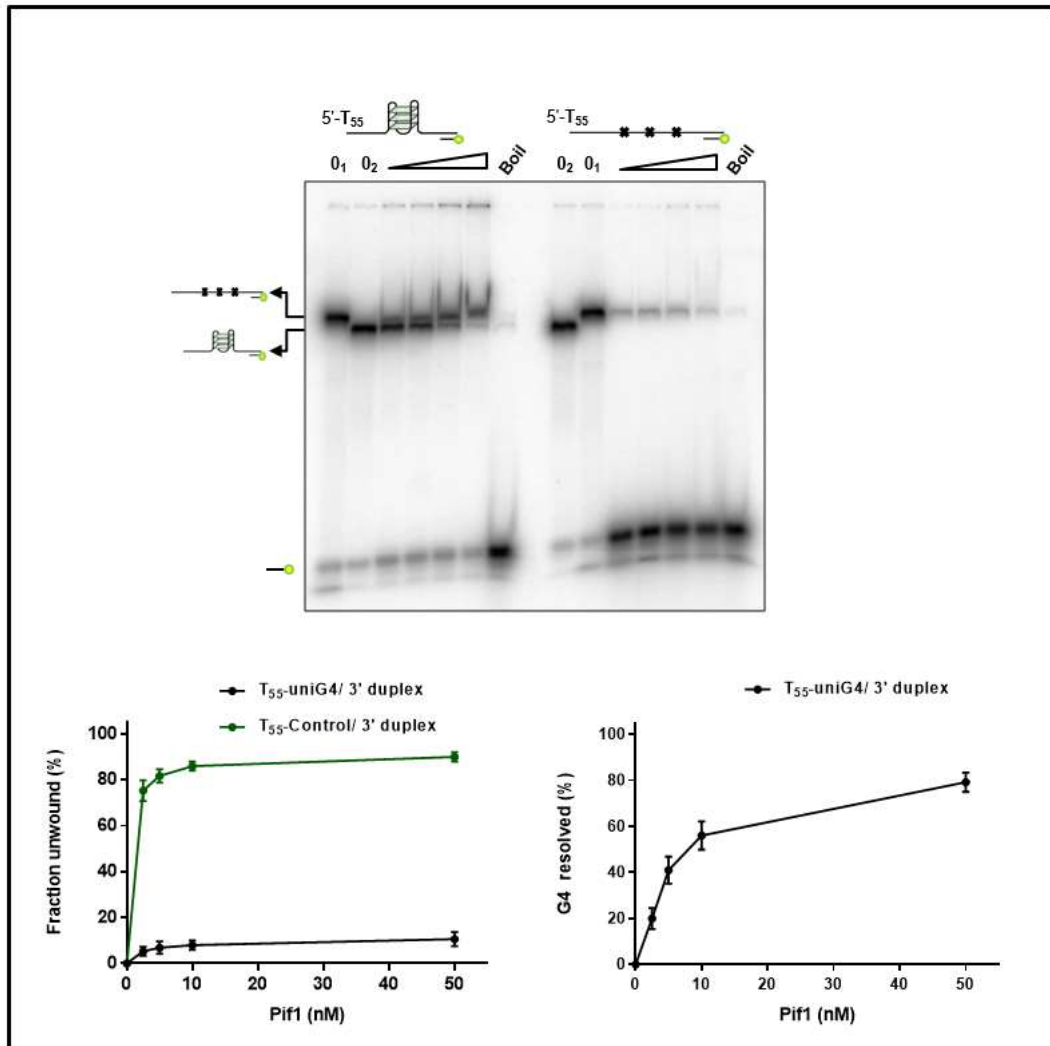
Figure 32. Unimolecular G4 helicase assay with FL-hPif1. Strand displacement assay for the unimolecular G4 DNA substrate with a 3′ duplex and the "mutant" control sequence with a 3′ duplex (**top**). 0_1 is the native mutant control substrate and 0_2 is the native G4 substrate. Note: both 0_1 and 0_2 are shown next to each other at the beginning of each set of titrations to show the mobility difference of the two. The thermally denatured substrate is shown as "Boil", which releases the radiolabelled 20 mer. The substrate concentration was 0.1 nM and the protein titration series 2.5 nM, 5 nM, 10 nM and 50 nM. The graphs below show the quantified results with standard error bars (N=3). The bottom-left graph is the quantified result of the unwound fraction of the duplex (fraction of the released 20 mer ssDNA). The bottom-right graph shows the quantified results of the putative G4 unwinding by hPif1 (intensity of shifted band).

This finding was further investigated by radio-labelling the G4 DNA-forming oligonucleotide directly (without the complimentary ssDNA annealed at the 3′ end) and the helicase assay was repeated.   As shown in Figure 33 (left), the fraction of G4 DNA substrate apparently resolved continues to increase with increasing protein concentration, as evident from the mobility shift of the product compared to substrate. Curiously, the boiled (heat-denatured) substrate displayed a minimal mobility shift, indicating that the G4 component was not dissociated or rapidly re-associated as expected. This raises the question as to why a significant amount of unwound G4 DNA product can be observed in the presence of FL-hPif1.

Next, a qualitative experiment based on DMS DNA footprinting was performed in parallel with the electrophoresis-based helicase assay on each of the protein titration points and with the no protein and thermally denatured substrate controls (Fig. 33, right).   As shown, the native G4 DNA substrate shows protection from DMS modification dependent cleavage. In the helicase reactions, protection over the G4 DNA forming sequence decreased as the protein concentration increases and the boiled control shows no modification-protection resulting in a high band intensity at the G-rich sites. To provide a quantitative approximation of the unwinding, the intensity of the G bands in the boiled control was taken as 100% modification and the native G4 substrate (0, no protein) as 0%, no modification.   A graph plotted for the quantified intensity of the protein titration series, normalized against the 0 and the control (intensity of 100%) shows that the G residues become less protected when the FL-hPif1 concentration increases. Taken together with the results of the electrophoresis-based assays, the data strongly suggest that FL-hPif1 can resolve unimolecular G4 DNA in a concentration dependent manner.
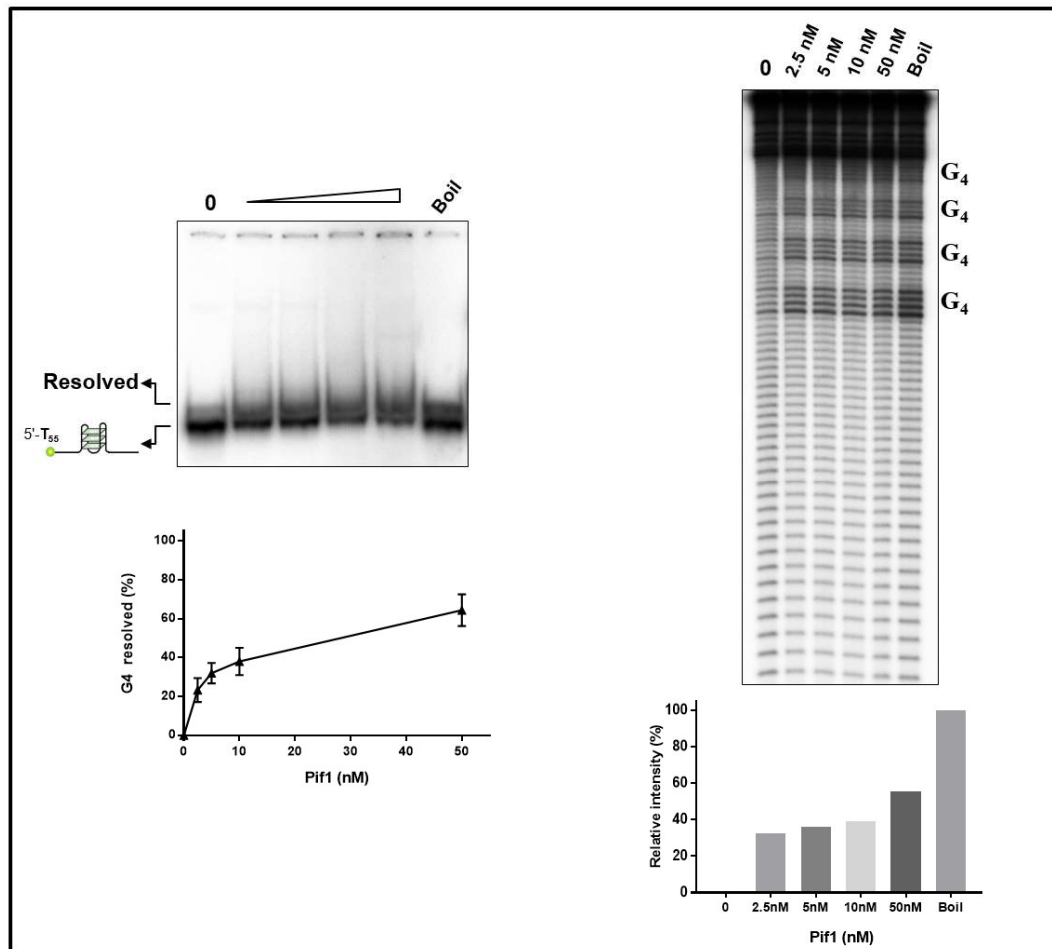
Figure 33. Unimolecular G4 DNA resolution by FL-hPif1. Left, strand displacement assay (0.1 nM substrate, 2.5, 5, 10 and 50 nM hPif1) showing the electrophoresis gel and the quantified results in a graph (N=3; resolved G4 DNA/fraction of intensity of shifted band). 0 shows the mobility of the native substrate, thermally denatured substrate indicated as Boil. Right top, sequencing gel showing the cleavage protection of the G residues of the samples after the strand displacement assay was performed. The graph plotted shows the relative intensity of the G-rich bands of each of the titration points (right, below).

3.4 Drug-like molecular inhibitor screening and structure-based drug design.

A drug-discovery campaign targeting hPif1 was initiated in this study. Several approaches were pursued: (i) A previous screen (S. D-T and Dr. Cyril Sanders, University of Sheffield) of a 9000 compound MRCT diversity set of drug-like molecules, representative of a wider group of compounds in a larger collection, via a radiometric helicase assays was exploited. (ii) A differential scanning fluorimetry (DSF) or protein thermal shift assay was established and several fragment libraries screened. (iii) Protein crystallisation and X-ray crystallography was exploited with the XChem fragment-based screening service run at the Diamond Light Source national facility (in collaboration with Dr. Ben Bax, Universities of York/ Cardiff and Prof. Fred Antson, University of York).

3.4.1 Full-length (FL) hPif1 purification

During the initial part of the study only the hPif1 helicase domain was available in sufficient quantity and purity for use as a screening tool. Ultimately, it was desirable to test all compounds against the FL-hPif1 (residues 1- 641) enzyme and to develop the expression and purification further so that it could be used as a screening reagent in its own right. Expression and purification trials of FL-hPif1 were performed prior to continuing further experiments on these compounds. Therefore, the first part of this section is focused on the purification of FL-hPif1.

3.4.1.1 Optimising FL-hPif1 protein expression

Attempts to purify a GST-tagged version of the FL-hPif1 enzyme in the ArcticExpress™ expression system proved unsatisfactory (results not shown, see Table 2). The full-length ORF was cloned into a modified version of the pET15b plasmid where the thrombin cleavage site was replaced with a TEV cleavage site, and this was employed in the ArcticExpress™ expression system that had been successful for hPif1-HD. A time-course was performed at 6°C, 9°C, 12°C and samples were collected at 4 time-points of 6h, 24h, 48h, and 72h. As the results show (Fig. 34), the optimum expression of FL-hPif1 was at 9°C for a maximum incubation time of 72 hours, although the majority of protein remained insoluble under all conditions tested.

A
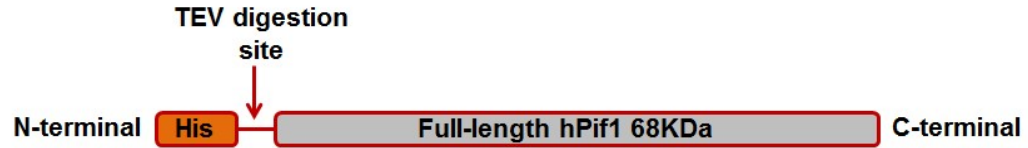
B
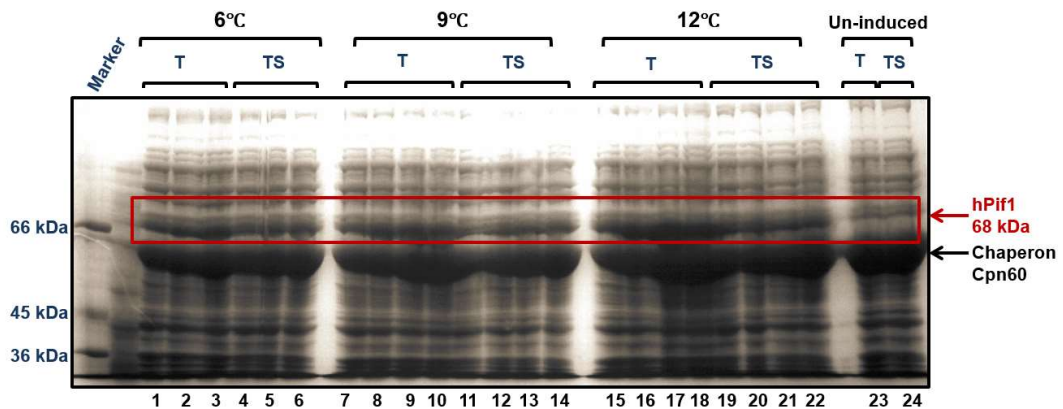
Figure 34. Expression trial of full length (FL) hPif1 in the ArcticExpress™ system. **(A)** Schematic figure of the FL-hPif1 construct fused to a histidine tag added to assist purification. **(B)** Time-course result of expression of FL-hPif1 (~68 kDa). Lanes 1, 2, 3 are total protein extract samples (T) harvested at 24, 48 and 72 hours respectively after IPTG induction at 6°C. Lanes 4, 5, 6 are the cleared lysate (total soluble, TS) samples at 6°C. Lanes 7, 8, 9,10 are total protein extract samples (T) harvested at 8, 24, 48 and 72 hours respectively after induction with IPTG at 9°C. Lanes 11,12,13,14 are the corresponding cleared lysate samples (TS, 9°C). Lanes 15,16,17,18 are total protein extract samples (T) harvested at 8, 24, 48 and 72 hours respectively after induction with IPTG at 12°C. Lanes 19, 20, 21, 22 are the corresponding cleared lysate samples (TS, 12°C). Lanes 23 and 24 are the uninduced (no IPTG) samples harvested at 72 hours, 9°C, as a negative control showing minimal/no expression of the FL-hPif1.

3.4.1.2 The finalized FL-hPif1 purification procedure

As the results of the time-course suggested, a temperature of 9°C and 72 hours incubation time was selected for large-scale culture. After several attempts and analysis of purification using different chromatogram columns and conditions (not shown) the finalized purification procedure of FL-hPif1 is summarized below. Figure 35 shows the final purification steps. Approximately 2 mg of FL-hPif1 was obtained from ~200 g of bacteria (~90 litres of culture).

Bacterial lysis, extract clarification and DNA removal with Polymin P

⇓

40% saturation ammonium sulphate cut

⇓

Application to an SP sepharose column and gradient elution (0.1-1 M NaCl)

⇓

First His-trap column (IMAC) and imidazole gradient elution (20-200 mM)

⇓

Removal of chaperone/high molecular weight material on a preparative Superdex S200 SEC column (shown in Fig. 35A)

⇓

Removing imidazole by dialysing overnight and cleavage of the His-tag by digestion with TEV protease

⇓

Second His-trap column (2-100 mM imidazole gradient) (shown in Fig. 35B)

⇓

Final polishing on a high-resolution Source S column (shown in Fig. 35C)

⇓

Concentration with Vivaspin centrifugal concentrator

Figure 35. The final three stages of FL-hPif1 purification. Chromatograms showing the UV Abs$_{280}$ (**left**) and the SDS-PAGE analysis of fractions corresponding to the UV peak (**right**). **(A)** SEC column showing the separation of the major contaminant (chaperone) from the FL-Pif1 protein after initial SP-sepharose and IMAC steps. **(B)** Second IMAC column showing the separation of the remaining contaminants after TEV protease cleavage. The undigested sample has been indicated as "U/D" and a sample of the protein after digestion is indicated with "D". **(C)** Final high-resolution Source S column showing the purity of the protein.

3.4.1.3 Optimisation of FL-hPif1 helicase assay condition

After purification, an initial test was done to determine the optimum salt concentration at which the helicase had the maximum activity. This was critical as hPif1-HD had previously been determine to be highly salt sensitive, with significant inhibition of activity above 20 mM NaCl. The experiment was done using different KCl concentrations since the dominating cation in cells is K^+. The results clearly show (Fig. 36) the protein is most active at the lower salt concentration tested (30 mM) and significantly inhibited above 70 mM KCl. Titrations of NaCl performed in parallel (Dr. C. M. Sanders, not shown) showed a very similar salt sensitivity profile, indicating that NaCl and KCl were interchangeable in assays. Also, comparing different incubation temperature the results suggested the optimum temperature for incubating the assay is at 37°C (not shown). This is in contrast to previous observations in the laboratory with hPif1-HD whose helicase activity decreases significantly above 25°C, and is inactive at 37°C.
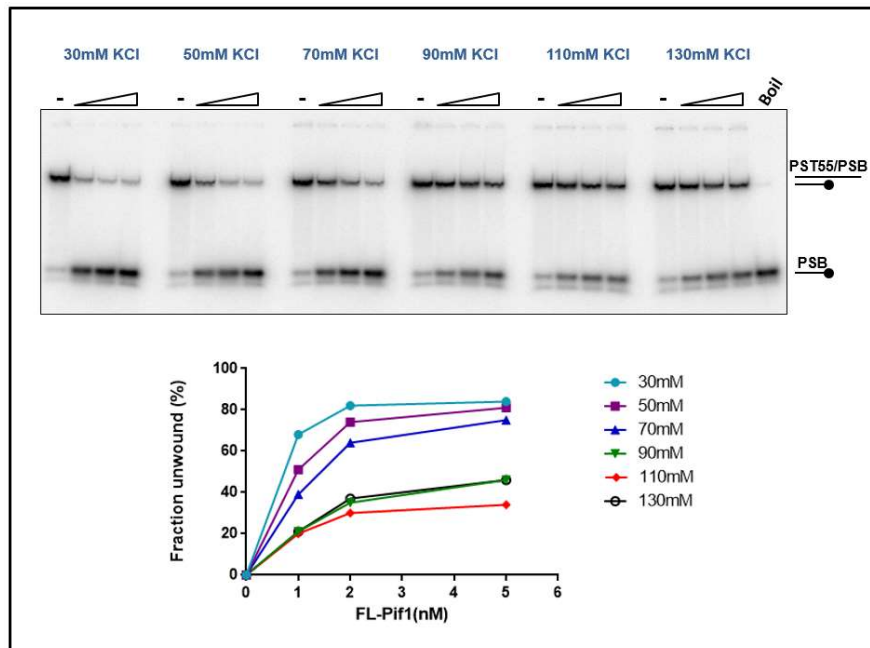
Figure 36. Salt concentration optimization of FL-hPif1 helicase activity. The gel (**top**) shows the strand displacement activity (0.1 nM substrate, 1, 2 and 5 nM FL-hPif1) as a function of KCl concentration in the otherwise standard reaction buffer (Materials and Methods). A graph of the quantified data is shown below.

3.4.2 Summary of the primary screen of the MRCT library

Initial screening of 9,000 compound
from MRCT library at 100 μM

⇓ ≃ **250**

Strong hits were re-screened
against other helicases to discard
the non-specific ones

⇓ ≃ **120**

The IC50 of the specific hits
was measured

⇓

IC50 <20 μM   ≃ **10**

IC50 <10 μM   **4**

⇓

78 similar compounds were resupplied
and screened using hPif1$_{1\text{-}641}$

A library of 9,000 compounds provided by the MRCT was initially screened using

hPif1-HD at a final compound concentration of 100 μM in the standard radiometric

helicase assay. Approximately 250 compounds were inhibitors of hPif1-HD helicase

activity and were then counter screened against the E1 helicase (a bovine

papillomavirus SF3 helicase) available in Dr. Sanders' lab. This counter screen was

used to eliminate a large number of non-specific inhibitors (180). The IC$_{50}$ (helicase

assay, the compound concentration giving 50% inhibition of unwinding under standard

assay conditions) of the remaining compounds was measured using the original (low

volume, 10-20 μl) material provided in a 384 well plate. Ten compounds were found to

have an $IC_{50} < 20$ µM and four had an IC_{50} of lower than 10 µM (Table 5). These compounds were then provided by the MRCT in larger amount (5 mg, DMSO stocks prepared) for cell-culture based assay (Appendix, Supplementary Figure 6), as well as 78 compounds with similar structural and chemical properties to the four strongest hits (analogues, $IC_{50} < 10$ µM). No structural information was provided by MRCT on any of the compounds.

MRCT Code	Destination Well number	Initial IC_{50}	Resupplied IC_{50}
MRT00105813	H2	4.4	2.9
MRT00203960	E3	7.3	4.7
MRT00153955	G3	4.1	1.7
MRT00044704	E4	8.7	16.4

Table 5. List of four compounds that had an IC50 < 10 µM and their IC50 values from re-supplied stock.

3.4.2.1 Screening of selected hits and their related analogues

The 78 analogues were provided by MRCT for further experiments in a 96 well plate format, ~10-20 µl at 10 mM in DMSO. An initial wide screening of these compounds was performed to select the strongest inhibitors for subsequent IC_{50} determination. All the compounds were screened first at 100 µM concentration (Fig. 37A) and then at a concentration of 25 µM in the standard radiometric helicase assay using FL-hPif1. An IC_{50} value was determined for the ones which show significant inhibition at 25 µM, as indicated in Figure 37B.

A
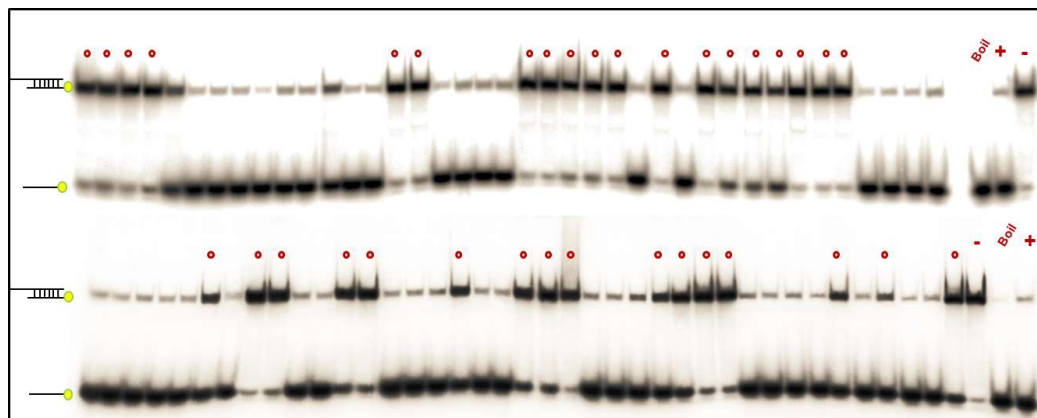
B
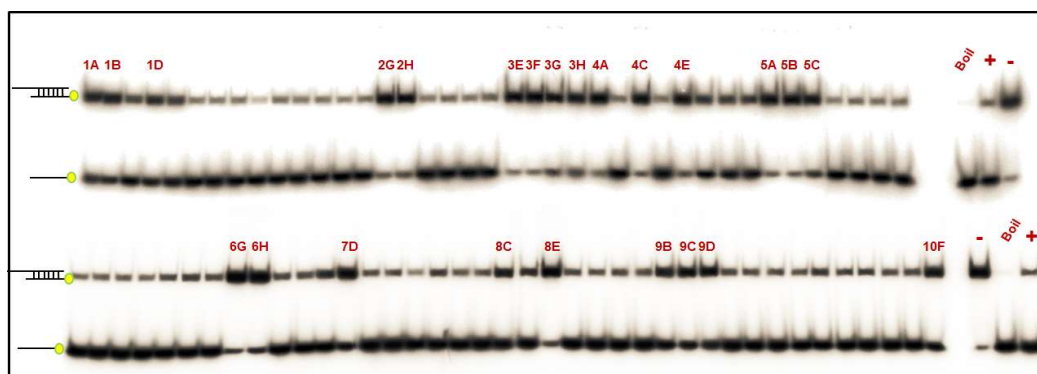
Figure 37. Secondary screen of the hit analogues from the MRCT library. The initial screening of 78 compounds structurally related to the 4 strongest inhibitors identified in the primary screening. **(A)** Screening at 100 μM compound concentration in the helicase assay. Strong inhibitors are marked by red dots. **(B)** Secondary screening at 25 μM compound concentration loaded in the same order as section A. The strongest inhibitors are indicated by their codes provided by the supplier (96 well plate location). No compound/DMSO only reaction (+) was used as a positive control (0.1 nM substrate, 4 nM FL-hPif1). Negative control (only substrate) is labelled as (**-**) and the heat denatured substrate is indicated as (Boil).

3.4.2.2 Measuring IC$_{50}$ values for selected compounds

The IC$_{50}$ (helicase assay) of the selected compounds from Figure 37B were measured, covering the compound titration range of 0 µM to 250 µM. Examples of an IC$_{50}$ measurement, which was performed for each of the selected 24 compounds, are shown in Figure 38. The results are summarized in Table 6, which includes additional information described below.

UPF1 is a related eukaryotic SF1 helicase that has multiple functions including roles in nonsense mRNA decay (NMD), telomere maintenance and a contribution to genome stability. The structural similarity, although distant, with hPif1 makes it a good candidate for counter screening. The full-length hUPF1 helicases was previously purified by myself in Dr. Sanders' lab (Dehghani-Tafti and Sanders, 2017). Counter screening of the selected most potent hits (lowest IC$_{50}$ values) was performed against hUPF1 under optimum helicase reaction condition at two compound concentrations of 100 µM and 25 µM to assess specificity (Table 6). In addition, a protein thermal shift experiment was performed on all of the selected 24 compounds using FL-hPif1 to identify direct binders (shown in column 6 of Table 6). Also included in column 5 are the results of preliminary tests (one measurement) to determine the possible inhibition of hPif1 strand annealing activity by the compounds assayed at 25 µM (+ indicates inhibition and – indicates no inhibition).

As there was no structural information provided by the MRCT the hits are labelled based on their well numbers. Eight compounds (highlighted in red in the table) are ones selected by the MRCT after analysis of the data and resupplied in dry powder format (~5 mg) for further experiments. The initial experiments were to measure the IC$_{50}$ (helicase assay) of the resupplied compounds using hPif1 and hUPF1 and also their

effect on the proteins strand annealing activity to confirm their reproducibility (Table 6 column 7, 8, 9 and Fig. 38). All except G2 and D9 produced very similar numerical values as they did when stocks from the original plates 96 well plate were assayed (column 2). These were then tested in triplicate in the protein thermal shift assay (PTS) using hPif1$_{206-641}$. As shown in column 10, only compound A5 (highlighted in green) gave a significant, reproducible, thermal shift (ΔTm = 2.1 $\pm$0.3°C). Several of the compounds resulted in uninterpretable fluorescent readout and thermal denaturation profiles as the software could not produce a melting curve for them. This is likely due to interfering fluorescent properties and these compounds are marked as (*) in Table 6. 27 analogues of compound A5 were supplied by MRCT and tested for thermal shift in 6 independent repeats. Other than the original compound (A5) only two analogues produced a significant thermal shift when assayed at 200 μM final concentration. However, no inhibition of enzymatic activity was observed when tested in the helicase assay (Table 7).

1	2	3	4	5	6	7	8	9	10
Well	IC$_{50}$ (µM) Original supply	% Inhibition of UPF1 at 100 µM	% Inhibition of UPF1 at 25 µM	Strand-annealing inhibition at 100 µM	ΔTm °C Pif1$_{1-641}$ at 100 µM	IC$_{50}$ (µM) Pif1$_{1-641}$ (resupply)	IC$_{50}$ (µM) UPF1	Strand-annealing inhibition at 100 µM	ΔTm °C Pif1$_{206-641}$ at 100 µM
A1	92.2	0	0	-	0				
B1	9.4	0	0	+	*	8.8	> 250	+	*
D1	55.3	45%	26%	-	1.1				
G2	9.6	53%	36%	+	1.2	> 250	> 250	-	0
H2	2.9	50%	43%	-	*				
E3	4.7	71%	52%	+	0				
F3	4.1	80%	72%	+	2.8				
G3	1.7	48%	64%	+	*				
H3	9.6	57%	42%	+	2.1				
A4	7.7	55%	38%	+	3.7	8.4	6.4	+	0
C4	2.8	80%	68%	+	0	4	5.3	+	0
E4	16.4	0	0	+	0				
A5	3.4	78%	56%	+	5.8	3.3	5.1	+	2.1 ±0.3
B5	3.6	80%	72%	+	0				
C5	9.7	71%	50%	+	0	8.9	8.1	+	0
G6	5.7	69%	65%	+	0	9.7	4.7	+	0
H6	40.5	26%	8%	-	0				
D7	99.7	0	0	-	0				
C8	49.5	0	0	-	0				
E8	40.8	17%	0	-	0				
C9	3	ND	ND	ND	*				
D9	6.9	51%	30%	+	*	> 250	> 250	-	0
B9	52.1	83%	38%	+	0				

Table 6. Summary of the *in vitro* activity results of the compounds in the secondary screen of the MRCT library. List of compounds selected from secondary screening and their measured IC$_{50}$ values and approximate strength of UPF1 inhibition. Columns 1-6 are results generated from the original stock (96 well plate format) and columns 7-10 are the results from resupplied dry stock (dissolved in DMSO). Note: although the IC$_{50}$ of compound C9 was measured, there was a significant amount of DNA substrate smearing and retention in the wells of the electrophoresis gel therefore the results weren't quantifiable for the other assays.

114

1B

$IC_{50} = 9.4\ \mu M$

4A

$IC_{50} = 8.4\ \mu M$

4C

$IC_{50} = 4\ \mu M$

5A

$IC_{50} = 3.3\ \mu M$

5C

$IC_{50} = 8.9\ \mu M$

6G
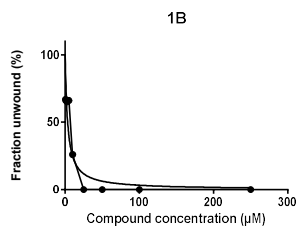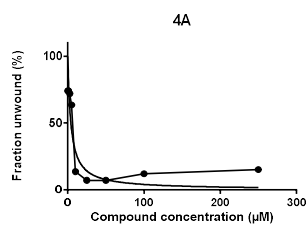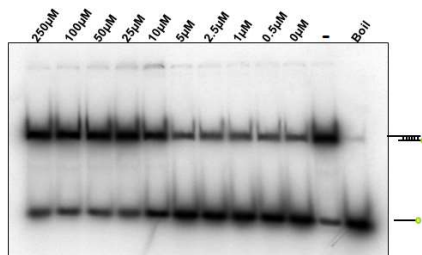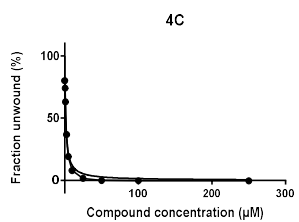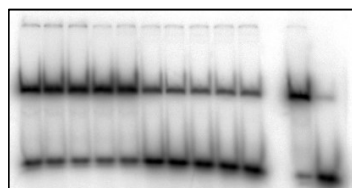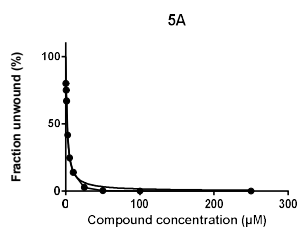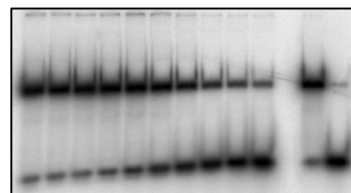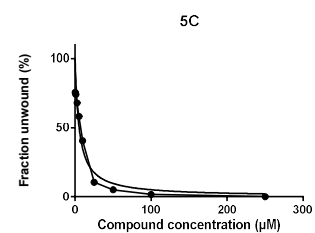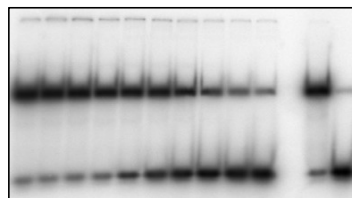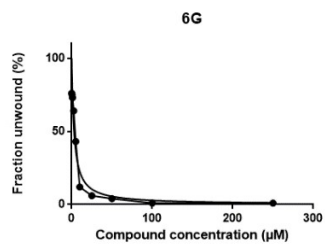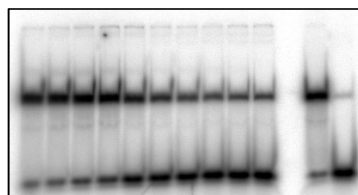
$IC_{50} = 9.5\ \mu M$
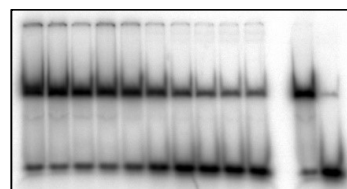
Figure 38. Determination of IC₅₀ values of selected compounds. Examples of helicase assays (0.1 nM substrate, 4 nM FL-hPif1) and graphs of plotted data for measuring the IC_{50} values of selected/ resupplied compounds from the original stock highlighted in Table 6. As labelled for compound 1B, the titration series of the compounds used was 250 µM, 100 µM, 50 µM, 25 µM, 10 µM, 5 µM, 2.5 µM, 1 µM, 0.5 µM, 0 µM (positive control +DMSO only). No enzyme negative control (only substrate) is indicated with (-) and heat denatured substrate is indicated as "Boil". Subsequent assays (not labelled) follow the same labelling and loading order.

	Structure	Formula/ Supplier ID (Enamine)		IC_{50}	ΔTm °C at 200 µM
A5		Formula $C_{14}H_{11}ClN_2O_3S$ MW 322.8 Z110096906		3.3 ±0.2 µM	2.1 ±0.3°C
1		Formula $C_{13}H_9ClN_2O_3S$ MW 308.7 Z29215461		> 250 µM	1.5 ±0.2°C
2		Formula $C_{13}H_{10}N_2O_4$ MW 258.2 Z29215571		> 250 µM	1.1 ±0.2°C

Table 7. Properties of a candidate inhibitor from the MRCT library. List of the selected homologues of compound 5A and their measured IC_{50} and thermal shift values (N=3).

3.4.3 Fragment library screening

Fragment library screening was undertaken to initiate a fragment-based drug design approach to hPif1 inhibition. In this approach, a library of small molecules is screened initially using a biophysical assay to select the molecules that directly bind to the protein. Subsequently a biochemical screen is performed to measure the effects of the bound molecules on the protein's enzymatic activity.

The protein thermal shift (PTS) assay was established and optimized for the hPif1$_{206-641}$ helicase domain construct and the full-length protein (FL-hPif1). As described previously, the 1.1Å crystal structure of the hPif1 helicase domain confirms the direct binding of AMP-PNP (Figs. 18 & 20), therefore this molecule was used throughout the screening as a positive control. Binding of AMP-PNP stabilises the protein, consequently increasing its melting temperature. Comparing the melting temperature of the protein alone to the melting temperature of the protein in the presence on AMP-PNP we observe an increase which is referred to as ΔTm.

A small molecule fragment library at 30 mM concentration in DMSO was supplied by the MRCT in 386 well plates (~20 µl volume). Consequently, a prior optimization of the PTS assay in the presence of different DMSO concentrations was performed, as shown in Figu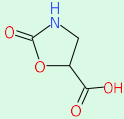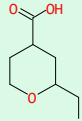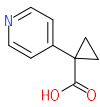re 39. The results show throughout that there is 2.1-2.3℃ difference in melting temperature between protein alone (hPif1$_{206-641}$) and hPif1$_{206-641}$ +AMP-PNP, at DMSO concentrations between 2-10%. However, the protein melting temperature decreases with increasing DMSO concentration, indicating that DMSO alone does affect the stability of the protein. The concentration of DMSO was adjusted to 4% in the reaction throughout the screening. Also the speed of increasing temperature was optimised by comparing two ramp speeds (1℃ per minute, 1%, and 2℃ per minute, 2%). As the results showed the higher speed (2%) was better in terms

of obtaining higher stability of the protein (Fig. 40), so this setting was used throughout the following PTS experiments.

Due to initial limitations in producing the FL- hPif1 protein in sufficient quantity, the primary screens couldn't be carried using this protein. Also, for the verification of binding of the hits to FL-hPif1 the protein could not be used at the standard concentration (minimum of 3 µM) of hPif1$_{206-641}$ used for the primary PTS screen. Therefore, minimizing the consumption of the full-length protein for use in the final assays was necessary. As the results in Figure 41 indicate, the concentration can be minimized to 0.25 µM FL-hPif1, although 0.5 µM was used for final rounds of screening as explained below.
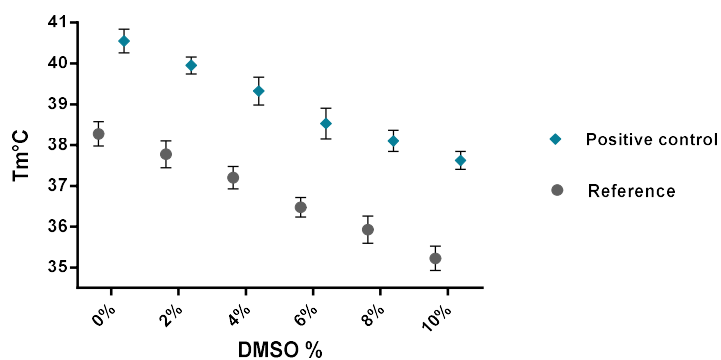
	Reference			+AMP-PNP			
DMSO %	Mean	SEM	N	Mean	SEM	N	ΔTm°C
0%	38.2	0.15	4	40.5	0.14	4	3.2
2%	37.7	0.15	4	39.9	0.10	4	2.2
4%	37.2	0.15	4	39.3	0.17	4	2.1
6%	36.4	0.11	4	38.5	0.18	4	2.1
8%	35.9	0.16	4	38.1	0.13	4	2.2
10%	35.2	0.15	4	37.6	0.11	4	2.3

Figure 39. Measured the melting temperature of hPif1-HD at different DMSO concentrations. "Reference" indicates the protein alone (with and without DMSO) and the positive control is shown as +AMP-PNP (with and without DMSO). SEM is standard error of the mean and N is number of independent repeats of the assay.
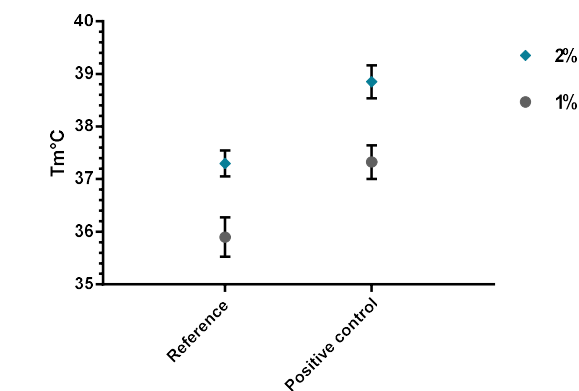
	2%			1%		
	Mean	**SEM**	**N**	**Mean**	**SEM**	**N**
Reference	37.3	0.12	4	35.9	0.18	4
+AMP-PNP	38.8	0.15	4	37.3	0.16	4

Figure 40. Optimizing the ramp speed in the PTS assay. Measured melting temperature of hPif1$_{206-641}$ at different ramp speeds (the rate of increase in temperature induced by the real time PCR machine), 2°C per minute (2%) and 1°C per minute (1%). "Reference" indicates the protein alone (+ DMSO) and the positive control is shown as +AMP-PNP (+ DMSO). SEM is the standard error of the mean and N is the number of independent repeats of the assay.
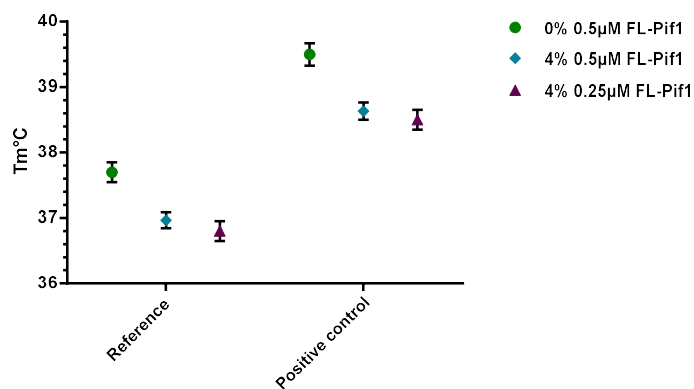
	0%DMSO, 0.5µM hPif1			4%DMSO, 0.5µM hPif1			4%DMSO, 0.25µM hPif1		
	Mean	**SEM**	**N**	**Mean**	**SEM**	**N**	**Mean**	**SEM**	**N**
Reference	37.7	0.15	3	36.9	0.12	3	36.8	0.15	3
+AMP-PNP	39.5	0.17	3	38.6	0.13	3	38.5	0.15	3

Figure 41. Measured melting temperature of full-length hPif1. "Reference" indicates the protein alone and the positive control is shown as +AMP-PNP. SEM is the standard error of the mean and N is number of independent repeats of the assay.
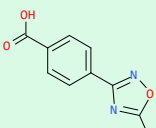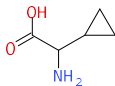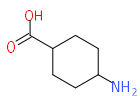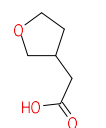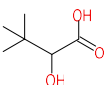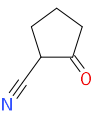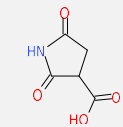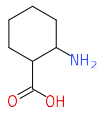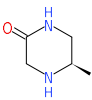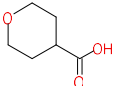
3.4.3.1 MRCT Fragment Library Screening
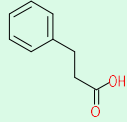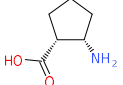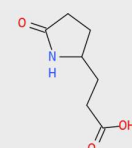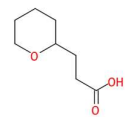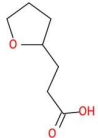
2200 chemical fragment molecules were screened by PTS as summarized in Figure 42. The entire library was initially screened twice at a final compound concentration of 1.2 mM using the hPif1$_{206-641}$ construct (the primary screen). The positive control molecule used was AMP-PNP which at 0.5 mM concentration and 4% DMSO gives a ~3°C thermal shift in melting temperature compare to the protein alone (the "Reference"). The DMSO concentration was adjusted in all control reactions to parallel the concentration determined by fragment addition (4%). The selection of fragments for further analysis after the two primary screens was based on the reproducibility of the thermal shift values and the scale of the shift (mostly ones having a ΔTm > 0.5°C). Approximately 300 hits were selected and re-located ("picked") into a new 384 well plate and diluted to half the initial concentration with DMSO and re-screened at a final concentration of 600 µM. Approximately 200 molecules were selected and re-screened at 200 µM final concentration using both hPif1$_{206-641}$ and FL-hPif1. Finally, 67 molecules with significant and reproducible ΔTm values were selected as listed in Table 8. Also, prior to this library another fragment library of approximately 1200 molecules was screened by me in collaboration with Prof. Roderick E Hubbard and his group at the University of York which were not further investigated after initial screening. The results are summarized in Appendix, Supplementary Table 2 & 3.
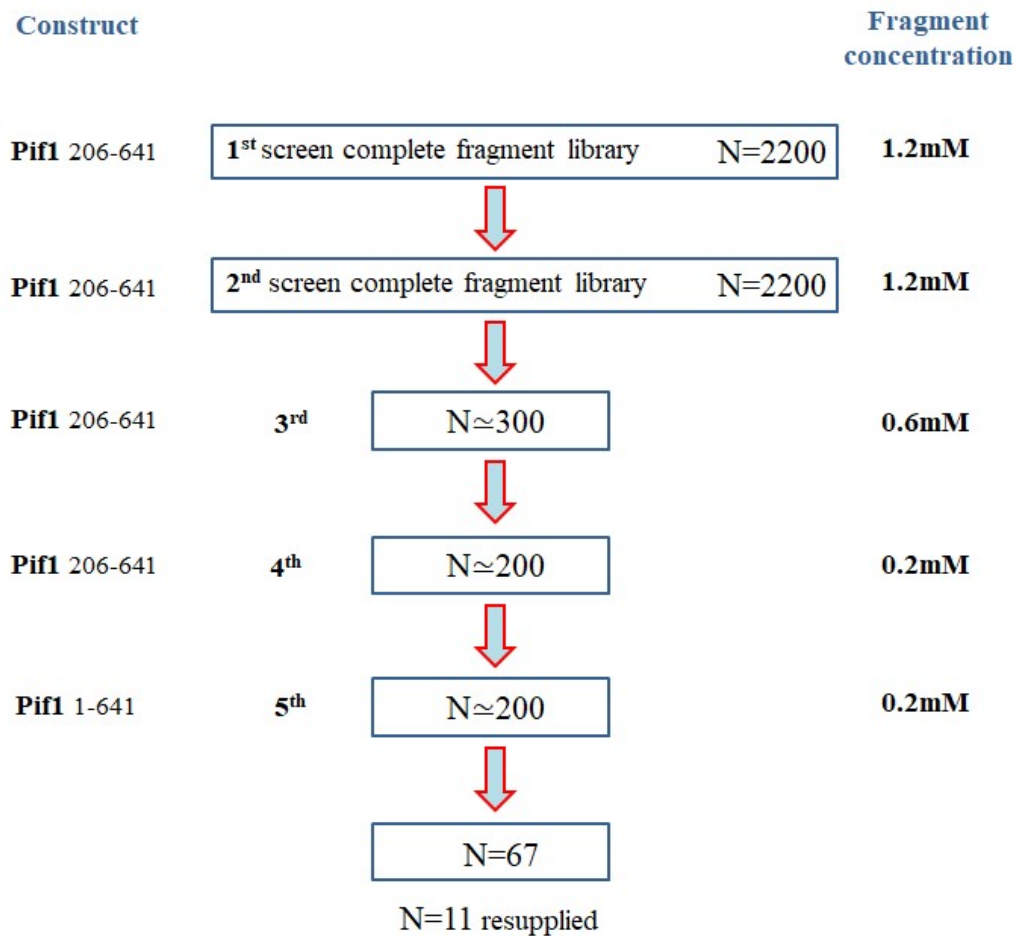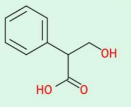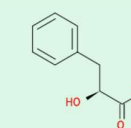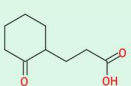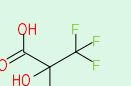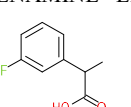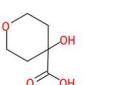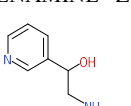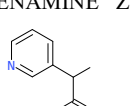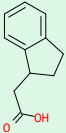
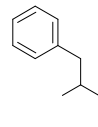**Figure 42. Work flow scheme of the fragment library screening by protein thermal shift assay.** Left, protein construct used for each round of screening. Middle, number of fragments selected and screened in each step. Right, final concentration of the fragments used in the assay at each step.
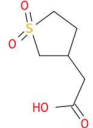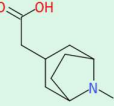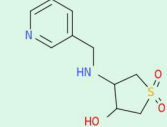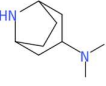
	Pif1 construct	206-641	206-641	206-641	206-641	1-641
	Chemical Fragment well number /supplier code and structure	ΔTm°C 1.2 mM	ΔTm°C 1.2 mM	ΔTm°C 600 μM	ΔTm°C 200 μM	ΔTm°C 200 μM
1	P1-P20 MRCT-INV13996	0.6	0.7	0.8	0.5	1.3
2 **8**	P2-B05 ENAMINE Z802856032	1.1	0.9	1.7	0.5	0.5
3	P2-D15 ENAMINEZ 336079706	0.8	0.8	1	0.7	0.3
4	P2-F17 ENAMINE Z1205495408	1	1	1.3	0.7	0.3
5	P2-F21 ENAMINE Z803056544	1.1	0.9	0.6	0.6	0.6
6	P2-J03 ENAMINE Z951172672	1.3	1.1	1	0.7	0.5
7	P2-L19 CHESS 3061	1.5	1.5	1.6	1.5	0.3

	Pif1 construct	206-641 ΔTm°C 1.2 mM	206-641 ΔTm°C 1.2 mM	206-641 ΔTm°C 600 μM	206-641 ΔTm°C 200 μM	1-641 ΔTm°C 200 μM
	Chemical Fragment well number /supplier code and structure					
8	P2-N15 LIFE F2102-0060	1	1	1	0.9	0.4
9	P2-D10 CHEMBRIDGE 4082079	1.3	0.9	1.7	0.7	0.3
10	P2-D14 ASINEX BAS 06502515	1.5	0.8	1.5	0.6	0.5
11	P2-F04 ENAMINE Z1203735066	1.1	1	1.3	0.7	0.9
12	P2-H06 ENAMINE Z1198171568	2.1	1.1	1.3	1	1.1
13	P2-H10 ENAMINE EN300-126862	1.3	1.9	2	0.9	0.9
14 **10**	P2-H16 ARK AK100124	1.4	1.3	1.2	0.9	0.5
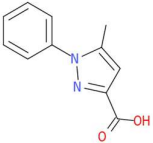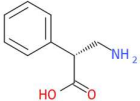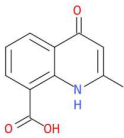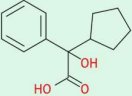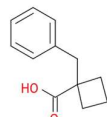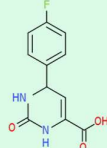
	Pif1 construct	206-641 ΔTm°C 1.2 mM	206-641 ΔTm°C 1.2 mM	206-641 ΔTm°C 600 µM	206-641 ΔTm°C 200 µM	1-641 ΔTm°C 200 µM
	Chemical Fragment well number /supplier code and structure					
15	P2-J10 ENAMINE Z57127332	1.3	1	2	1.6	0.8
16 **6**	P2-J14 ENAMINE Z985201642	1.6	1.2	2	1.5	0.5
17	P2-J20 ASINEX BAS 08835363	1.4	1.2	2	1.5	0.5
18	P2-L04 VITAS M LAB STK133208	1.3	1.1	1.2	0.7	0.8
19	P2-L06 ENAMINE EN300-74198	1	0.9	0.8	0.6	0.7
20	P2-L08 ENAMINE Z1192360086	1.2	1	1.1	0.7	0.8
21	P2-L10 ENAMINE Z362564576	0.9	1	1.6	1	0.6
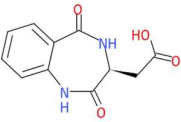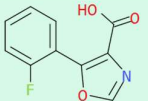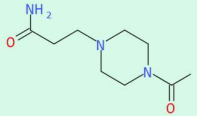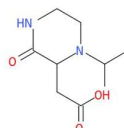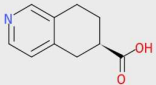
	Pif1 construct	206-641	206-641	206-641	206-641	1-641
	Chemical Fragment well number /supplier code and structure	ΔTm°C 1.2 mM	ΔTm°C 1.2 mM	ΔTm°C 600 μM	ΔTm°C 200 μM	ΔTm°C 200 μM
22	P2-L14 ARK AK-90070	0.9	0.9	1.7	1	0.9
23	P2-L16 CHEM-IMPE 9427	1	1	2.2	1	0.6
24	P2-L20 ASINEX BAS 10189614	0.9	1.4	1	0.6	1.1
25	P2-N12 APOLLO PC7674	1.4	1	1.7	1.1	0.6
26	P2-N14 ENAMINE EN300-90542	0.9	0.9	1.2	0.6	0.8
27	P2-N16 ENAMINE Z788000910	1.2	0.7	1.3	1	0.7
28	P2-P08 ENAMINE Z1192359706	1.3	1	1.6	0.9	0.5
29	P2-P14 ENAMINE Z792377268	1.1	0.7	1.3	0.5	0.5
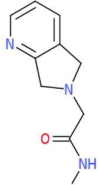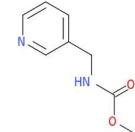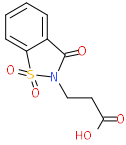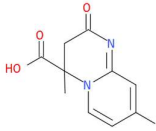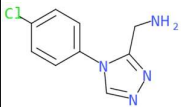
	Pif1 construct	206-641	206-641	206-641	206-641	1-641
	Chemical Fragment well number /supplier code and structure	ΔTm°C 1.2 mM	ΔTm°C 1.2 mM	ΔTm°C 600 μM	ΔTm°C 200 μM	ΔTm°C 200 μM
30	P3-E15 ENAMINE Z352782144	1.3	1.3	1.9	1	0.8
31	P3-E19 ENAMINE Z826527398	1.6	0.7	1.5	0.6	0.8
32	P3-G03 ENAMINE Z316258914	1.2	1.2	1.2	0.5	1.2
33	P3-G09 ENAMINE EN300-39278	0.9	0.9	1.2	0.6	0.8
34	P3-G19 VITAS M LAB STK682998	1.5	1	1.8	1	2.8
35	P3-I19 SYNTHONIX A15702	1.3	0.9	1.2	0.8	1.1
36	P3-K05 ENAMINE Z792377274	0.9	0.9	0.9	0.7	0.8
37	P3-K11 ENAMINE Z513624710	0.6	0.8	1.1	0.8	0.8
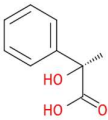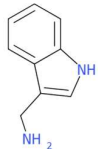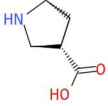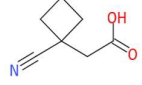
	Pif1 construct	206-641	206-641	206-641	206-641	1-641
	Chemical Fragment well number /supplier code and structure	ΔTm°C 1.2 mM	ΔTm°C 1.2 mM	ΔTm°C 600 µM	ΔTm°C 200 µM	ΔTm°C 200 µM
38	P3-O05 ENAMINE Z57162030	1	0.8	1.3	0.9	0.6
39	P3-O09 ENAMINE EN300-70609	1.3	0.9	2.4	1.1	0.9
40	P3-C04 ENAMINE EN300-116391	1.5	0.9	1.3	0.7	1.1
41	P3-G10 ASINEX BAS 03013071	0.6	1.5	2.5	1.5	2
42 **1**	P3-G16 ENAMINE 2213780	1.2	1.5	1.5	0.7	0.9
43 **5**	P3-I08 MATRIX 21798	1.7	0.9	1.4	0.6	0.8
44	P3-I12 VITAS M LAB STK397419	1.2	1.7	1.7	0.6	0.7

	Pif1 construct	206-641	206-641	206-641	206-641	1-641
	Chemical Fragment well number /supplier code and structure	ΔTm°C 1.2 mM	ΔTm°C 1.2 mM	ΔTm°C 600 µM	ΔTm°C 200 µM	ΔTm°C 200 µM
45	P3-I14 LIFE F2164-0026	1	1.6	1.5	0.7	0.4
46 **9**	P3-K06 ENAMINE Z1033088270	0.8	1.4	1.6	0.6	0.6
47	P3-K16 OTAVA 7019081226	1.1	1.3	1.5	1	0.8
48	P3-M08 VITAS M LAB STK506752	0.7	0.7	1.4	0.7	1.1
49 **2**	P3-M10 ALFA AESAR H34382	1.5	1.3	1.6	0.5	0.6
50	P3-O14 OTAVA 7020300451	1.3	1	1.5	0.7	0.9

	Pif1 construct	206-641	206-641	206-641	206-641	1-641
	Chemical Fragment well number /supplier code and structure	ΔTm°C 1.2 mM	ΔTm°C 1.2 mM	ΔTm°C 600 μM	ΔTm°C 200 μM	ΔTm°C 200 μM
51	P3-J17 PRINCETON 3DF0086	2.1	1.8	1.2	0.7	0.6
52	P3-L19 COMBI BLOCKS 3DF0037	0.7	1	1.2	0.5	1
53	P3-P03 ENAMINE 3DF0062	1	0.9	1.1	0.9	0.9
54	P3-P06 IBS 3DF0028	1.1	0.7	1.3	0.5	1
55	P3-P12 CHEMBRIDGE 3DF0033	0.8	0.7	1.1	1	0.7
56	P4-G03 PRINCETON 3DF0165	0.9	1.1	1.4	0.5	1.1
57	P4-G15 TAKEDA_FRAG B0257765	1.4	1.1	1.3	1	0.5

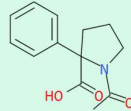|	Pif1 construct	206-641	206-641	206-641	206-641	1-641
	Chemical Fragment well number /supplier code and structure	ΔTm°C 1.2 mM	ΔTm°C 1.2 mM	ΔTm°C 600 µM	ΔTm°C 200 µM	ΔTm°C 200 µM
58	P4-M09 TAKEDA_FRAG B0245526	0.9	0.6	1	0.4	0.9
59	P5-J08 TAKEDA_FRAG B0270256	1.1	1	1.1	0.4	0.5
60 **4**	P6-G13 LIFE F0900-3328	1	1.3	0.9	0.5	0.6
61	P6-M13 LIFE F0918-7077	1.1	1.2	0.9	0.6	0.7
62	P6-E10 LIFE F2123-0002	0.8	1.3	1.3	0.4	0.6
63 **7**	P6-G16 LIFE F2147-0864	1	1.4	1.1	0.8	0.6

	Pif1 construct	206-641 ΔTm°C 1.2 mM	206-641 ΔTm°C 1.2 mM	206-641 ΔTm°C 600 µM	206-641 ΔTm°C 200 µM	1-641 ΔTm°C 200 µM
	Chemical Fragment well number /supplier code and structure					
64	P2-H22 ENAMINE Z1154245575	1.3	0.6	1.4	0.8	0.7
65 **3**	P6-P20 LIFE F2135-1150	3.2	3.5	2.3	1	1
66 **11**	P3-E08 BIOBLOCKS AA698-2	1.4	1.3	1.1	0.9	2.6
67	P2-H13 ENAMINE EN300-187151	0.9	1.3	1.4	0.7	0.6

Table 8. Results of the MRCT fragment library screening. Final fragments list and values (ΔTm) related to them at each step of the screening. The fragments numbered in red are the ones resupplied for final evaluation. Row highlighted in green are compounds which gave a significant and reproducible thermal shift with FL-hPif1 as well as the helicase domain used in the initial screening. Rows in grey are compounds that also gave a significant and reproducible thermal shift with the helicase domain but a reduced thermal shift with FL-hPif1.

3.4.3.2 Biochemical inhibition of hPif1 by selected fragment molecules

The 67 fragment hits were all tested 3 times independently at a final concentration of 1.2 mM in the helicase assay, under standard conditions, to determine whether they had any inhibitory activity against FL-hPif1. These assays were performed using the low volume of remaining compound from the original library which constrained the scope of the initial analysis. Each repeat was incubated for different times, 1 hour, 45 min and 30 min. The initial data for one hour's incubation showed no inhibition by any fragment which is why the decision was made to incubate for a shorter time of 45 min and then 30 min. Since the unwinding extent observed in the assays is time-dependent and because the inhibitors are likely to be of low affinity and reversible it was considered possible that extended incubation times could mask inhibitory activity. Figure 43 shows the electrophoretogram of the helicase assays (30 minutes incubation) for the 67 fragments. The graph below is plotted for the values obtained from the 30 and 45 minutes' incubation periods for fragments demonstrating a level of inhibitory activity which reproduced at both time points (the top value is for 30 min incubation and bottom one is for 45 min incubation).

Selection of fragment molecules for further investigation was based on the highest thermal shift values (Table 8) and the biochemical inhibition observed (Fig. 43) and further constraints such as affordability and availability. Fragments 2, 14, 16, 42, 43, 46, 49, 60, 63, 65, 66 were purchased as dry powders (re-numbered 1-11 in red in Table 8). Although fragment 16 (6) and 66 (11) showed no inhibition of helicase activity they were selected because of their relatively high and reproducible thermal shift values. These 11 fragments are henceforth referred to as the "final hits".
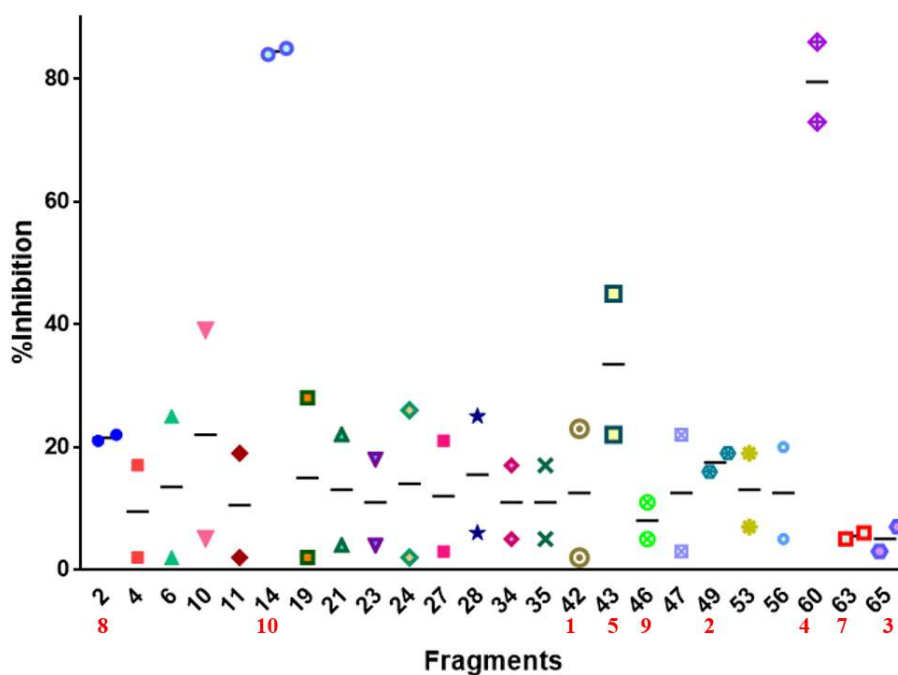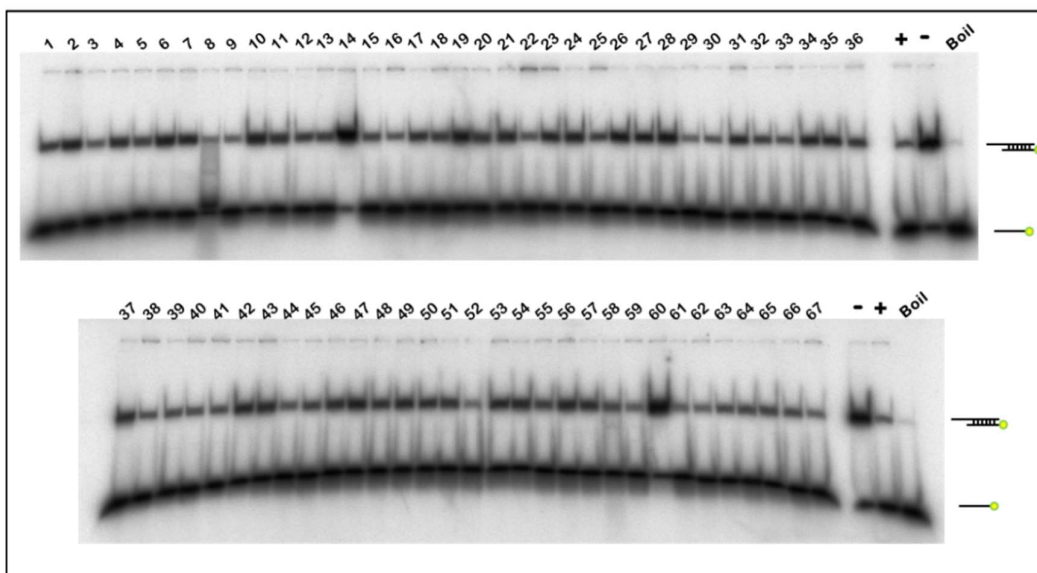
Figure 43. Inhibition of FL-hPif1 helicase activity by selected fragments from the MRCT library. Helicase assay (30 min incubation) for the final 67 selected fragment molecules (Table 8), (0.1 nM substrate, 4 nM FL-hPif1). + indicated protein only (no inhibitor), - indicates native substrate and "Boil" indicates heat denatured substrate (**top**). Graph plotted for the values obtained from the two repeats (only for the fragments that showed a significant level of inhibition) (**bottom**). The top value was obtained from 30 minutes' incubation of the assay and the bottom value is for 45 minutes' incubation. The numbers in red identify the 11 compounds selected for re-supply (refer to Table 8).

3.4.3.3 Re-testing the selected final hits from the fragment screen

All 11 final hits were purchased and first re-tested in the thermal shift assay at 1 mM concentration. As the results in Figure 44 show, fragments 1-10 all induced a ΔTm for hPif1 but the ΔTm values were not as high as originally observed. Curiously, fragment 11, which gave the second highest thermal shift value in the original library screen (Table 8, row 66), was destabilizing the protein (negative ΔTm value). Although fragment 1 and 2 did re-produce a modest thermal shift (Fig. 44) the values were not similar and much diminished compared to those originally observed (Table 8).

All 11 compounds were also tested for inhibition of helicase activity (Fig. 45) in the standard strand displacement assay at a concentration of 1 mM (N=3). IC_{50} values for inhibition of helicase activity of FL-hPif1 and FL-hUPF1 were determined for five selected fragments (3, 4, 5, 7 and 8), as summarised in Figure 45 and Table 9). IC_{50} values for FL-hPif1 inhibition ranged from 0.8 ±0.1 mM (fragment 8) to 3.9 ±0.5 mM (fragment 5). Fragments 3, 4 and 7 demonstrated a high degree of specificity towards FL-Pif1 compared to UPF1, while fragment 8 demonstrated an intermediate level of specificity (~10-fold differences in IC_{50} values). Fragment 5 demonstrated no specificity.
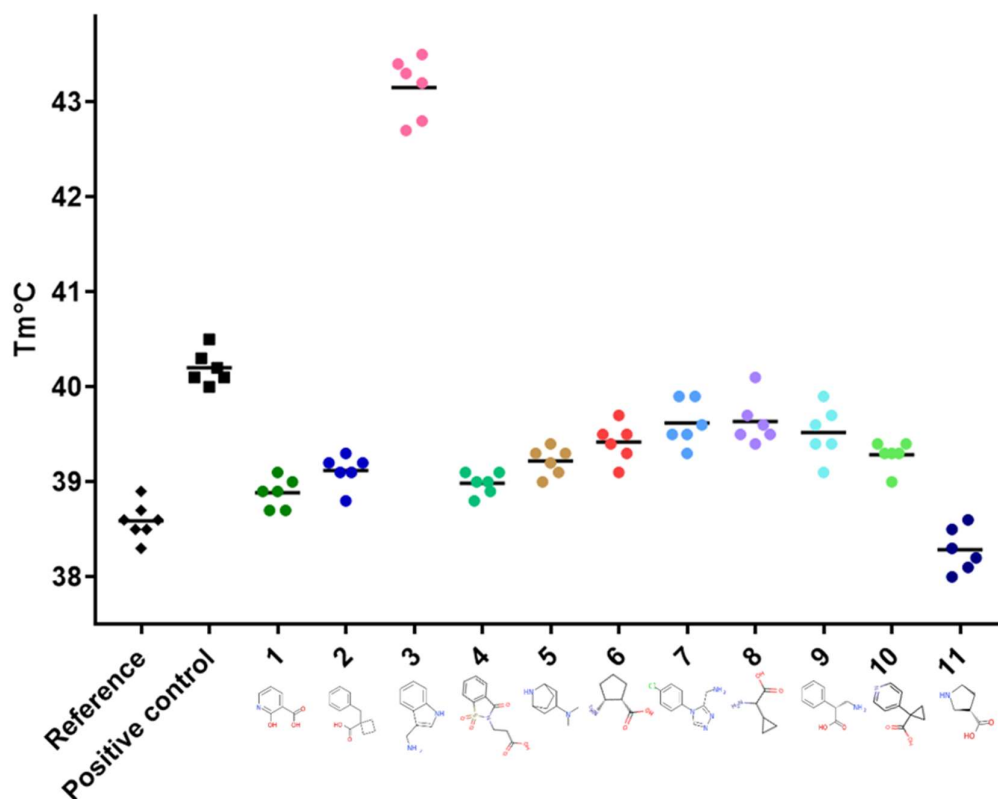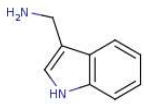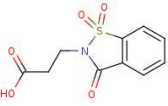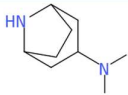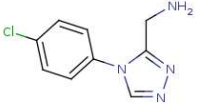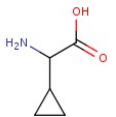
	Mean °C	SEM	N	ΔTm °C
Reference	38.5	0.18	7	0
+AMP-PNP	40.2	0.17	6	1.7
1	38.8	0.16	6	0.3
2	39.1	0.17	6	0.6
3	43.1	0.32	6	4.6
4	39	0.11	6	0.5
5	39.2	0.14	6	0.7
6	39.4	0.2	6	0.9
7	39.6	0.24	6	1.1
8	39.6	0.25	6	1.1
9	39.5	0.27	6	1
10	39.2	0.14	6	0.7
11	38.2	0.23	6	-0.3

Figure 44. Protein thermal shift data for the 11 resupplied fragment molecules. Each dot plotted on the graph represents the melting temperature of FL-hPif1 (0.5 μM) value induced by each fragment molecule (1 mM) in one independent assay (N=6). Horizontal bars are the mean of the melting temperature observed in the presence of each fragment molecule. "Reference" indicates the protein alone and the positive control is shown as +AMP-PNP in the table below (with matched DMSO concentrations). SEM is the standard error of the mean and N is the number of independent repeats of the assay.
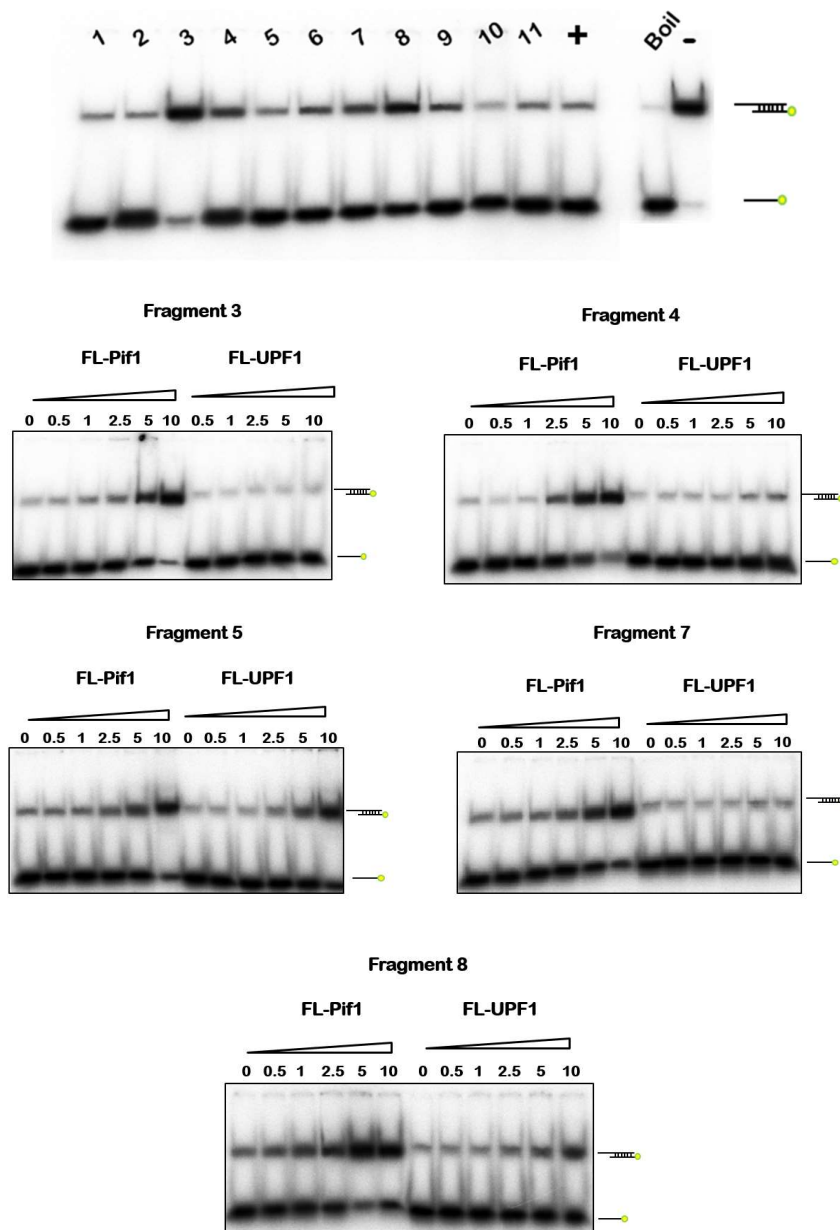
Figure 45. Inhibition of helicase activity by selected fragment molecules. Measuring the helicase inhibition effect of the fragment molecules (final hits, numbers indicated above the gel lanes) at 1 mM concentration (**top**) and examples of measured IC_{50} (0.5 to 10 mM compound as indicated) for the selected molecules (**bottom**). + indicates protein only (no inhibitor, DMSO concentration matched), - indicates native substrate and "Boil" indicates heat denatured substrate. All assays 0.1 nM substrate, 4 nM FL-hPif1, 10 nM FL-UPF1 and 30 minutes' incubation.

Fragment	ΔTm °C (at 1mM)	IC$_{50}$ Pif1 (mM)	IC$_{50}$ UPF1 (mM)
3	4.6 ±0.32°C	3.3 ±1.4	>10
4	0.5 ±0.11°C	3.4 ±0.3	>10
5	0.7 ±0.14°C	4.7 ±0.3	6.2 ±0.4
7	1.1 ±0.24°C	4.5 ±0.2	>10
8	1.1 ±0.25°C	1.1 ±0.1	>10

Table 9. Summary of the results of the fragment library screen. Thermal shift assay and helicase inhibition IC$_{50}$ values measured for each fragment molecule (the final hits) using FL-hPif1 and FL-hUPF1 (N=4). Standard error of the mean values is shown by (±).

3.4.4 XChem fragment screening by crystallography

Automated crystal soaking with ~900 fragment molecules was performed in collaboration with Dr. Ben Bax and Prof. Fred Antson (University of Cardiff and York) through the XChem screening programme at the Diamond Light Source national facility. hPif1$_{206-620}$-AMP-PNP co-crystals were used for this screening since they were more reliably re-producible compared to the apo crystals. As the ATP binding site was occupied, the hit rate of the screening was significantly low and possibly lower than may be expected for the apo-form of the enzyme. However, the outcome of the screening was three hit molecules (hereafter referred to as XChem hits) which showed direct binding to the protein. One compound, EN300-02473, was bound in a pocket between domains 2A and 2B close to the ssDNA binding site (Fig. 46). The resolution of the structure was 1.73Å with good density for the bound ligand, showing well defined interactions between the heterocyclic thiazole ring of EN300-02473 and the protein. Also, the same molecule was bound at another site on the protein surface between domain 1A and the C-terminal helix.

The original XChem hit (1, EN300-02473) and 5 structurally similar compounds with modifications to the carboxyl end (Fig. 47) were purchased and tested in biochemical assays to investigate inhibitory activity. Compounds were prepared as 100 mM stocks in DMSO. Of these compounds 3, 4 and 5 (Fig. 47) had solubility problems in aqueous solutions and also caused precipitation of the protein in the assay. The remaining three molecules and the original (which did stay soluble under assay conditions) were taken to the next step and IC$_{50}$ values for inhibition of helicase activity were measured for both full-length hPif1 and hUPF1 (Fig. 48). These compounds had IC$_{50}$ of 0.08 ±0.02 mM to 0.11 ±0.02 mM and demonstrated a high degree of specificity

for hPif1 compared to hUPF1 *e.g.* minimum 28 fold difference in IC_{50} for compound 2

(hPif1 0.09 ± 0.02 mM compared to UPF1, 3.9 ± 0.4 mM).
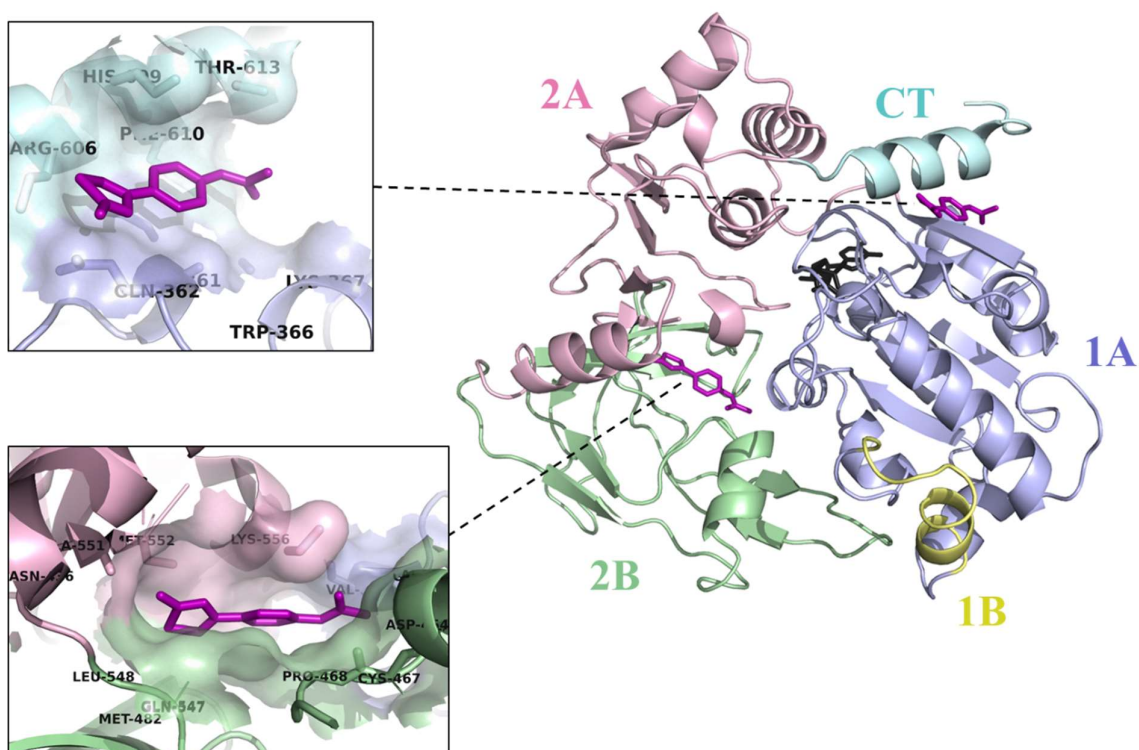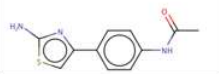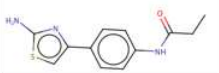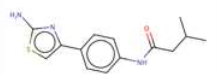
Figure 46. Binding of EN300-02473 to hPif1. Structure of the EN300-02473 fragment molecule (shown in purple) bound and the surrounding residues (created in PyMol). Protein domains are labelled and coloured accordingly in the structure and the AMP-PNP molecule is shown in black.
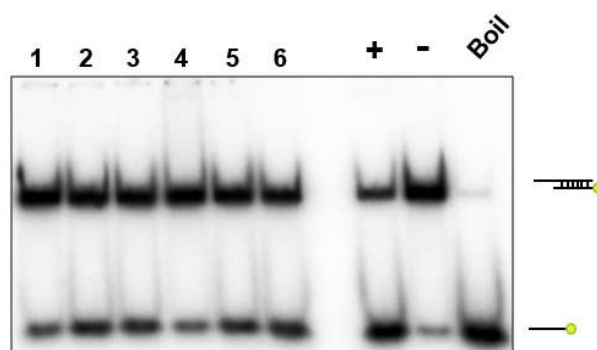
Figure 47. Inhibition of helicase activity by the XChem hit and analogues. DNA strand displacement assay (0.1 nM substrate, 0.4 nM hPif1, 30 minute incubation) in the presence of the XChem hit (1, EN300-02473) and structurally related compounds (2, 3, 4, 5 & 6) at a single concentration of 2.5 mM. + indicates protein only (no inhibitor/DMSO matched), - indicates native substrate and "Boil" indicates heat denatured substrate.

	Structure	IC$_{50}$ hPif1	IC$_{50}$ hUPF1
1		0.11 ±0.02 mM	> 5 mM
2		0.09 ±0.02 mM	3.9 ±0.4 mM
6		0.08 ±0.02 mM	> 5 mM
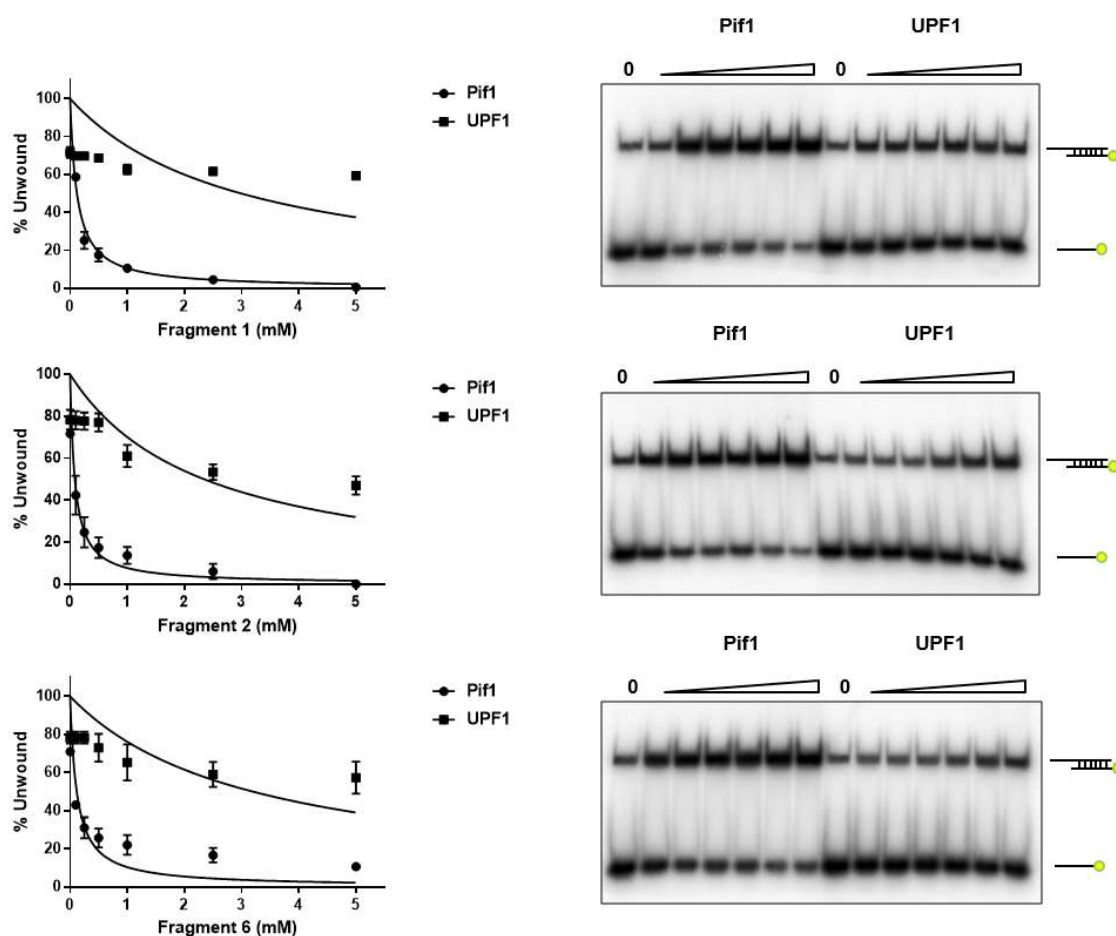
Figure 48. IC$_{50}$ determination for EN300-02473 and selected analogues. IC$_{50}$ values for helicase activity inhibition measured for the selected analogues and the original hit EN300-02473 (fragment 1) using FL-hPif1 and FL-hUPF1. Graphs show the normalized curves plotted for extracting IC$_{50}$ values and the actual values obtained for each titration point are shown with their error bars on the same graph (N=3).

4 Discussion

The important role of helicases in maintaining genome stability has made them attractive targets for treating cancer by causing cell death via the DNA damage response. To date, inhibitors of five human DNA helicases, WRN, BLM, DNA2, DDX3 and MCM have been introduced and are being investigated or developed as drugs with cancer therapy applications (Datta and Brosh, 2018). As discussed in section 1.3, hPif1 is potentially a highly attractive therapy target compared to other helicases for several reasons. First, its biology fits neatly into the concept of synthetic lethality and, in addition, it is not an essential helicase (mouse knock-out mice are normal) (Snow et al., 2007), indicating that anti-hPif1 drugs could avoid some toxicity problems. Success in targeting any protein relies on understanding its basic biological function/s and mechanism/s of action in detail. Here, an investigation of hPif1's structure and function and a drug discovery campaign were initiated. The latter has resulted in the identification of a small-molecule inhibitors of hPif1 and other chemical entities that merit follow-up investigation.

4.1   Production of recombinant hPif1

One of the main objectives of this study was to obtain a high-resolution structure of the hPif1 helicase to gain detail knowledge of its mechanism to inform the structure-based drug design aspect of the project, as discussed further below.   The initial challenge was the production of the hPif1 protein in a sufficient amount for structure and function studies. Since hPif1's identification, many attempts by several laboratory groups to obtain this human helicase in sufficient abundance and purity have failed. Consequently therefore, very little has been added to our knowledge of hPif1's biochemistry in the literature. However, many *in vivo* and *in vitro* studies have been

carried out using the orthologues from *Saccharomyces cerevisiae* and *Bacteroides.* Milestones in these investigations include the recent determination of structures of the *Bs*Pif1 and *Sc*Pif1 helicase domain (Lu et al., 2018, Zhou et al., 2016, Chen et al., 2016). In 2010, hPif1-HD was produced in the Sanders' lab, facilitated by a GST fusion that was subsequently cleaved and removed from preparations (high quality, low/moderate yield) as well as FL-hPif1 (lower quality and lower yield) expressed in *E. coli* BL21(DE3) (Sanders, 2010). These expression constructs formed the basis for initial investigation but proved unsatisfactory in facilitating structural work and biophysics.

In this study the initial focus for X-ray crystallography was the hPif1 helicase domain (residues 206-620) because several protein structure prediction software packages suggested that the N-terminal section (residues 1-205) and a small C-terminal piece (residues 621-641) are relatively un-structured (Appendix, Supplementary Figure 1). Also, the protein modelling software (Phyr^2; http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) failed to predict a model structure for residues 1-205. Importantly, initial studies indicated that helicase core (residues 206-620) retains all DNA binding and processing activities (George et al., 2009). Furthermore, it was known from several lines of evidence that the hPif1 N-terminus is susceptible to proteolysis: (i) Purification of N-terminal residues 1-205 was problematic, even though it can be produced in high yields, it co-purifies with a significant ladder of degradation products when analysed by SDS-PAGE (unpublished observations, Sanders lab). (ii) Degradation species were consistently visualized in purification attempts of the FL-hPif1, but not hPif1-HD, and these species are greatly increased in the presence of the site specific protease thrombin. Significantly, the latter issue was solved by replacing the thrombin site with a TEV protease cleavage site

which considerably reduced the degradation products that were produced during protease digestion to remove the tag.

Purification of the hPif1-HD via the established route (GST fusion, expression in *E. coli* BL21(DE3)) had several problems including; (i) low protein solubility during expression, (ii) low stability of the purified protein at room-temperature and (iii) co-purification with the endogenous chaperone GroEL/ES which significantly reduced the yield of protein. It was therefore proposed to produce the protein at low temperatures using the ArcticExpress™ system to, possibly, improve solubility and stability. In this system the strain of BL21(DE3) employed overexpresses a GroEL/ES chaperone homologue, Cpn10/60, from the arctic bacterium *O. antarctica* (Ferrer et al., 2003). Fortuitously, unlike the native *E. coli* GroEL/ES, Cpn10/60 happened to separate easier from hPif1 during the chromatography purification steps, while the low temperature expression significantly elevated the production of soluble protein. From expression to purification and crystallization of hPif1-HD proteins the temperature was constantly maintained at 4 - 6°C. This decision was made because the experimental observations of the hPif1-HD protein's stability (melting temperature measurements by protein thermal shift) and solubility showed that if kept in the "ideal" buffer conditions at room temperature compared to 4°C the protein readily starts to precipitate as well as the indirect observation that helicase activity reduces precipitously above 25°C.

As discussed above, although the purity and yield of all hPif1 protein constructs did improve significantly using ArcticExpress™, the yield of FL-hPif1 was still insufficient to proceed with protein crystallisation attempts, although high purity was achieved. Therefore, the protein crystallization screens were initiated using the hPif1-HD (206-620) and later a construct extending to the natural C-terminus (residue 641). Clearly, structural information on the intact FL-hPif1 remains a highly desirable goal.

Additional expression trials employing the ArcticExpress™ system should be prioritised as it seems generally conducive for hPif1 expression. However, it was generally observed that the choice of tag can greatly affect the outcome (yield and purity), so others should be trialled.

4.2 Structure determination and analysis

The structural and functional analysis of human hPif1 has proven challenging compared to its yeast and bacterial orthologues, principally on account of the difficulties in producing protein (see above). However, the conservation of biochemical functions (Gu et al., 2008, George et al., 2009, Sanders, 2010) and a role in replication fork progression in intact cells (Gagou et al., 2014) and telomere maintenance (Zhang et al., 2006) indicate a conserved function in genome stability. Pif1 proteins have multiple DNA binding and processing activities including binding to G4 DNA and replication fork-like structures and the catalysis of DNA unwinding and strand annealing. Since hPif1 may be involved in tumourogenesis and could be a cancer therapy target (Gagou et al., 2011, Chisholm et al., 2012) it is important to know which of these activities, if not all, are critical for genome stability. A structure-function analysis could aid the design of separation-of-function mutants that could ultimately assist the dissection of activities in intact cells by a gene-replacement/reverse genetics type approach. In this regard it should be noted that the drug discovery campaign initiated here focused on targeting hPif1 helicase activity. Obtaining a high-resolution structure at 1.1Å resolution with the ATP analogue (AMP-PNP) bound revealed highly detailed information on the ATP binding site and the interactions/ bonds with the residues surrounding the nucleotide (Fig. 20). Even though a low resolution crystal structure of

hPif1-HD was determined by (Zhou et al., 2016), the quality and the overall structure is notably different from the high resolution structure presented in this study (Appendix, Supplementary Table 1.2). The coordinates released for this structure in the Protein Data Bank is the highest resolution helicase structure in the database to date. Such detailed information could facilitate targeting the helicase with nucleotide analogues or chemical entities bound at other identified critical functional sites.

Many attempts were made to co-crystallise hPif1 with various DNA structures without success. Short single-stranded DNA substrates of varying length (8-15 bases) and sequence, self-complimentary partially-single and double-stranded fork-like substrates and both uni- and tetramolecular G4 DNA were employed. Attempts often resulted in diffracting crystals from which structures were obtained but without DNA bound.

One approach was to pre-screen DNA substrate binding by EMSA. Generally, it was observed that for ssDNA the binding affinity increased with length (Fig. 29). Unlike previous observation made in the absence of nucleotide cofactors (George et al 2009), in the presence of ADP•AlF$_4^-$ increasing ssDNA length favoured the formation of additional complexes where formation of one species was observed with a ~10 base substrate and the formation of a second higher order species was observed with a 20 base sequence. Binding to 30 and 35mers suggested three independent binding event and hence a nominal ssDNA binding site size of 10-12 bases. This would be consistent with the length of the ssDNA binding channel observed in *Sc* and *Bs*Pif1 ssDNA co-crystal structures.

When comparing different sequences, a binding preference of hPif1-HD for poly d(C) was observed. No binding to poly d(A) was seen and surprisingly the affinity for poly d(C) was much greater than for poly d(T) or mixed dT-dG-rich sequence (*e.g*

TTTGTGTGTGTT or a telomere repeat, TTTTAGGGTTT; Figs. 26, 27, 30). Thus, the hPif1 ssDNA sequence binding does not match that of *Sc*Pif1 which is reported to be a preference for G-rich ssDNA (Lu et al., 2018). Importantly, the observed preference for d(C)-rich sequences could indicate a role in processing the unpaired DNA that would be present with G4 DNA (opposite complementary strand). This should be investigated in the future. It is also possible that the affinity for dC-rich ssDNA could indicate a binding preference for secondary structures such as DNA helices (Chen and Li, 2007) or the I-motif (Fleming et al., 2017) that can form in poly dC. However, these structures are highly buffer condition dependent and, more importantly, I-motif formation is length-dependent and unlikely to form in the short (~10 base) sequences tested.

The overall structure of hPif1-HD resembles the *Sc* and *Bs*Pif1 orthologous, and has the domain architecture characteristic of the RecD2 family (Singleton et al., 2007). The accurate structural information obtained for hPif1 and the analysis of already available microbial Pif1-ssDNA complexes allowed the generation and testing of a model for hPif1 ssDNA binding (Fig. 21 and Appendix, Supplementary Figure 5). Structures of *Bs*Pif1 demonstrated that a relatively large movement of domain 2B accompanies ssDNA binding. Thus far, although Pif1 proteins bind ssDNA without nucleotide cofactors, no Pif1-ssDNA structures have been obtained in the absence of ADP•AlF$_4^-$. While structures with and without nucleotides bound are almost identical, as are the apo and nucleotide-bound hPif1 structures determined in this study, higher affinity of *Bs*Pif1 towards ssDNA was observed only for the ADP•AlF$_4^-$ transition-state mimic (Zhou et al., 2016). It was observed that ADP•AlF$_4^-$ induces a similar increase in ssDNA binding affinity towards the human protein, hPif1, while in the presence of AMP-PNP (ground-state mimic) or ADP (product) cofactors ssDNA binding affinities

are lower than those observed in the absence of cofactor (Fig. 28). Why structures for hPif1 with ssDNA have not yet been obtained is not clear. However, although all the Pif1 family of proteins bind and process similar DNA substrates, the biochemical parameters of each are significantly different. For example, the hPif1 DNA-dependent ATP*ase* activity determined (Fig. 23) corresponds to a K_{cat} value of 73 ±5.2 min^{-1}, while the value determined for *Bs*Pif1 under similar conditions was 810 min^{-1}. Also, as mentioned above, in contrast to the *Sc*Pif1 (Lu et al., 2018), we observed no preferential binding towards G-rich ssDNA.

It is also possible that hPif1 has a significantly lower affinity for ssDNA compared to its microbial orthologues and this has prevented co-crystallisation. Attempts to measure accurate K_d values independent of EMSA by microscale thermophoresis (MST) proved problematic (C.M. Sanders unpublished), but initial binding isotherms indicate solution K_d values well below apparent values estimated from EMSA, and much lower than for *Bs*Pif1 (50 nM with ADP•AlF$_4^-$, Zhou et al. 2016). It is also possible that the presence of aluminium ions has a negative effect on crystallisation. The hPif1-ADP•AlF$_4^-$ structures obtained, with and without ssDNA, were all low resolution (>3.5 Å). Recently C.M. Sanders has observed that the AlF$_4^-$ ion is not required and high affinity binding can be observed in the absence of AlCl$_3$ but with ADP, Mg^{2+} and NaF. This suggests that trifluoromagnesate (MgF$_3^-$) is approximately isosteric with and can substitute for AlF$_4^-$, as well as being isoelectric with PO$_3^-$ (Jin et al., 2017). Subsequent crystallisation trials should be performed without AlCl$_3$. It is also noteworthy that some of the highest resolution helicase-ssDNA structures have been obtained for the bacterial SF1 family helicase UvrD in the presence of ADP•MgF$_3^-$ (Lee and Yang, 2006).

4.3 Exploitation of structural information for functional analysis

All substitutions of conserved residues in the putative hPif1 ssDNA binding channel resulted in a 5-10 fold reduction in ssDNA binding in the absence of nucleotide cofactors. As expected, variant E307Q, targeting the ATP binding site (Walker B) had near wild-type activity, as did K414 which is conserved in mammalian and *Bs*Pif1 but outside of the ssDNA binding channel and coordinates a sulphate ion in the apo hPif1 structure. With the exception of K485E, which retained ~10% activity, all substitutions in the ssDNA binding site, as well as the Walker B variant E307Q were inactive in helicase assays. All DNA binding defective mutants had decreased DNA-dependent ATP*ase* activity, although reduced ssDNA binding without nucleotide cofactor was not strictly commensurate with reduced ATP*ase*. For example, variants N495A, K485E and N486A retain significant DNA-dependent ATP*ase* compared to others. These residues are all in the conserved B and C motifs (residues 477-486 and residues 494-499 respectively) of mobile domain 2B (Bochman et al., 2010). In the 2B domain of *Bs*Pif1, only N288 (hPif1 N486) and N296 (hPif1 N495) interact with DNA, forming hydrogen bonds with the phosphate backbone. The equivalent residue to hPif1 K485 in *Bs*Pif1 is V287. Lysine at this position is largely conserved in mammalian Pif1 and RecD from *E. coli* and *D. radiodurans*. Mutation and deletion of the short linker region between the *Bs*Pif1 B and C motifs (Chen et al., 2016) significantly impairs DNA unwinding but not ssDNA binding and ATP*ase* activity. *Bs*Pif1 N296A retains 60% DNA binding activity while the corresponding substitution in hPif1 (N495A) retains only ~15%. Taken together, the functional data indicate that the mode of ssDNA binding is substantially conserved in Pif1 proteins. They also indicate that additional residues in the B and C motifs make a more defined contribution to ssDNA binding in the mammalian compared to bacterial Pif1.

The separation wedge, domain 1B, is at the entrance to the ssDNA binding channel. The position of the helix-turn-helix structure of the wedge is believed to be held in place by interactions with the α-helical Pif1 family signature motif (Chen et al., 2016, Zhou et al., 2016). In yeast Pif1 this motif is sensitive to amino acid substitutions and its integrity is essential for ATP*ase* activity (Geronimo et al., 2018, Mohammad et al., 2018). As noted above, the sequence of the wedge is variable and non-conserved. Compared to human and bacterial Pif1, in yeast species the wedge is Lys and Arg rich. The apical residues of the helical turns where strand separation is assumed to occur are ^{87}KFSEYK^{92} in *Bs*Pif1, ^{289}QRPGVR^{294} in hPif1 and ^{323}RRSRKH^{328} in *Sc*Pif1. Although several residues of this segment in *Bs*Pif1 interact with ssDNA, which in the crystal structure is observed to bend sharply as it enters the ssDNA binding channel, alanine substitution of the residues results in only modest reductions in ssDNA binding and unwinding (Zhou et al., 2016). In hPif1, while variants P291A and R294A have near wild-type ssDNA binding and unwinding activities, R290A results in a significant decrease in ssDNA binding (~40% of wild-type), unwinding (~25% of wild-type) and DNA-stimulated ATP hydrolysis (~30% wild-type). The closest corresponding residue in *Bs*Pif1 is K87 and the variant K87A has only a small defect in ssDNA binding but no defect in dsDNA unwinding. These data suggest that functional sites may be optimised differently in Pif1 proteins, so caution should be exercised in extrapolating directly from microbial to human Pif1.

As noted above, the biological function and mechanism of the DNA strand-annealing activity of Pif1 proteins are not fully understood. A DNA strand-annealing activity resides in helicase core (George et al., 2009), but the N-terminal (~ 200 residues) may also have an independent annealing function (Gu et al., 2008). The annealing activity of hPif1 variant proteins was assayed here under stringent conditions

(135 mM NaCl) without nucleotide cofactors. Unexpectedly, all mutants other than F573A and N495A displayed annealing activities equivalent to or greater than wild-type. N495 (B motif) and the equivalent residue in *Bs*Pif1 is on the outer surface of the DNA-free Pif1, where it coordinates a sulphate ion in our apo hPif1 structure, and moves ~16Å upon binding ADP•AlF$_4^-$ and ssDNA. F573 (*Bs*Pif1 F379) is at the exit of the ssDNA binding channel and binds the terminal 5′ residue of ssDNA. Although the nucleotide-free ssDNA bound structure of Pif1 is not known, without cofactors hPif1-ssDNA binding was only observed with poly-T oligonucleotides >30 bases. Direct and indirect observations indicate that monomeric SF1 and SF2 (Eoff and Raney, 2006, Buttner et al., 2007) and oligomeric helicase (*e.g.* papillomavirus E1 (Chaban et al., 2015)) as well as binding to the translocating (active) strand they may also engage the displaced "passive" ssDNA strand along a distinct binding path. *Sc*Pif1 can interact simultaneously with two ssDNA molecules (Ramanagoudr-Bhojappa et al., 2014), suggesting that at least two low affinity binding sites exist that could allow interactions with long oligonucleotides on the protein surface or facilitate complementary strand annealing. Residue N495 could define a second ssDNA binding site that can shuttle between the active ssDNA binding channel and the outer surface in an ATP-dependent manner, to chaperone the displaced strand or mediate DNA strand-annealing.

The structural basis of G4 DNA binding in Pif1 proteins is not known. The recent structure of the G4 DNA binding and unwinding SF2 helicase DHX36 (RHAU/G4R1) bound to a G4 DNA substrate with a 3′ ssDNA tail (Chen et al., 2018) shows the G4 DNA engaged by a specific α-helical motif at the entrance to ssDNA binding channel where ssDNA is bound. Based on indirect observations, it has been proposed that in *Sc*Pif1 G4 DNA is clamped at the entrance to the ssDNA binding in "pliers" composed of two sets of positively charged residues (Lu et al., 2018). However, one arm of the

plier in domain 2C (residues R594 and K595) is completely absent in hPif1 (and *Bs*Pif1) while the second is the highly positive segment of the wedge described above. Of the three residues implicated in *Sc*Pif1 G4 DNA binding (K321, R324 and R326), only R324 is conserved in an analogous position in hPif1 residue R290. The analysis of hPif1 G4 DNA (Fig. 25) binding demonstrated only small defects (minimum ~70% of wild-type, variant H269A in the ssDNA binding channel) in G4 DNA binding for all wedge domain and ssDNA binding channel mutants. As such these data do not identify a G4-specific DNA binding segment in hPif1 but indicate that the G4 DNA binding mechanism is not conserved. Again, caution is therefore needed in extrapolating from yeast and bacterial to human Pif1. Curiously, the mutant K414A demonstrated increased G4 DNA binding at low protein concentrations. The reasons for this are unclear. However, they are probably indirect and more likely related to non-specific effects such as, for example, protein stability or domain flexibility.

In conclusion, mutational data indicate that the mechanism of DNA binding and unwinding of human Pif1 are similar to bacterial and yeast Pif1 proteins. However, sequence variability in functional motifs (*e.g.* the wedge region), the presence of additional domains (*e.g.* the 2C domain insertion in *Sc*Pif1) and the biochemical analysis of variant hPif1 forms indicate that extrapolating directly from the analysis of yeast and bacterial Pif1 proteins to human is unreliable.

4.4   Unimolecular G4 DNA processing by hPif1

As described previously in section 1.2.1, several helicases involved in the DNA damage response are known to resolve G4 DNA structures *in vitro*. In fact, the ability of a helicase to resolve G4 structures is not a particularly uncommon property, but G4 structure-specific binding is, as is the case with Pif1. The role of yeast Pif1 in processing G4 DNA has been well studied *in vivo* and *in vitro*. The analysis confirms the importance of *Sc*Pif1 in maintaining genome stability where DNA breakage, deletion and recombination are significantly elevated in the G4 DNA forming sequences in the Pif1 depleted genome (Ribeyre et al., 2009, Paeschke et al., 2011). Because of the complications in creating a homogeneous species of unimolecular G4 DNA and the difficulty in visualization of unimolecular G4 DNA unwinding events by gel electrophoresis, most *in vitro* experiments have employed synthetic tetramolecular G4 DNA substrates (Sanders, 2010, Wu et al., 2008). Although useful, such substrates usually also have four corresponding ssDNA segments that are required for helicase engagement and can complicate the analysis. Alternatively, single molecule unwinding events using tethered substrates have been probed in FRET based assays (Zhou et al., 2014).

As shown in Figure 32, the unimolecular G4 substrate with a 3′ duplex employed in helicase assays with FL-hPif1 shows a mobility shift (same mobility as the mutant sequence that doesn't form a G4 DNA structure) that was dependent on the increase in hPif1 concentration. However, the short duplex 3′ of the unimolecular G4 DNA forming sequence was barely unwound by the helicase (less than 10% displacement at the highest protein concentration tested, Fig. 32). This observation was intriguing, since unimolecular G4 DNA unwinding has not previously been observed by gel electrophoresis and encouraged further investigation. To confirm the unwinding of the

unimolecular G4 DNA by hPif1 a DMS modification and footprinting assay was coupled to the helicase reaction. Since N7 of guanine is protected from DMS modification and subsequent cleavage by piperidine in G4 DNA, its unwinding should result in enhanced reactivity. As the results show in Figure 33, G-specific DNA cleavage over the G4 DNA forming sequence of the substrate increased with increasing hPif1 concentration. Together with the electrophoresis based unwinding assay performed in parallel, this finding supports the hypothesis that the helicase resolves the G4 DNA structure but does not continue to unwind the following duplex. However, since the boiled substrate (Fig. 33 left) does not show a mobility shift, it was rather surprising to observe that the G4 DNA resolved in the presence of hPif1 does not re-form the secondary structure, even after denaturing the protein with SDS-containing stop buffer. Note, the DMS modification assay confirms that the G4 DNA structure is dissociated by heat denaturation (Fig. 33).

The observation discussed above are intriguing, but suggests that the helicase may cause some other chemical or physical modifications to the unwound G4 substrate to maintain this state. However, this is an initial hypothesis that can be tested in the first instance by comparing, in the same system, hPif1 with other helicases that are known to resolve G4 DNA. If substantiated as an assay measuring a *bona fide* hPif1-dependent unimolecular unwinding event the electrophoresis-based assay could be employed as a simple tool to screen G4 DNA unwinding mutants with a physiologically relevant substrate.

4.5 Screening of the MRCT diversity set library

The initial approach to screen the MRCT library for drug-like molecular inhibitors of hPif1 was based on screening for inhibition of helicase activity. By immediately counter-screening "hits" against the bovine papillomavirus E1 helicase the majority of potential non-specific inhibitors such as DNA binding compounds and nucleotide mimics should have been eliminated. IC_{50} values were measured for the remaining compounds and for four (H2, E3, G3 and E4, Table 5) that had an IC_{50} <10 µM. 78 related compounds (including the original four) were resupplied by the MRCT for further investigation (note: structural information was not provided, but the choices made by the MRCT were based on a structure-activity relationship (SAR) and knowledge of non-specific effects in the library). Of these, 23 showed inhibition when analysed at 100 and 25 µM (Fig. 37) and were selected for IC_{50} (helicase assay) determination and counter-screening against the human SF1 helicase hUPF1 (Table 6, columns 1-6). Most compounds had low IC_{50} values as expected, including the original four that produced a similar IC_{50} value to that determined from the initial library material (less than 2-fold difference, Table 6). However, many compounds also inhibited hUPF1 that was introduced as a counter-screen at this point, except A1, B1, E4, D7, E8 and C9. A1 and D7 were the weakest hPif1 inhibitors among that group (IC_{50} close to 100 µM) and they also did not inhibit strand annealing activity nor did they produce a thermal shift, so these could simply be weak inhibitors.

After reporting these results, the MRCT selected and resupplied, as dry powder stocks, eight compounds for further analysis (B1, G2, A4, C4, A5, C5, G6 and D9, highlighted in red in Table 6). Initially, IC_{50} values of these compounds were re-measured twice to confirm their reproducibility. All but two compounds (G2 and D9) produced almost the exact same IC_{50} values and had similar effects on strand annealing

activity (Table 6, columns 7-10) as the original library material. Failure of G2 and D9 to reproduce could be due to incorrect labelling / re-supply of the compound or product quality control. Careful determination of IC_{50} values showed that four of the other compounds (A4, C4, C5 and G6) were near equivalent or more potent inhibitors of hUPF1 compared to hPif1 (Table 6), and no thermal shifts were observed. Although it is possible that these molecules are genuine helicase inhibitors, significant effort including co-crystal structure determination would be required to engineer specificity.

Two compounds remain of interest from the MRCT library screen, B1 and A5 in Table 6. B1 is interesting because other than having a relatively low IC_{50} it shows no inhibition of hUPF1. Although, no stabilization of hPif1 could be detected with the PTS assay (Table 6, column 6 & 10), this should not deter follow up as the PTS fluorescent curves were uninterpretable indicating interfering fluorescence signals. Indeed, this is one of the problems associated with the PTS assay which is prone to false positives and negatives (Davis and Erlanson, 2013). It is worth investigating the binding of these molecules by ITC or NMR techniques (Leavitt and Freire, 2001, Klein et al., 1999, Zega, 2017). Compound 5A was the only one which was an inhibitor of the hPif1 helicase activity and a reproducible protein stabilizer in the PTS assay. However, it does not show a significant level of specificity for hPif1. 27 analogues of this compound were tested in thermal shift and biochemical assays. Even though all were structurally similar to A5, only two compounds generated a thermal shift but they failed to inhibit the helicase activity of hPif1. The main structural difference between compound A5 and the two analogues is the methyl group on the bicyclic portion of the molecule. In principle, this and other differences could cause the slight variation of thermal shift values observed (A5 > analogue 1 > analogue 2). The low IC_{50} value of analogues 1 and 2 also suggests that the inhibitory effect on hPif1 helicase activity is

potentially dependent on the methyl group. Although these candidate compounds have reproduced their biochemical and thermal shift effect, it is worth confirming direct binding by NMR or ITC before work proceeds with these candidate inhibitors.

The main limitation of the compound library screening was that there were no structural data made available for us to make decisions on which compounds to pick for further investigations. Second, the amount of the initial supply material was not enough to do multiple experiments (*e.g.* establishing accurate dilution series for IC_{50} determination) during initial selection (all the initial data presented in Table 6 were generated from ~10 µl of 10 mM of each chemical). The fact that the initial selection was based on a biochemical assay and not on the detection of physical binders could also be a limitation of the approach. Many of the inhibition events observed could be induced by factors other than the direct binding of the compounds to hPif1. Also, re-screening the entire library could have helped improve the selection of initial hits. The experience gained with the subsequent screening of a fragment library by PTS, to be discussed later, proved PTS screening to be a more reliable approach since the majority of binders that were selected after initial screening by PTS inhibited helicase activity.

It should be mentioned that four compounds with IC_{50} values less than 10 µM (selected from the initial screening) were tested in cell based assays for their effect on HCT116 cancer cells which originally demonstrated sensitivity to hPif1 siRNA knock-down (Gagou et al., 2011, Gagou et al., 2014). None of these compounds reproduced the effects of siRNA mediated hPif1 knock-down, *i.e.* we did not observe a specific apoptotic cell death response (increase in sub-G1 cells) in the HCT116 cells compared to the HEK293 non-tumour cell control (Appendix, Supplementary Figure 6). Compound MRT00044704 (E4) had no apparent cell cycle effects on HCT116 or HEK293 at 50 or 100 µM. Similarly, MRT00203960 (E3) had no effect on HCT116

but there was a modest effect on the cell cycle distribution of the HEK293 cells (slight increase in sub-G1, decrease in G2/M cell populations). However, the biochemical counter screen indicated that this compound may inhibit related helicases or other ATP*ases*. Compound MRT00153955 (G3) (50 µM) had a weak but discernible effect on the cell cycle (slight increase in the sub-G1 cell population) but this was not specific to tumour cells. Compound MRT00105813 (H2) was highly toxic to both HCT116 and HEK293 even at the lowest dose tested (5 µM) and therefore does not discriminate between normal and tumour cells. This compound also strongly inhibited the related helicase protein hUPF1 used in the biochemical counter screen. MRT00105813 (H2) and MRT00044704 (E4) demonstrated insolubility (crystalline deposits) at all concentrations tested, illustrating that even when biochemical inhibitors are identified further barriers to complete evaluation will need to be overcome.

4.6   Fragment library screening

Fragments are low molecular weight compounds usually 100-250 Da, with generally low binding affinities (from high micro-molar to millimolar). Often, they are derived from known drug-like molecules with already characterised pharmaceutical properties. Sometimes they are referred to as corresponding to the "Rule of Three", having a molecular weight < 300 Da; clogP $\leq$ 3 and $\leq$ 3 hydrogen bond donors or acceptors (Congreve et al., 2003, Rees et al., 2004). Fragment based drug designing (FBDD) is part of structure-based drug design which, ideally, requires a high-resolution crystal structure of the target protein with the potential binding molecule (Lamoree and Hubbard, 2017).

This approach initially requires a screening of small chemical fragments starting with a biophysical assay such as PTS and subsequently confirming the hits with a more

precise techniques such as NMR, ITC or X-ray crystallography, which can also determine the exact binding site of the ligand (Mashalidis et al., 2013). Sometimes an inhibitory effect on the protein's activity *in vitro* is also observable. Ideally, with knowledge of the fragment binding site at sufficient resolution (ideally < 2Å) it is possible to make a more potent binder (higher affinity and specificity). This is done by modifying the initial fragment *in silico,* according to the chemical space, by addition of chemical groups (fragment growing) or by adding an appropriate linker to two or more semi-potent fragments with identified, proximal, binding sites (fragment linking) (Hung et al., 2009, Sledz et al., 2010) (Fig. 49A). However, most FBDD attempts use the fragment growing approach since it is more straightforward and requires only one starting point. An example of this kind of drug design was the development of a drug-like inhibitor of cyclin dependent kinase 2 (CDK2), N-(4-piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide (AT7519), which was subsequently tested in phase I clinical trial for cancer therapy (Wyatt et al., 2008). This compound was developed in 6 steps from a small and low-affinity fragment (Indazole) with an IC_{50} of 185 μM. The final molecule has an IC_{50} of 0.047 μM and has multiple interactions with the surrounding residues (Fig. 49B).

(A)

Fragment growing

Fragment linking

(B)

Fragment-Based Drug Design

$IC_{50} = 185\ \mu M$ $IC_{50} = 0.047\ \mu M$
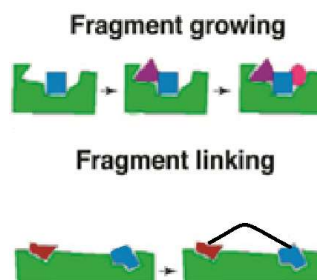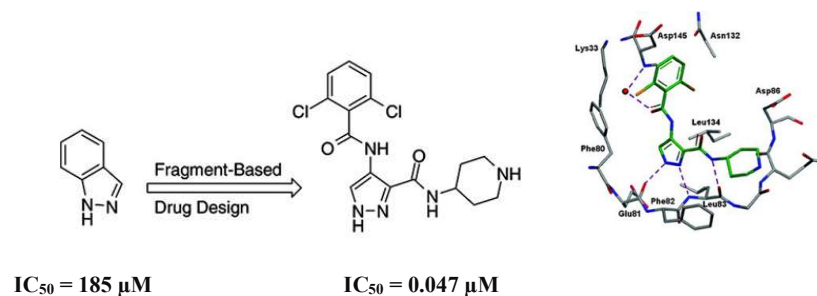
Figure 49. **(A)** Schematic figure illustrating the basic concepts of fragment growing and fragment linking adapted from (Jubb et al., 2012). **(B)** An example of FBDD using the fragment growing approach of a CDK2 inhibitor showing the small starting fragment (Indazole) and the final developed drug molecule (N-(4-Piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H-pyrazole-3-carboxamide) and the crystal structure of the final molecule (shown in green) interacting with the surrounding residues of the CDK2ATP binding site. Adapted from (Wyatt et al., 2008).

The MRCT fragment library was screened multiple times and at different ligand concentrations by PTS assay to identify potential hPif1-HD binding molecules as discussed in section 3.4.3.1 and Table 8. Usually, it is suggested to use a 10 mM concentration of fragments and selection should be based on a minimum of 0.5 °C shift compare to the base control (protein only) (Silvestre et al., 2013). Here, the screening was initilized at a compound concentration of 1.2 mM (the supplied stock concentrations were 30 mM imposing a limitation) and then narrowed to 200μM with multiple rounds of screening. The general observation of these hits showed that the majority of the molecules are acidic which can be ratonalized since the protein is a nucleic acid binding protein and overall has a positive surface charge. This general specificity was also observed on the hits from the 1200 molecule fragment library screened at the University of York in conjunction with the laboratory of Prof. R. Hubbard (Appendix, Suplementary Table 3). As shown in Table 8, some initial hPif1-HD hits that consistently produced a thermal shift failed to stabilise FL-hPif1 (Table 8 highlighted in gray). These molecules may bind to sites on hPif1-HD that are exposed due to the removal of the N-terminal piece of the protein (1-205), but cannot bind in the context of FL-hPif1 if the N-terminal residues share the interaction surface. The biochemical assay (helicase inhibition) on the selected 67 molecules (Fig. 43) was carried on to help narrow down the selection of the fragments for re-supply. However, the decision was not based solely on the results of biochemical inhibition but mostly focused on the magnitude of the PTS result. None the less, there are still a few interesting molecules with high PTS values (highlighted in green, Table 8) which could not be obtained at the time but are marked to be prioritised in future work.

Eleven molecules were obtained based on their thermal shift values, inhibitory effect and also cost and availability from suppliers. These fragments were re-tested by

PTS assay multiple times using FL-Pif1 to confirm the results of the primary screen. Even though the secondary PTS values were slightly lower than the values obtained at initial screening, all but two of the molecules produced a thermal shift (fragments 1 and 11). The generally lower PTS values could be due to concentration variations between the library of compounds and those accurately reconstituted from re-supply. Failure of 1 and 11 to reproduce could potentially be due to mislabeling, or poor quality control of the supply. All eleven of the selected hits were tested in biochemical assay to identify any inhibitory effect on hPif1 helicase activity. Although it is not necessarily expected to see activity for such small molecules, inhibition was observed and IC$_{50}$ values were determined for both hPif1 and hUPF1 (Fig. 45 and Table 9). The IC$_{50}$ values were in the milli-molar range, which is expected for small fragment molecules (Silvestre et al., 2013), with five of the fragments (3, 4, 5, 7, 8) showing inhibition of hPif1 and four of them specificity for hPif1, *i.e.* minimal inhibition of hUPF1 (Fragments 3, 4, 7, 8) (Table 9). Fragment 8 had the lowest IC$_{50}$ (1.1 ±0.1 mM) for hPif1 and more that 10-fold higher IC$_{50}$ for hUPF1. Fragment 3,4,7 and 8, and possibly others from the initial screening not yet re-evaluated, could be good starting points for FBDD. This approach should start with identifying the exact binding site/s by saturation transfer difference (STD)NMR spectroscopy (Klein et al., 1999) or X-ray crystalography by co-crystalizing of protein and fragment/s.

As described in section 3.4.4, after the XChem automatic fragment library screening (with Dr. Ben Bax University of York and Cardiff), structures for three molecules bound to hPif1-HD were determined. Only one, EN300-02473, is described here since the other two are surface-bound and are not considered to be potential inhibitors of hPif1's catalytic activities. EN300-02473 was found to bind at two different sites on the protein (Fig. 46). One site was particulary interesting because the

molecule was bridging domains 2B and 2A and inteacting with residues from both (Fig. 50). This site is very close to the conserved DNA binding site of the helicase and where the structure analysis shows that domain 2B undergoes a major movement in the presence of ssDNA (Appendix, Suplementary Figure 5). Also as described in mutagenesis analysis of hPif1, the mutation in residue K556 had a significant effect on the DNA binding, unwinding and ATP*ase* activity of the helicase. Thus, there is a rationale behind the inhibitory effect observed.
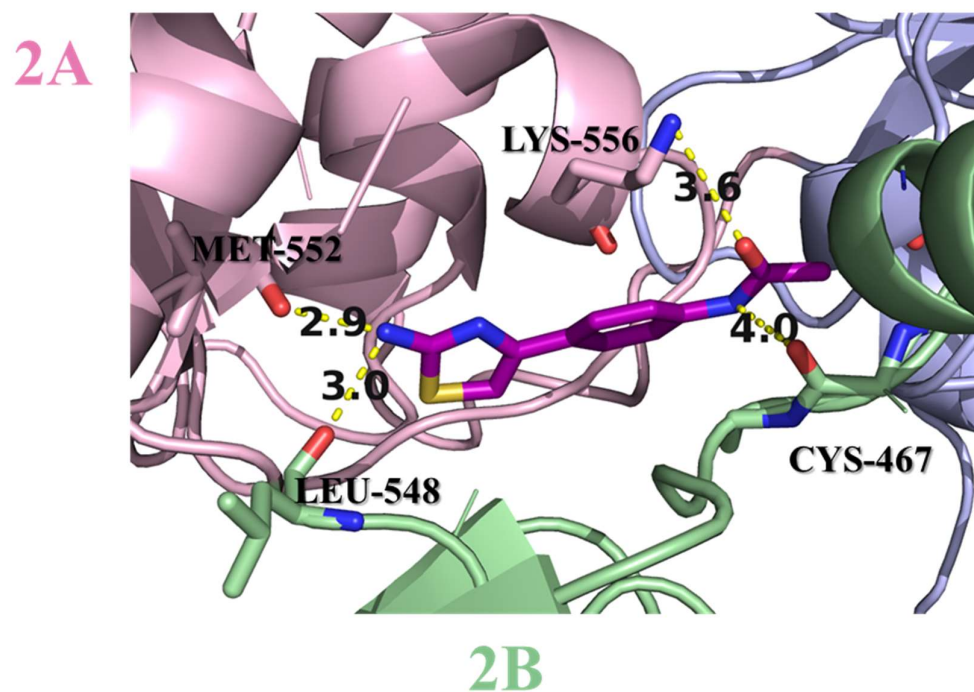
Figure 50. Interactions of EN300-02473 with residues from domains 2A and 2B of hPif1. Interacting residues are shown in sticks and labelled. Distance between the interacting atoms (Å) are shown with yellow lines.

The analysis of the hPif1 structure with EN300-02473 bound suggests that this molecule is potentially locking the protein in a specific conformation and perhaps preventing the natrual movement of these two domains, so disrupting the helicase activity. This type of inhibitory effect was also discovered for a HCV virus protein (NS3) inhibitor molecule for which a detailed study has shown that the inhibitor binds in a pocket between the helicase and the protease domain of NS3. The inhibitor interacts with residues from both protein domains and disrupts the function of the protein (Saalau-Bethell et al., 2012). Drug discovery usually tends to target an enzyme's catalytic sites, very often a nucleotide binding site but also other substrate interaction sites. Although nucelotide binding sites may be a good target for initiating drug design as they are naturaly configured to bind ligands, their structural conservation among many protein/enzyme groups means it is harder to gain specificity. The less the specificity the more likely there will be undesired effects of the drug.

EN300-02473 was found bound at a second site. However it doesn't seem to be an interesting site because it is on the surface of the protein in a structural domain and not close to the any of the protein's conserved catalytic or substrate binding sites. Analysis of the structure showed only a few weak interactions between the molecule and surrounding residue. To make a robust statement on the validity of persuing the candidate inhibitor for further modification mutants that disrupt one or other of the EN300-02473 binding sites could be generated and tested for activity (*i.e.* DNA binding, unwinding, annealing, G4 binding ect.) with and without EN300-02473. Ideally, modulation of inhibitor binding should also be investigated by NMR or X-ray crystallography.

Five analogues of EN300-02473 were obtained and tested in the helicase inhibition assay (Fig. 47). These analogues were selected to have modifications on the

carboxyl end since the amino end is shown to have the strongest interactions between M552 from 2A domain and L548 of the 2B domain (Fig. 50). Three of the analogues had solubility problems or caused protein precipitation when added to the helicase reaction mix. The IC_{50} values (helicase assay) of the remaining three, including the original, were carefully measured for both FL-hPif1 and FL-hUPF1 helicases. Remarkably, all three showed specificity for inhibition of hPif1 with IC_{50} values of less than or equal to 100 μM (Fig. 48). Analogue number 2 produced the lowest IC_{50} (90 μM), it showed more than a ~40-fold lower inhibitory effect on hUPF1, while the other two had at least a 50-fold lower effect on hUPF1. These results suggest that the EN300-02473 binding site is potentially specific to hPif1 and can be developed into a more potent inhibitor/drug with higher affinity/ lower IC_{50}.

4.7 Conclusions

hPif1 has been identified as a potential cancer therapy target, based on the observation that a subset of human cancer cell lines are sensitive to siRNA-dpendent Pif1 depletion and die by apoptosis, while control non-tumour cells are not. The effect is exacerbated by DNA damaging agents such as the cancer chemotherapy drug gemcitabine and thymidine depletion, indicating that the effect is at the level of replication stress. Here, a discovery campaign was initiated to identify small molecule inhibitors of hPif1. Initially these would act as tool compounds to verify the results of genetic knock-down assay and then, potentially, progress to drug-like molecules for testing in animal models.

High-resolution crystal structures of the hPif1 helicase domain were obtained, including apo and an AMP-PNP bound form. Functional residues for DNA binding

were identified in the protein and tested for functionality by site directed mutagenesis and the analysis of variant forms by biochemical assays. The data reveal a conserved ssDNA binding channel in hPif1 that is critical for ssDNA binding during unwinding but not G4 DNA binding. Also, while the ssDNA binding channel is important for helicase activity, the data suggest that it is not used for strand annealing. Accurate structural information and the anlaysis of variants described here will facilitate an in depth *in vitro* dissection of hPif1 activities for subsequent testing in intact cells. Several small molecule inhibitors of hPif1 helicase activity were identified that have the potential to progress to effective tool compounds and perhaps drug-like molecules.

5 Appendix
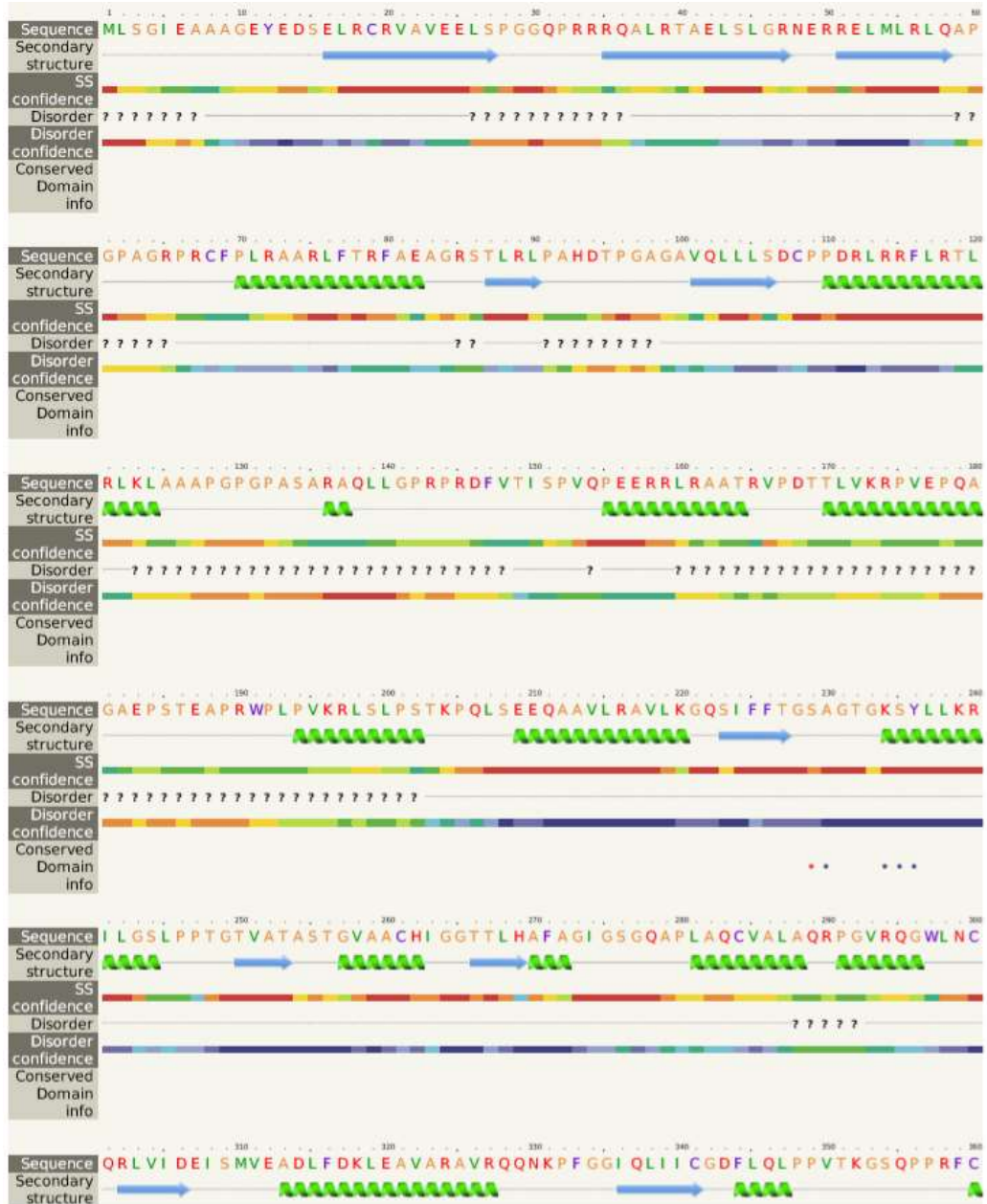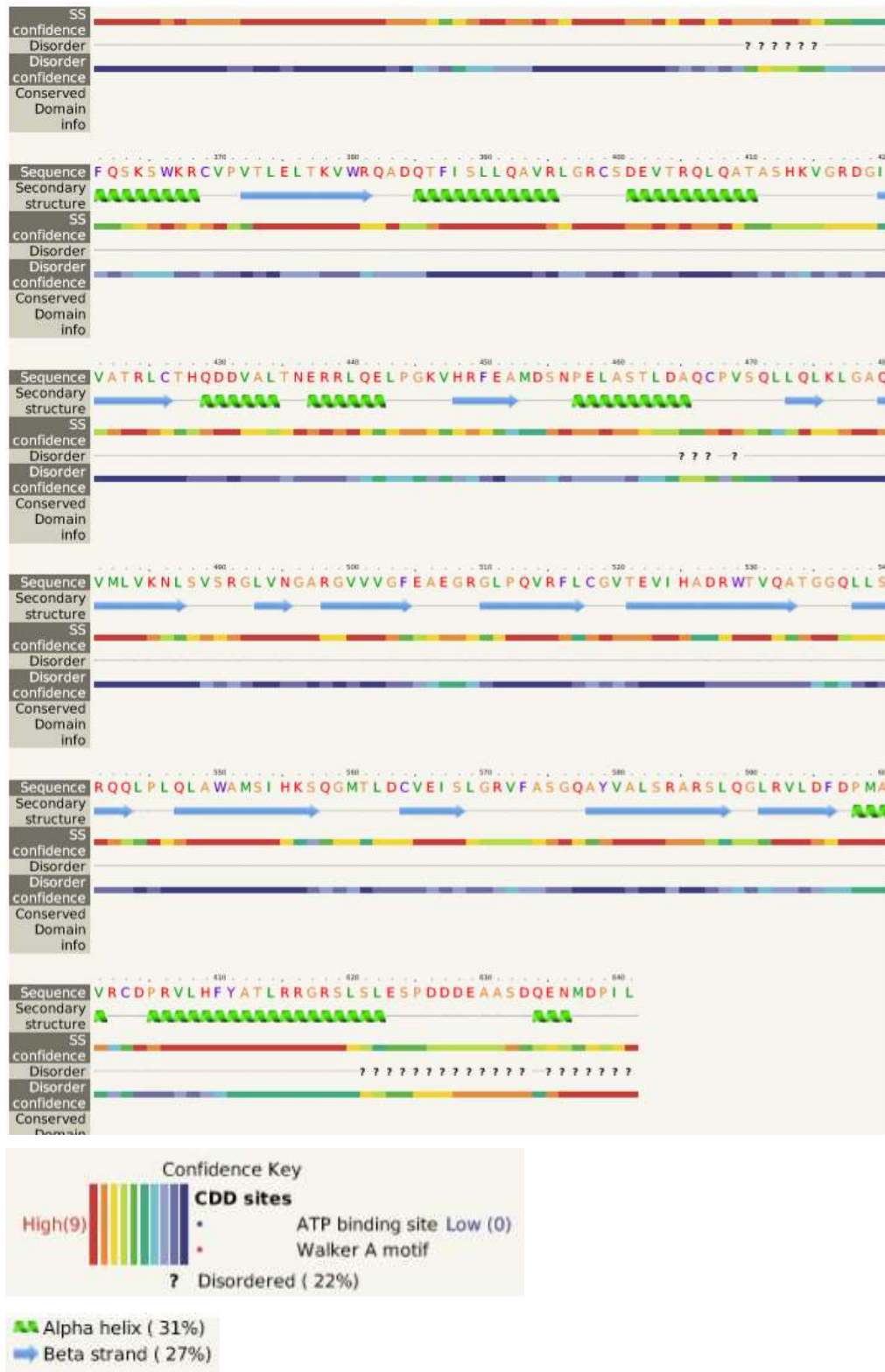
Supplementary Figure 1. Structure prediction of human Pif1 helicase created by Phyre 2 online software.
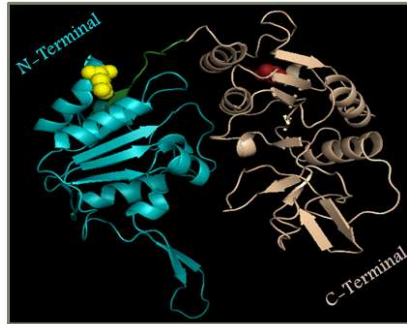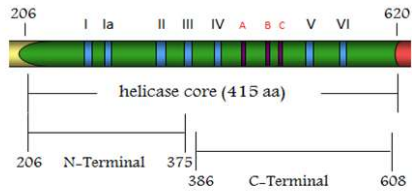
Supplementary Figure 2. Predicted model of hPif1 helicase domain (206-608) structure by Phyre2 online software and viewed by PyMol. The hPif1-HD N-terminal (206-375) segment shown in cyan and C-terminal (386-608), shown in wheat. The end residues are indicated in Yellow (N-terminal) and Red (C-terminal).
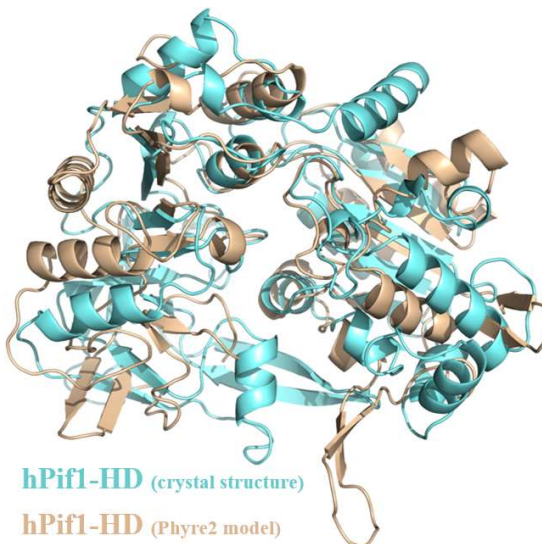
Supplementary Figure 3. Structure alignment of the predicted secondary structure of hPif1-HD (206-620) by Phyre2 software (wheat) and the actual 1.4 Å structure of apo hPif1-HD (cyan) viewed by PyMol software.

	hPif1_AMP-PNP	hPif1_Br			hPif1_Apo	hPif1_ADF
Ligand	AMP-PNP	Br, AMP-PNP			SO_4^{2-}	ADP-AlF4
Sequence fragment	206-620	206-620			206-641	206-620

Data collection

	hPif1_AMP-PNP	hPif1_Br			hPif1_Apo	hPif1_ADF
X-ray source	I02, DLS	I03, DLS			I04, DLS	I04, DLS
		peak	inflection	remote		
Wavelength (Å)	0.97949	0.91966	0.91992	0.91912	0.92819	0.92819
Resolution range (Å)	49.95 – 1.13	73.25 – 1.43	76.98 – 1.39	76.94 – 1.87	81.40 – 1.44	181.74 – 3.96
Space group	C222₁	P2₁2₁2₁	P2₁2₁2₁	P2₁2₁2₁	P2₁2₁2₁	P3₁21
Unit-cell parameters (Å)	$a = 73.3$, $b = 143.3$, $c = 77.6$ α= β = γ =90°	$a = 73.3$, $b = 76.8$, $c = 142.5$ α= β = γ =90°	$a = 73.2$, $b = 77.0$, $c = 142.5$ α= β = γ =90°	$a = 73.2$, $b = 76.9$, $c = 142.1$ α= β = γ =90°	$a = 63.5$, $b = 81.4$, $c = 91.5$ α= β = γ =90°	$a = b = 209.9$, $c = 78.9$ α= β = 90°, γ = 120°
Number of unique reflections, overall/outer shell[a]	149927 / 9913	148682 / 10867	161273 / 11722	67095 / 4888	86350 / 6315	17633 / 1273
Completeness (%), overall/outer shell[a]	98.7 / 89.63	99.9 / 99.8	99.5 / 98.9	100.0 / 100.0	99.9 / 99.9	100.0 / 100.0
Redundancy, overall/outer shell[a]	6.3 / 4.0	12.8 / 10.7	12.4 / 8.3	12.8 / 13.8	6.6 / 6.8	12.3 / 12.8
$I/\sigma(I)$, overall/outer shell[a]	11.2 / 1.4	9.6 / 1.0	10.5 / 1.0	5.2 / 1.0	15.9 / 1.3	17.1 / 2.9
R_{merge}[b] (%), overall/outer shell[a]	6.8 / 84.1	14.6 / 231.6	12.5 / 185.1	37.0 / 259.5	4.5 / 132.9	8.0 / 85.7
$CC(1/2)$ (%), overall/outer shell[a]	99.7 / 89.6	99.9 / 59.8	99.9 / 55.2	98.5 / 68.2	99.9 / 51.4	100.0 / 89.2

Refinement and model statistics

	hPif1_AMP-PNP	hPif1_Br			hPif1_Apo	hPif1_ADF
Resolution range (Å)	49.95 – 1.13	71.27 – 1.43			60.89 - 1.44	91.03 – 3.96
R-factor[c] (R-free[d]) (%)	12.9 (16.3)	23.3 (27.7)			13.7 (19.5)	17.9 (25.3)
Reflections (working/free)	142572 / 7307	141225 / 7357			81905 / 4365	16607 / 794
Outer shell R-factor[c] (R-free[d])) (%)	30.7 (32.5)	48.5 (51.9)			30.7 (33.9)	33.3 (39.5)
Outer shell reflections (working/free)[g]	9436 / 464	10330 / 532			5981 / 331	1134 / 85
Molecules/asymmetric unit	1	2			1	2
Number of protein non hydrogen atoms	3424	6519			3218	6464
Number of water and small molecules atoms	1754	1260			1002	66
R.m.s. deviation from target[f]						
Bond lengths (Å)	0.029	0.025			0.029	0.010
Bond angles (°)	2.496	2.236			2.452	1.558
Average B-factor (Å²)	21.1	20.2			34.9	151.7
Ramachandran plot[g] (%)	97.9/1.9/0.3	97.4/2.13/0.5			96.8/2.7/0.5	87.8/8.7/3.5

[a]The outer shell corresponds to 1.16 – 1.13 Å (hPif1_AMP-PNP), 1.47 – 1.43 Å (hPif1_Br peak), 1.43 – 1.39 Å (hPif1_Br inflection), 1.92 – 1.87 Å (hPif1_Br remote), 1.48 – 1.44 Å (hPif1_Apo), 4.06 – 3.96 Å (hPif1_ADF).

[b]$R_{merge} = \sum hkl \sum i |Ii - <I>| / \sum hkl \sum i <I>$ where Ii is the intensity of the ith measurement of a reflection with indexes hkl and $<I>$ is the statistically weighted average reflection intensity.

[c]R-factor $= \sum ||Fo| - |Fc|| / \sum |Fo|$ where Fo and Fc are the observed and calculated structure factor amplitudes, respectively.

[d]R-free is the R-factor calculated with 5 % of the reflections chosen at random and omitted from refinement.

[e]Outer shell for refinement corresponds to 1.16 – 1.13 Å (hPif1_AMP-PNP), 1.47 – 1.43 Å (hPif1_Br peak), 1.43 – 1.39 Å (hPif1_Br inflection), 1.92 – 1.87 Å (hPif1_Br remote), 1.48 – 1.44 Å (hPif1_Apo), 4.06 – 3.96 Å (hPif1_ADF).

[f]Root-mean-square deviation of bond lengths and bond angles from ideal geometry.

[g]Percentage of residues in most-favoured/ allowed/ outliers regions of the Ramachandran plot.

Supplementary Table 1.1. X-ray Data Collection and Refinement Statistics.

The reported resolution of this entry is 1.13 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.
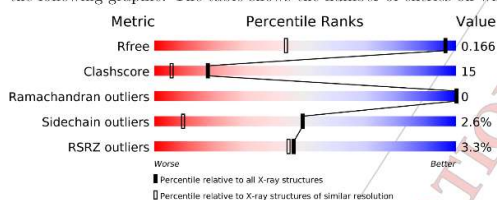
Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	111664	1189 (1.16-1.08)
Clashscore	122126	1004 (1.14-1.10)
Ramachandran outliers	120053	1188 (1.16-1.08)
Sidechain outliers	120020	1185 (1.16-1.08)
RSRZ outliers	108989	1165 (1.16-1.08)

The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5% The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.

Mol	Chain	Length	Quality of chain
1	A	418	

The reported resolution of this entry is 3.60 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.
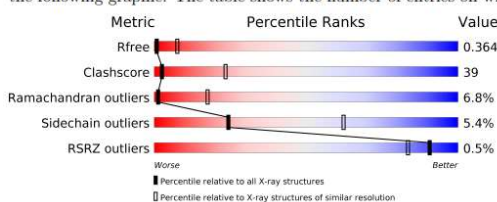
Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	111664	1053 (3.70-3.50)
Clashscore	122126	1141 (3.70-3.50)
Ramachandran outliers	120053	1102 (3.70-3.50)
Sidechain outliers	120020	1102 (3.70-3.50)
RSRZ outliers	108989	1009 (3.72-3.48)

The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5% The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.

Mol	Chain	Length	Quality of chain
1	A	442	
1	B	442	

Supplementary Table 1.2. Structure quality comparison. Structure validation reports of the 1.13Å of hPif1 residues 206-620 (**top**) and the 3.6Å hPif1 structure (residues 200-641) by (Zhou et al., 2016), Pdb code: 5FHH (**bottom**).
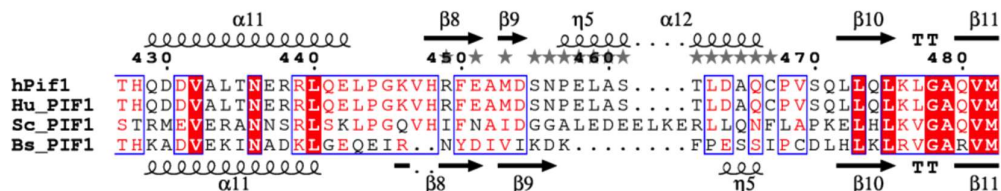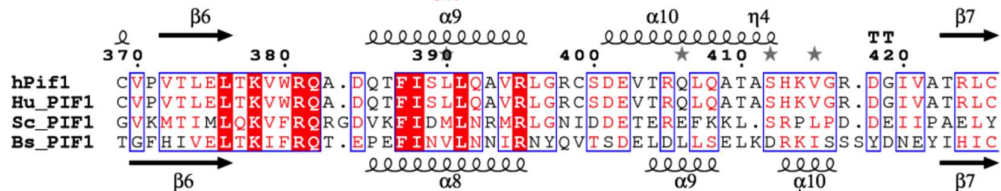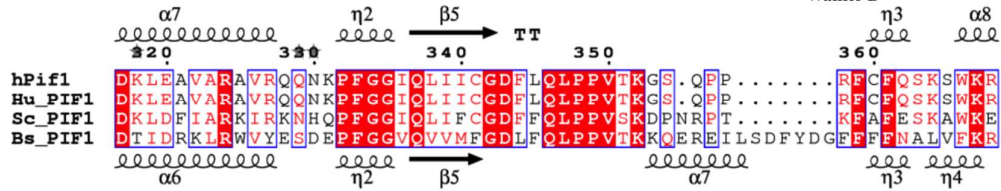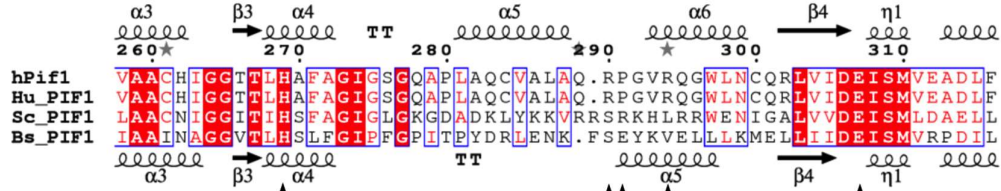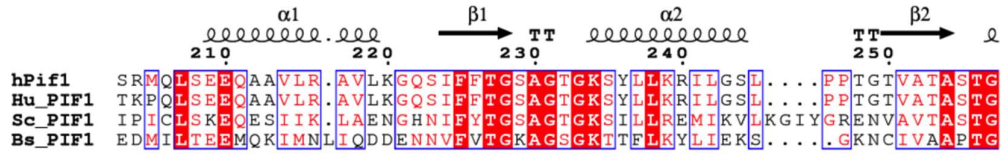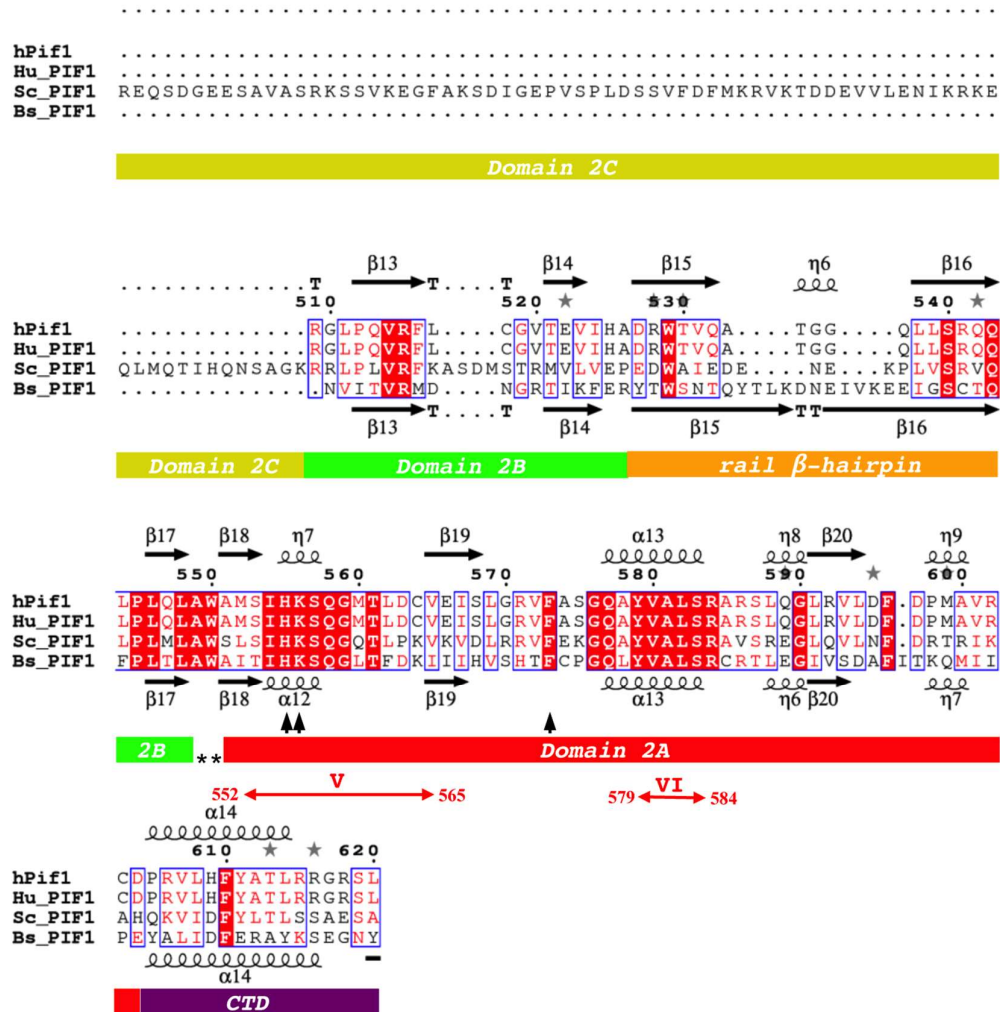
Supplementary Figure 4. Pif1 sequence alignment. The first line, hPif1, is the protein sequence of the human Pif1 protein construct for which electron density was modelled in our structures. The human Pif1 constructs used start at residue 206 and are preceded by four residues from the tag (GSRM) but the electron density for the glycine is not visible. The sequence indicated as hPif1 is aligned against the native human sequence (Hu_PIF1) beginning at residue 203 (residues 203-205 aligning with SRM of the tag), as well as the corresponding residues for Bacteroides spp. (Bs_PIF1) and the yeast (Sc_PIF1) helicase domain. Identical residues are boxed in red, similar residues are red and none conserved residues are black. Regions framed in blue have a global similarity >0.7. Secondary structural elements are indicated as α1, β1, η1 (3/10 helix), TT (β turn) and dark grey stars are residue with alternative conformations. Note, linker regions between domains are indicated by *s in the domain definition line (below sequences). Walker A (GXXXXGK[TS]) and Walker B (hhhDE) sequence motifs are also shown in the domain line (h=hydrophobic residue). The twelve residues indicated by black arrows ↑ were mutated in this study (see Table 1, main manuscript). Top and bottom secondary structures are from ESPript 3.0 (Robert, X. and Gouet, P. (2014) Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids
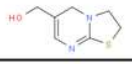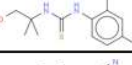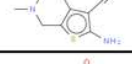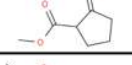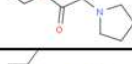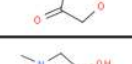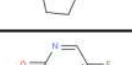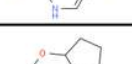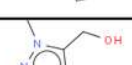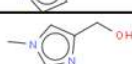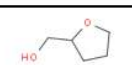
Res. 42(W1), W320-W324. doi: 10.1093/nar/gku316) for the 1.13 Å AMP-PNP complex (hPIF1) and 5FTB (BsPif1), respectively. We define domain 1B (the separation wedge) as residues 280-300 (cyan in Fig. 1, main manuscript). The residue numbers above the alignment are from the native Hu_PIF1 sequence.
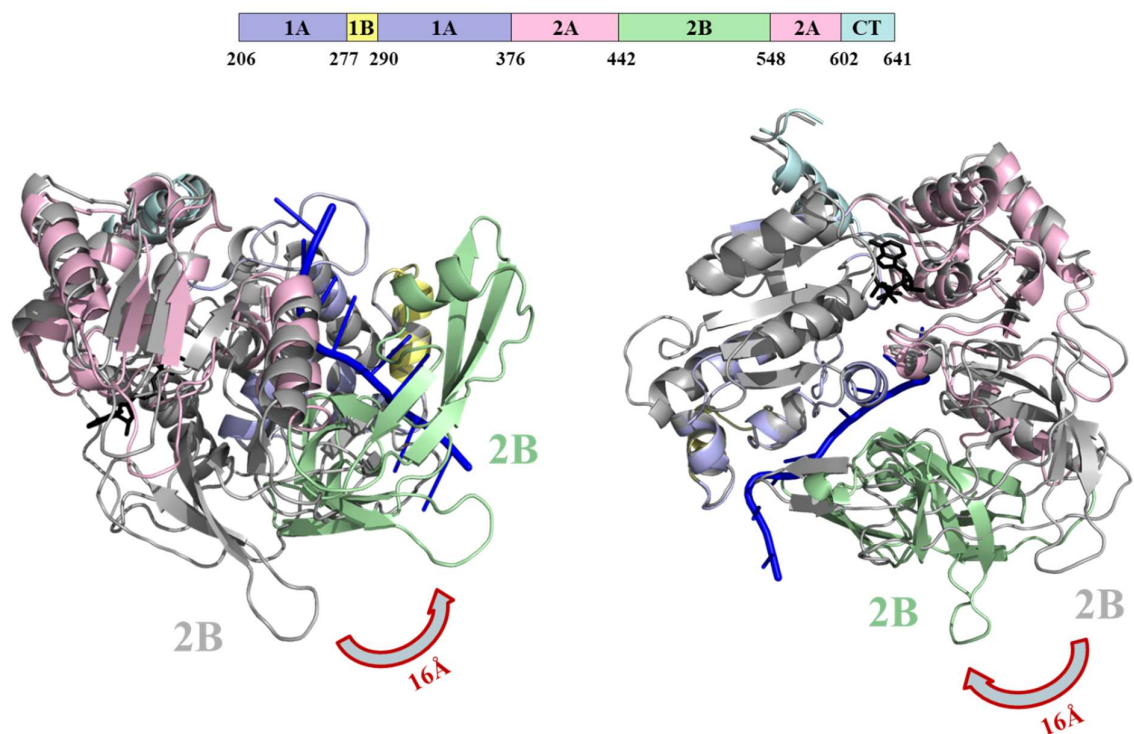
Supplementary Figure 5. Alignment of hPif1-HD AMP-PNP structure to the DNA bound model showing the 16Å movement of the domain 2B folding towards the ssDNA in two angular views. The AMP-PNP bound hPif1-HD structure is shown in gray and the modeled DNA bound hPif1 structure shown in color (domains colored accordingly). The AMP-PNP molecule is shown in black and the ssDNA colored in blue.

Sample ID	Barcodes	Rack	Row	Column	Initial ΔTm (4mM)	Retest ΔTm [4mM-triplicates] (stdev)	True/False
477	FR03623904	Rack 1	B	02	0.69	0.39 (0.02)	-
554	FR03625245	Rack 2	A	09	0.50	0.45 (0.13)	-
569	FR03625260	Rack 2	C	04	1.16	1.63 (0.07)	True
647	FR03622762	Rack 3	C	02	0.73	0.81 (0.015)	True
650	FR03622765	Rack 3	C	05	0.46	0.49 (0.06)	True
756	FR03621828	Rack 4	E	11	0.46	0.64 (0.06)	True
772	FR03621844	Rack 4	G	07	0.41	0.89 (0.88)	False
1446	FR03625641	Rack 4	H	09	0.69	-	False
827	FR03622203	Rack 5	E	02	0.64	-1.14 (0.24)	False
1428	FR03625623	Rack 5	E	09	0.42	0.04 (0.07)	False
862	FR03622238	Rack 5	H	07	1.31	-	False
947	FR03623702	Rack 7	A	02	0.67	0.88 (0.08)	True
954	FR03623709	Rack 7	A	09	0.62	0.23 (0.72)	False
1047	FR03623434	Rack 8	C	02	0.47	-	False
1069	FR03623456	Rack 8	E	04	0.48	0.59 (0.05)	True
1112	FR03625819	Rack 9	A	07	0.46	0.72 (0.04)	True
1145	FR03625852	Rack 9	D	10	0.54	-	False
1176	FR03625883	Rack 9	G	11	0.47	0.59 (0.05)	True
1184	FR03625891	Rack 9	H	09	1.32	0.56 (0.09)	True
1253	FR03624716	Rack 10	G	08	1.10	0.03 (0.18)	False
1254	FR03624724	Rack 10	G	09	0.75	-0.33 (0.05)	False
1263	FR03624717	Rack 10	H	08	1.23	-0.61 (0.07)	False
1264	FR03624725	Rack 10	H	09	1.00	0.52 (0.12)	-
1391	FR03624122	Rack 12	E	06	4.26	0.34 (0.2)	False
1691	FR03622738	Rack 13	A	05	0.41	-1.81 (0.11)	False
1648	FR03625324	Rack 13	B	05	0.46	0.54 (0.15)	-
1650	FR03624657	Rack 13	C	02	0.54	-0.18 (0.02)	False

Supplementary Table 2. Primary results of the 1200 fragment molecule library screening obtained at the University of York. Fragments were tested at 4 mM concentration in both the initial screen and re-test. Fragment 581 precipitated upon re-testing (Re-testing uses the stock from the vials, whereas the initial screen uses the daughter plates). Fragments 318, 396 and 679 gave flat thermal curves.

Number	Barcodes	Initial ΔTm (4mM)	Retest ΔTm [4mM-triplicates] (stdev)	Structure
1	FR03623904	0.69	0.39 (0.02)	
2	FR03625245	0.50	0.45 (0.13)	
3	FR03625260	1.16	1.63 (0.07)	
4	FR03622762	0.73	0.81 (0.015)	
5	FR03622765	0.46	0.49 (0.06)	
6	FR03621828	0.46	0.64 (0.06)	
7	FR03623702	0.67	0.88 (0.08)	
8	FR03623456	0.48	0.59 (0.05)	
9	FR03625819	0.46	0.72 (0.04)	
10	FR03625883	0.47	0.59 (0.05)	
11	FR03625891	1.32	0.56 (0.09)	
12	FR03624725	1.00	0.52 (0.12)	
13	FR03625324	0.46	0.54 (0.15)	
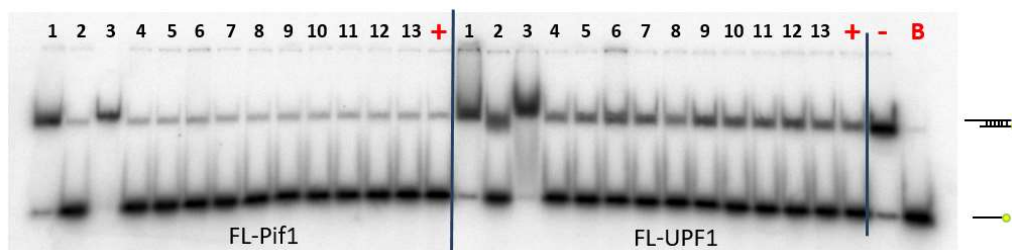
Supplementary Table 3. Final fragment molecules selected after secondary screening (at the University of York) (**top**) and their inhibitory effect on both FL-Pif1 and FL-UPF1 at 1 mM concentration (**bottom**).
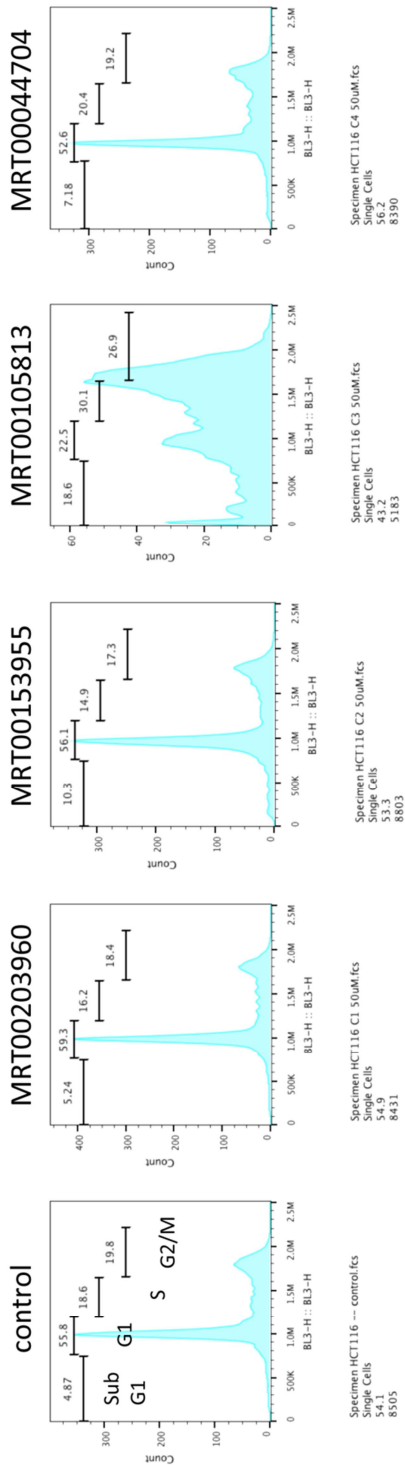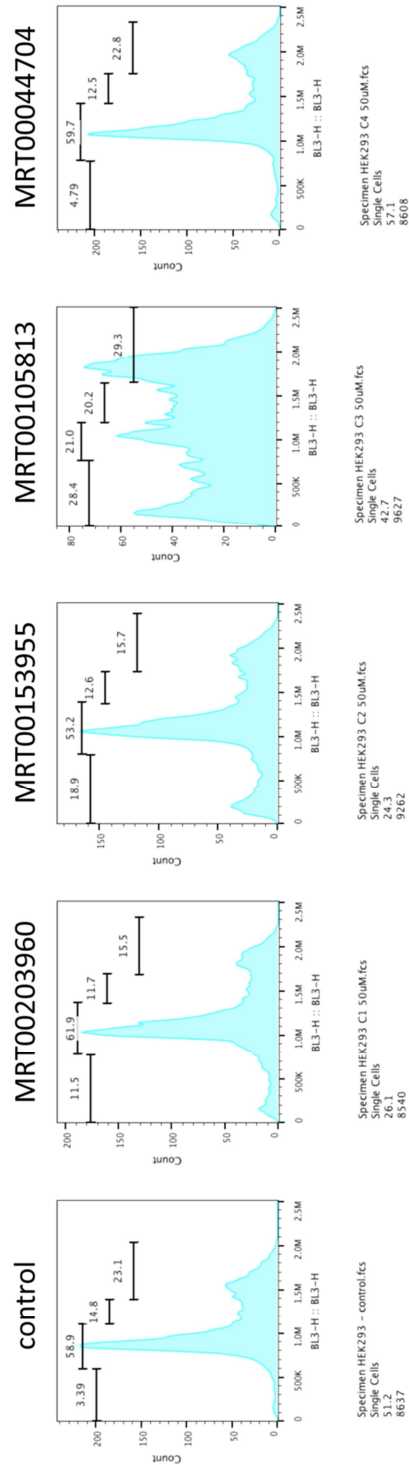
Supplementary Figure 6. Cells analysed by FACS (fluorescence activated cell sorting) for cell-cycle distribution following treatment with Pif1 inhibitors for 48 hours.

6 References

ABDELHALEEM, M. 2004. Over-expression of RNA helicases in cancer. *Anticancer Research,* 24**,** 3951-3953.

AGGARWAL, M., BANERJEE, T., SOMMERS, J. A. & BROSH JR, R. M. 2013. Targeting an Achilles' heel of cancer with a WRN helicase inhibitor. *Cell Cycle,* 12**,** 3329-3335.

ANAND, S. P. & KHAN, S. A. 2004. Structure-specific DNA binding and bipolar helicase activities of PcrA. *Nucleic Acids Research,* 32**,** 3190-3197.

ARAI, A., CHANO, T., FUTAMI, K., FURUICHI, Y., IKEBUCHI, K., INUI, T., TAMENO, H., OCHI, Y., SHIMADA, T., HISA, Y. & OKABE, H. 2011. RECQL1 and WRN Proteins Are Potential Therapeutic Targets in Head and Neck Squamous Cell Carcinoma. *Cancer Research,* 71**,** 4598-4607.

AYLON, Y. & KUPIEC, M. 2004. New insights into the mechanism of homologous recombination in yeast. *Mutation Research - Reviews in Mutation Research,* 566**,** 231-248.

BACOLLA, A., COLLINS, J. R., GOLD, B., CHUZHANOVA, N., YI, M., STEPHENS, R. M., STEFANOV, S., OLSH, A., JAKUPCIAK, J. P., DEAN, M., LEMPICKI, R. A., COOPER, D. N. & WELLS, R. D. 2006a. Long homopurine center dot homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Research,* 34**,** 2663-2675.

BACOLLA, A. & WELLS, R. D. 2009. Non-B DNA conformations as determinants of mutagenesis and human disease. *Molecular Carcinogenesis,* 48**,** 273-285.

BACOLLA, A., WOJCIECHOWSKA, M., KOSMIDER, B., LARSON, J. E. & WELLS, R. D. 2006b. The involvement of non-B DNA structures in gross chromosomal rearrangements. *DNA Repair,* 5**,** 1161-1170.

BANASIK, M. & UEDA, K. 1994. Inhibitors and activators of ADP-ribosylation reactions. *Mol Cell Biochem,* 138**,** 185-197.

BERGER, J. M. 2008. SnapShot: nucleic acid helicases and translocases. *Cell,* 134**,** 888-888.e1.

BERNSTEIN, C., BERNSTEIN, H., PAYNE, C. M. & GAREWAL, H. 2002. DNA repair/pro-apoptotic dual-role proteins in five major DNA repair pathways: Fail-safe protection against carcinogenesis. *Mutation Research - Reviews in Mutation Research,* 511**,** 145-178.

BIFFI, G., TANNAHILL, D., MCCAFFERTY, J. & BALASUBRAMANIAN, S. 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature Chemistry,* 5**,** 182-186.

BIRNBOIM, H. & DOLY, J. A. 1979. A rapid alkaline extraction prodedure for screening recombinant plasmid DNA. *Nucleic Acids Research*, 7, 859-869.

BOCHMAN, M. L., JUDGE, C. P., ZAKIAN, V. A. & KELLOGG, D. 2011. The Pif1 family in prokaryotes: what are our helicases doing in your bacteria? *Molecular Biology of the Cell,* 22**,** 1955-1959.

BOCHMAN, M. L., SABOURI, N. & ZAKIAN, V. A. 2010. Unwinding the functions of the Pif1 family helicases. *DNA Repair,* 9**,** 237-249.

BOULE, J. B., VEGA, L. R. & ZAKIAN, V. A. 2005. The yeast Pif1p helicase removes telomerase from telomeric DNA. *Nature,* 438**,** 57-61.

BRADFORD, M. M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry,* 72**,** 248-254.

BRAI, A., FAZI, R., TINTORI, C., ZAMPERINI, C., BUGLI, F., SANGUINETTI, M., STIGLIANO, E., ESTÉ, J., BADIA, R., FRANCO, S., MARTINEZ, M. A.,

MARTINEZ, J. P., MEYERHANS, A., SALADINI, F., ZAZZI, M., GARBELLI, A., MAGA, G. & BOTTA, M. 2016. Human DDX3 protein is a valuable target to develop broad spectrum antiviral agents. *Proceedings of the National Academy of Sciences,* 113**,** 5388-5393.

BROSH JR, R. M. 2013. DNA helicases involved in DNA repair and their roles in cancer. *Nature Reviews Cancer,* 13**,** 542-558.

BRUENING, J.-G., HOWARD, J. L. & MCGLYNN, P. 2014. Accessory replicative helicases and the replication of protein-bound DNA. *Journal of Molecular Biology,* 426**,** 3917-3928.

BRYAN, T. M., ENGLEZOU, A., DALLA-POZZA, L., DUNHAM, M. A. & REDDEL, R. R. 1997. Evidence for an alternative mechanism for maintaining telomere length in human tumors and tumor-derived cell lines. *Nature Medicine,* 3**,** 1271-1274.

BRYANT, H. E., SCHULTZ, N., THOMAS, H. D., PARKER, K. M., FLOWER, D., LOPEZ, E., KYLE, S., MEUTH, M., CURTIN, N. J. & HELLEDAY, T. 2005. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature,* 434**,** 913-917.

BRÁZDA, V., LAISTER, R. C., JAGELSKÁ, E. B. & ARROWSMITH, C. 2011. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Molecular Biology,* 12**,** 33.

BUDD, M. E., REIS, C. C., SMITH, S., MYUNG, K. & CAMPBELL, J. L. 2006. Evidence suggesting that Pif1 helicase functions in DNA replication with the DNA2 helicase/nuclease and DNA polymerase delta. *Molecular and Cellular Biology,* 26**,** 2490-2500.

BUTTNER, K., NEHRING, S. & HOPFNER, K. P. 2007. Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nature Structural & Molecular Biology,* 14**,** 647-652.

CASKEY, C. T., PIZZUTI, A., FU, Y. H., FENWICK, R. G. & NELSON, D. L. 1992. Triplet repeat mutations in human-disease. *Science,* 256**,** 784-789.

CERVELLI, T., BORGHINI, A., GALLI, A. & ANDREASSI, M. G. 2012. DNA damage and repair in atherosclerosis: current insights and future perspectives. *International journal of molecular sciences,* 13**,** 16929-16944.

CHABAN, Y., STEAD, J. A., RYZHENKOVA, K., WHELAN, F., LAMBER, E. P., ANTSON, A., SANDERS, C. M. & ORLOVA, E. V. 2015. Structural basis for DNA strand separation by a hexameric replicative helicase. *Nucleic Acids Research,* 43**,** 8551-8563.

CHAN, K. K. L., ZHANG, Q. M. & DIANOV, G. L. 2006. Base excision repair fidelity in normal and cancer cells. *Mutagenesis,* 21**,** 173-178.

CHAPMAN, J. R., TAYLOR, M. R. & BOULTON, S. J. 2012. Playing the end game: DNA double-strand break repair pathway choice. *Mol Cell,* 47**,** 497-510.

CHEN, M. C., TIPPANA, R., DEMESHKINA, N. A., MURAT, P., BALASUBRAMANIAN, S., MYONG, S. & FERRÉ-D'AMARÉ, A. R. 2018. Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature,* 558**,** 465-469.

CHEN, P. & LI, C. M. 2007. Nanopore unstacking of single-stranded DNA helices. *Small,* 3**,** 1204-1208.

CHEN, W.-F., DAI, Y.-X., DUAN, X.-L., LIU, N.-N., SHI, W., LI, N., LI, M., DOU, S.-X., DONG, Y.-H., RETY, S. & XI, X.-G. 2016. Crystal structures of the BsPif1 helicase reveal that a major movement of the 2B SH3 domain is required for DNA unwinding. *Nucleic Acids Research,* 44**,** 2949-2961.

CHISHOLM, K. M., AUBERT, S. D., FREESE, K. P., ZAKIAN, V. A., KING, M.-C. & WELCSH, P. L. 2012. A genomewide screen for suppressors of Alu-mediated rearrangements reveals a role for PIF1. *PloS one,* 7**,** e30748-e30748.

CONGREVE, M., CARR, R., MURRAY, C. & JHOTI, H. 2003. A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today,* 8**,** 876-877.

DAMIA, G. & D'INCALCI, M. 2007. Targeting DNA repair as a promising approach in cancer therapy. *European Journal of Cancer,* 43**,** 1791-1801.

DATTA, A. & BROSH, R. M. 2018. New insights into DNA helicases as druggable targets for cancer therapy. *Frontiers in Molecular Biosciences,* 5**,** 59.

DAVIS, B. J. & ERLANSON, D. A. 2013. Learning from our mistakes: The 'unknown knowns' in fragment screening. *Bioorganic & Medicinal Chemistry Letters,* 23**,** 2844-2852.

DE, S. & MICHOR, F. 2011. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nature Structural and Molecular Biology,* 18**,** 950-955.

DEHGHANI-TAFTI, S. & SANDERS, C. M. 2017. DNA substrate recognition and processing by the full-length human UPF1 helicase. *Nucleic Acids Research,* 45**,** 7354-7366.

DILLEY, R. L. & GREENBERG, R. A. 2015. ALTernative telomere maintenance and cancer. *Trends in Cancer,* 1**,** 145-156.

DILLINGHAM, M. S., SPIES, M. & KOWALCZYKOWSKI, S. C. 2003. RecBCD enzyme is a bipolar DNA helicase. *Nature,* 423**,** 893-897.

DONG, Z. & FASULLO, M. 2003. Multiple recombination pathways for sister chromatid exchange in *Saccharomyces cerevisiae*: Role of RAD1 and the RAD52 epistasis group genes. *Nucleic Acids Research,* 31**,** 2576-2585.

DONMEZ, I. & PATEL, S. S. 2008. Coupling of DNA unwinding to nucleotide hydrolysis in a ring-shaped helicase. *The EMBO Journal,* 27**,** 1718-1726.

DONOVAN, S., HARWOOD, J., DRURY, L. S. & DIFFLEY, J. F. X. 1997. Cdc6p-dependent loading of Mcm proteins onto pre-replicative chromatin in budding yeast. *Proceedings of the National Academy of Sciences of the United States of America,* 94**,** 5611-5616.

EDWARDS, S. F., SIRITO, M., KRAHE, R. & SINDEN, R. R. 2009. A Z-DNA sequence reduces slipped-strand structure formation in the myotonic dystrophy type 2 (CCTG).(CAGG) repeat. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 3270-3275.

ELLES, L. M. S. & UHLENBECK, O. C. 2008. Mutation of the arginine finger in the active site of *Escherichia coli* DbpA abolishes ATPase and helicase activity and confers a dominant slow growth phenotype. *Nucleic Acids Research,* 36**,** 41-50.

EMSLEY, P. & COWTAN, K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography,* 60**,** 2126-2132.

ENEMARK, E. J. & JOSHUA-TOR, L. 2006. Mechanism of DNA translocation in a replicative hexameric helicase. *Nature,* 442**,** 270-275.

EOFF, R. L. & RANEY, K. D. 2006. Intermediates revealed in the kinetic mechanism for DNA unwinding by a monomeric helicase. *Nature Structural & Molecular Biology,* 13**,** 242-249.

ERBA, E., CAVALLARO, E., DAMIA, G., MANTOVANI, R., DI SILVIO, A., DI FRANCESCO, A. M., RICCARDI, R., CUEVAS, C., FAIRCLOTH, G. T. & D'INCALCI, M. 2004. The unique biological features of the marine product Yondelis (ET-743, trabectedin) are shared by its analog ET-637, which lacks the C ring. *Oncol Res,* 14**,** 579-587.

FAIRMAN-WILLIAMS, M. E., GUENTHER, U. P. & JANKOWSKY, E. 2010. SF1 and SF2 helicases: Family matters. *Current Opinion in Structural Biology,* 20**,** 313-324.

FANG, H., NIE, L., CHI, Z., LIU, J., GUO, D., LU, X., HEI, T. K., BALAJEE, A. S. & ZHAO, Y. 2013. RecQL4 helicase amplification is involved in human breast tumorigenesis. *Plos One,* 8**,** e69600-e69600.

FERRER, M., CHERNIKOVA, T. N., YAKIMOV, M. M., GOLYSHIN, P. N. & TIMMIS, K. N. 2003. Chaperonins govern growth of Escherichia coli at low temperatures. *Nature Biotechnology,* 21**,** 1266-1267.

FIORINI, F., BAGCHI, D., LE HIR, H. & CROQUETTE, V. 2015. Human UPF1 is a highly processive RNA helicase and translocase with RNP remodelling activities. *Nature Communications,* 6**,** 7581.

FLEMING, A. M., DING, Y., ROGERS, R. A., ZHU, J., BURTON, A. D., CARLISLE, C. B. & BURROWS, C. J. 2017. 4n-1 Is a "sweet spot" in DNA i-motif folding of 2'-deoxycytidine homopolymers. *J Am Chem Soc,* 139**,** 4682-4689.

FOURY, F. & KOLODYNSKI, J. 1983. Pif mutation blocks recombination between mitochondrial rho+ and rho- genomes having tandemly arrayed repeat units in *Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences,* 80**,** 5345-5349.

FREES, S., MENENDEZ, C., CRUM, M. & BAGGA, P. 2014. QGRS-Conserve: A computational method for discovering evolutionarily conserved G-quadruplex motifs. *Hum Genomics.* 2014; 8(1): 8.

FUTAMI, K. & FURUICHI, Y. 2014. RECQL1 and WRN DNA repair helicases: potential therapeutic targets and proliferative markers against cancers. *Front Genet,* 5**,** 441.

FUTAMI, K., SHIMAMOTO, A. & FURUICHI, Y. 2007. Mitochondrial and nuclear localization of human Pif1 helicase. *Biological and Pharmaceutical Bulletin,* 30**,** 1685-1692.

GAGOU, M. E., GANESH, A., PHEAR, G., ROBINSON, D., PETERMANN, E., COX, A. & MEUTH, M. 2014. Human PIF1 helicase supports DNA replication and cell growth under oncogenic-stress. *Oncotarget,* 5**,** 11381-11398.

GAGOU, M. E., GANESH, A., THOMPSON, R., PHEAR, G., SANDERS, C. & MEUTH, M. 2011. Suppression of apoptosis by PIF1 helicase in human tumor cells. *Cancer Research,* 71**,** 4998-5008.

GELLERT, M., ODEA, M. H. & MIZUUCHI, K. 1983. Slow cruciform transitions in palindromic DNA. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences,* 80**,** 5545-5549.

GEORGE, T., WEN, Q., GRIFFITHS, R., GANESH, A., MEUTH, M. & SANDERS, C. M. 2009. Human Pif1 helicase unwinds synthetic DNA structures resembling stalled DNA replication forks. *Nucleic Acids Research,* 37**,** 6491-6502.

GEORGESCU, R., YUAN, Z., BAI, L., DE LUNA ALMEIDA SANTOS, R., SUN, J., ZHANG, D., YURIEVA, O., LI, H. & O'DONNELL, M. E. 2017. Structure of eukaryotic CMG helicase at a replication fork and implications to replisome architecture and origin initiation. *Proc Natl Acad Sci U S A,* 114**,** e697-e706.

GERONIMO, C. L., SINGH, S. P., GALLETTO, R. & ZAKIAN, V. A. 2018. The signature motif of the *Saccharomyces cerevisiae* Pif1 DNA helicase is essential *in vivo* for mitochondrial and nuclear functions and in vitro for ATP*ase* activity. *Nucleic Acids Research***,** 8357-8370.

GONG, Z., KIM, J. E., LEUNG, C. C., GLOVER, J. N. & CHEN, J. 2010. BACH1/FANCJ acts with TopBP1 and participates early in DNA replication checkpoint control. *Mol Cell,* 37**,** 438-46.

GU, Y., MASUDA, Y. & KAMIYA, K. 2008. Biochemical analysis of human PIF1 helicase and functions of its N-terminal domain. *Nucleic Acids Research,* 36**,** 6295-6308.

GUO, M., HUNDSETH, K., DING, H., VIDHYASAGAR, V., INOUE, A., NGUYEN, C.-H., ZAIN, R., LEE, J. S. & WU, Y. 2015. A distinct triplex DNA unwinding activity of ChlR1 helicase. *The Journal of Biological Chemistry,* 290**,** 5174-5189.

HA, S. C., LOWENHAUPT, K., RICH, A., KIM, Y. G. & KIM, K. K. 2005. Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature,* 437**,** 1183-1186.

HAHN, W. C., STEWART, S. A., BROOKS, M. W., YORK, S. G., EATON, E., KURACHI, A., BEIJERSBERGEN, R. L., KNOLL, J. H. M., MEYERSON, M. & WEINBERG, R. A. 1999. Inhibition of telomerase limits the growth of human cancer cells. *Nature Medicine,* 5**,** 1164-1170.

HALL, M. C. & MATSON, S. W. 1999. Helicase motifs: the engine that powers DNA unwinding. *Molecular Microbiology,* 34**,** 867-877.

HAMADA, H. & KAKUNAGA, T. 1982. Potential Z-DNA forming sequences are highly dispersed in the human genome. *Nature,* 298**,** 396-398.

HANAHAN, D. & WEINBERG, R. A. 2000. The Hallmarks of Cancer. *Cell,* 100**,** 57-70.

HEFFERIN, M. L. & TOMKINSON, A. E. 2005. Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair (Amst),* 4**,** 639-648.

HEGI, M. E., DISERENS, A. C., GODARD, S., DIETRICH, P. Y., REGLI, L., OSTERMANN, S., OTTEN, P., VAN MELLE, G., DE TRIBOLET, N. & STUPP, R. 2004. Clinical trial substantiates the predictive value of O-6-Methylguanine-DNA methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. *Clinical Cancer Research,* 10**,** 1871-1874.

HICKSON, I. D. 2003. RecQ helicases: caretakers of the genome. *Nature Reviews Cancer,* 3**,** 169-187.

HO, S. N., HUNT, H. D., HORTON, R. M., PULLEN, J. K. & PEASE, L. R. 1989. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene,* 77**,** 51-59.

HTUN, H. & DAHLBERG, J. E. 1988. Single strands, triple strands, and kinks in H-DNA. *Science,* 241**,** 1791-1796.

HU, Y., RAYNARD, S., SEHORN, M. G., LU, X., BUSSEN, W., ZHENG, L., STARK, J. M., BARNES, E. L., CHI, P., JANSCAK, P., JASIN, M., VOGEL, H., SUNG, P. & LUO, G. 2007. RECQL5/Recql5 helicase regulates homologous recombination and suppresses tumor formation via disruption of Rad51 presynaptic filaments. *Genes and Development,* 21**,** 3073-3084.

HUBER, M. D., LEE, D. C. & MAIZELS, N. 2002. G4 DNA unwinding by BLM and Sgs1p: Substrate specificity and substrate-specific inhibition. *Nucleic Acids Research,* 30**,** 3954-3961.

HUNG, A. W., SILVESTRE, H. L., WEN, S., CIULLI, A., BLUNDELL, T. L. & ABELL, C. 2009. Application of fragment growing and fragment linking to the discovery of inhibitors of *Mycobacterium tuberculosis* pantothenate synthetase. *Angew Chem Int Ed Engl,* 48**,** 8452-8456.

HUPPERT, J. L. & BALASUBRAMANIAN, S. 2005. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research,* 33**,** 2908-2916.

HUPPERT, J. L. & BALASUBRAMANIAN, S. 2007. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Research,* 35**,** 406-413.

HUPPERT, J. L., BUGAUT, A., KUMARI, S. & BALASUBRAMANIAN, S. 2008. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Research,* 36**,** 6260-6268.

IGGO, R. D. & LANE, D. P. 1989. Nuclear protein p68 is an RNA-dependent ATPase. *The EMBO journal,* 8**,** 1827-1831.

INAGAKI, H., OHYE, T., KOGO, H., KATO, T., BOLOR, H., TANIGUCHI, M., SHAIKH, T. H., EMANUEL, B. S. & KURAHASHI, H. 2009. Chromosomal instability mediated by non-B DNA: Cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Research,* 19**,** 191-198.

INOUE, H., NOJIMA, H. & OKAYAMA, H. 1990. High efficiency transformation of Escherichia coli with plasmids. *Gene,* 96**,** 23-28.

IVESSA, A. S., ZHOU, J. Q. & ZAKIAN, V. A. 2000. The *Saccharomyces* Pif1p DNA helicase and the highly related Rrm3p have opposite effects on replication fork progression in ribosomal DNA. *Cell,* 100**,** 479-489.

JACKSON, R. N., LAVIN, M., CARTER, J. & WIEDENHEFT, B. 2014. Fitting CRISPR-associated Cas3 into the helicase family tree. *Current Opinion in Structural Biology,* 24**,** 106-114.

JAIN, A., BACOLLA, A., DEL MUNDO, I. M., ZHAO, J., WANG, G. & VASQUEZ, K. M. 2013. DHX9 helicase is involved in preventing genomic instability induced by alternatively structured DNA in human cells. *Nucleic Acids Research,* 41**,** 10345-10357.

JAIN, A., WANG, G. & VASQUEZ, K. M. 2008. DNA triple helices: Biological consequences and therapeutic potential. *Biochimie,* 90**,** 1117-1130.

JIN, Y., MOLT, R. W. & BLACKBURN, G. M. 2017. Metal fluorides: Tools for structural and computational analysis of phosphoryl transfer enzymes. *Topics in Current Chemistry,* 375**,** 36.

JIRICNY, J. 2006. The multifaceted mismatch-repair system. *Nature Reviews Molecular Cell Biology,* 7**,** 335-346.

JUBB, H., HIGUERUELO, A. P., WINTER, A. & BLUNDELL, T. L. 2012. Structural biology and drug discovery for protein-protein interactions. *Trends in Pharmacological Sciences,* 33**,** 241-248.

KABSCH, W. 2010. XDS. *Acta Crystallographica Section D-Biological Crystallography,* 66**,** 125-132.

KAMILERI, I., KARAKASILIOTI, I. & GARINIS, G. A. 2012. Nucleotide excision repair: new tricks with old bricks. *Trends in Genetics,* 28**,** 566-573.

KAROW, J. K., WU, L. & HICKSON, I. D. 2000. RecQ family helicases: roles in cancer and aging. *Current Opinion in Genetics & Development,* 10**,** 32-38.

KAWABE, T., TSUYAMA, N., KITAO, S., NISHIKAWA, K., SHIMAMOTO, A., SHIRATORI, M., MATSUMOTO, T., ANNO, K., SATO, T., MITSUI, Y., SEKI, M., ENOMOTO, T., GOTO, M., ELLIS, N. A., IDE, T., FURUICHI, Y. & SUGIMOTO, M. 2000. Differential regulation of human RecQ family helicases in cell transformation and cell cycle. *Oncogene,* 19**,** 4764-4772.

KHUU, P., SANDOR, M., DEYOUNG, J. & HO, P. S. 2007. Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 16528-16533.

KIM, D., REDDY, S., KIM, D. Y., RICH, A., LEE, S., KIM, K. K. & KIM, Y.-G. 2009. Base extrusion is found at helical junctions between right- and left-handed forms of DNA and RNA. *Nucleic Acids Research,* 37**,** 4353-4359.

KIM, D.-E., NARAYAN, M. & PATEL, S. S. 2002. T7 DNA helicase: A molecular motor that processively and unidirectionally translocates along single-stranded DNA. *Journal of Molecular Biology,* 321**,** 807-819.

KIM, N. W. 1997. Clinical implications of telomerase in cancer. *European Journal of Cancer Part A,* 33**,** 781-786.

KITAO, S., SHIMAMOTO, A., GOTO, M., MILLER, R. W., SMITHSON, W. A., LINDOR, N. M. & FURUICHI, Y. 1999. Mutations in RECQL4 cause a subset of cases of Rothmund-Thomson syndrome. *Nature Genetics,* 22**,** 82-84.

KLEIN, J., MEINECKE, R., MAYER, M. & MEYER, B. 1999. Detecting binding affinity to immobilized receptor proteins in compound libraries by HR-MAS STD NMR. *Journal of the American Chemical Society,* 121**,** 5336-5337.

KRONTIRIS, T. G. 1995. Minisatellites and human-disease. *Science,* 269**,** 1682-1683.

LABIB, K., TERCERO, J. A. & DIFFLEY, J. F. X. 2000. Uninterrupted MCM2-7 function required for DNA replication fork progression. *Science,* 288**,** 1643-1647.

LAHAYE, A., STAHL, H., THINES-SEMPOUX, D. & FOURY, F. 1991. PIF1: A DNA helicase in yeast mitochondria. *EMBO Journal,* 10**,** 997-1007.

LAMOREE, B. & HUBBARD, R. E. 2017. Current perspectives in fragment-based lead discovery (FBLD). *Essays in biochemistry,* 61**,** 453-464.

LANSDORP, P. M. 2009. Telomeres and disease. *The EMBO Journal,* 28**,** 2532-2540.

LAUGHLAN, G., MURCHIE, A. I. H., NORMAN, D. G., MOORE, M. H., MOODY, P. C. E., LILLEY, D. M. J. & LUISI, B. 1994. The high-resolution crystal-structure of a parallel-stranded guanine tetraplex. *Science,* 265**,** 520-524.

LAUPER, J. M., KRAUSE, A., VAUGHAN, T. L. & MONNAT, R. J., JR. 2013. Spectrum and risk of neoplasia in Werner syndrome: a systematic review. *PloS one,* 8**,** e59709-e59709.

LEAVITT, S. & FREIRE, E. 2001. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current Opinion in Structural Biology,* 11**,** 560-566.

LEE, J. Y. & YANG, W. 2006. UvrD Helicase unwinds DNA one base pair at a time by a two-part power stroke. *Cell,* 127**,** 1349-1360.

LEVIN, M. K., WANG, Y.-H. & PATEL, S. S. 2004. The functional interaction of the hepatitis C virus helicase molecules is responsible for unwinding processivity. Vol. 279 (2004) 26005–26012. *Journal of Biological Chemistry,* 279**,** 42350-42350.

LIN, H., LIU, W., FANG, Z., LIANG, X., LI, J., BAI, Y., LIN, L., YOU, H., PEI, Y., WANG, F. & ZHANG, Z.-Y. 2015. Overexpression of DHX32 contributes to the growth and metastasis of colorectal cancer. *Scientific reports,* 5**,** 9247-9247.

LIPPS, H. J. & RHODES, D. 2009. G-quadruplex structures: *In vivo* evidence and function. *Trends in Cell Biology,* 19**,** 414-422.

LOHMAN, T. M. & BJORNSON, K. P. 1996. Mechanisms of helicase-catalyzed DNA unwinding. *Annual Review of Biochemistry,* 65**,** 169-214.

LOHMAN, T. M., TOMKO, E. J. & WU, C. G. 2008. Non-hexameric DNA helicases and translocases: mechanisms and regulation. *Nature Reviews Molecular Cell Biology,* 9**,** 391-401.

LONDON, T. B. C., BARBER, L. J., MOSEDALE, G., KELLY, G. P., BALASUBRAMANIAN, S., HICKSON, I. D., BOULTON, S. J. & HIOM, K. 2008. FANCJ is a structure-specific DNA helicase associated with the maintenance of genomic G/C tracts. *Journal of Biological Chemistry,* 283**,** 36132-36139.

LOPEZ CASTEL, A., CLEARY, J. D. & PEARSON, C. E. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol,* 11**,** 165-170.

LU, K.-Y., CHEN, W.-F., RETY, S., LIU, N.-N., WU, W.-Q., DAI, Y.-X., LI, D., MA, H.-Y., DOU, S.-X. & XI, X.-G. 2018. Insights into the structural and mechanistic basis of multifunctional *S. cerevisiae* Pif1p helicase. *Nucleic Acids Research,* 46**,** 1486-1500.

LYUBIMOV, A. Y., STRYCHARSKA, M. & BERGER, J. M. 2011. The nuts and bolts of ring-translocase structure and mechanism. *Current Opinion in Structural Biology,* 21**,** 240-248.

MALUF, N. K., FISCHER, C. J. & LOHMAN, T. M. 2003. A dimer of *Escherichia coli* UvrD is the active form of the helicase *in vitro*. *Journal of Molecular Biology,* 325**,** 913-935.

MANOSAS, M., XI, X. G., BENSIMON, D. & CROQUETTE, V. 2010. Active and passive mechanisms of helicases. *Nucleic Acids Research,* 38**,** 5518-5526.

MAO, F. J., SIDOROVA, J. M., LAUPER, J. M., EMOND, M. J. & MONNAT, R. J. 2010. The human WRN and BLM RecQ helicases differentially regulate cell proliferation and survival after chemotherapeutic DNA damage. *Cancer Research,* 70**,** 6548-6555.

MASHALIDIS, E. H., ŚLEDŹ, P., LANG, S. & ABELL, C. 2013. A three-stage biophysical screening cascade for fragment-based drug discovery. *Nature Protocols,* 8**,** 2309.

MATEYAK, M. K. & ZAKIAN, V. A. 2006. Human PIF helicase is cell cycle regulated and associates with telomerase. *Cell Cycle,* 5**,** 2796-2804.

MAXAM, A. M. & GILBERT, W. 1977. New method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 560-564.

MCEACHERN, M. J., KRAUSKOPF, A. & BLACKBURN, E. H. 2000. Telomeres and their control. *Annual Review of Genetics,* 34**,** 331-358.

MENDOZA-MALDONADO, R., FAORO, V., BAJPAI, S., BERTI, M., ODREMAN, F., VINDIGNI, M., IUS, T., GHASEMIAN, A., BONIN, S., SKRAP, M., STANTA, G. & VINDIGNI, A. 2011. The human RECQ1 helicase is highly expressed in glioblastoma and plays an important role in tumor cell proliferation. *Molecular cancer,* 10**,** 83-83.

MIRKIN, S. M. 2007. Expandable DNA repeats and human disease. *Nature,* 447**,** 932-940.

MOHAMMAD, J. B., WALLGREN, M. & SABOURI, N. 2018. The Pif1 signature motif of Pfh1 is necessary for both protein displacement and helicase unwinding activities, but is dispensable for strand-annealing activity. *Nucleic acids research,* 46**,** 8516-8531.

MOLES, R., BAI, X. T., CHAIB-MEZRAG, H. & NICOT, C. 2016. WRN-targeted therapy using inhibitors NSC 19630 and NSC 617145 induce apoptosis in HTLV-1-transformed adult T-cell leukemia cells. *Journal of hematology & oncology,* 9**,** 121-121.

MOYZIS, R. K., TORNEY, D. C., MEYNE, J., BUCKINGHAM, J. M., WU, J. R., BURKS, C., SIROTKIN, K. M. & GOAD, W. B. 1989. The distribution of interspersed repetitive DNA-sequences in the human genome. *Genomics,* 4**,** 273-289.

MURSHUDOV, G. N., VAGIN, A. A. & DODSON, E. J. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D-Biological Crystallography,* 53**,** 240-255.

NAG, D. K. & PETES, T. D. 1991. 7-Base-pair inverted repeats in DNA form stable hairpins *in vivo* in *saccharomyces-cerevisiae. Genetics,* 129**,** 669-673.

NANDURI, B., BYRD, A. K., EOFF, R. L., TACKETTT, A. J. & RANEY, K. D. 2002. Pre-steady-state DNA unwinding by bacteriophage T4 Dda helicase reveals a monomeric molecular motor. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 14722-14727.

NGUYEN, G. H., DEXHEIMER, T. S., ROSENTHAL, A. S., CHU, W. K., SINGH, D. K., MOSEDALE, G., BACHRATI, C. Z., SCHULTZ, L., SAKURAI, M., SAVITSKY, P., ABU, M., MCHUGH, P. J., BOHR, V. A., HARRIS, C. C., JADHAV, A., GILEADI, O., MALONEY, D. J., SIMEONOV, A. & HICKSON, I. D. 2013. A small molecule inhibitor of the BLM helicase modulates chromosome stability in human cells. *Chemistry and Biology,* 20**,** 55-62.

NOGUCHI, Y., YUAN, Z., BAI, L., SCHNEIDER, S., ZHAO, G., STILLMAN, B., SPECK, C. & LI, H. 2017. Cryo-EM structure of Mcm2-7 double hexamer on DNA suggests a lagging-strand DNA extrusion model. *Proceedings of the National Academy of Sciences of the United States of America,* 114**,** e9529-e9538.

OH, D. B., KIM, Y. G. & RICH, A. 2002. Z-DNA-binding proteins can act as potent effectors of gene expression in vivo. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 16666-16671.

OUSSATCHEVA, E. A., PAVLICEK, J., SANKEY, O. F., SINDEN, R. R., LYUBCHENKO, Y. L. & POTAMAN, V. N. 2004. Influence of global DNA topology on cruciform formation in supercoiled DNA. *Journal of Molecular Biology,* 338**,** 735-743.

PADEKEN, J., ZELLER, P. & GASSER, S. M. 2015. Repeat DNA in genome organization and stability. *Curr Opin Genet Dev,* 31**,** 12-9.

PAESCHKE, K., BOCHMAN, M. L., DANIELA GARCIA, P., CEJKA, P., FRIEDMAN, K. L., KOWALCZYKOWSKI, S. C. & ZAKIAN, V. A. 2013. Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature,* 497**,** 458-462.

PAESCHKE, K., CAPRA, J. A. & ZAKIAN, V. A. 2011. DNA replication through G-Quadruplex motifs is promoted by the *Saccharomyces cerevisiae* Pif1 DNA Helicase. *Cell,* 145**,** 678-691.

PANAYOTATOS, N. & WELLS, R. D. 1981. Cruciform structures in supercoiled DNA. *Nature,* 289**,** 466-470.

PANTOLIANO, M. W., PETRELLA, E. C., KWASNOSKI, J. D., LOBANOV, V. S., MYSLIK, J., GRAF, E., CARVER, T., ASEL, E., SPRINGER, B. A., LANE, P. & SALEMME, F. R. 2001. High-density miniaturized thermal shift assays as a general strategy for drug discovery. *Journal of Biomolecular Screening,* 6**,** 429-440.

PANYUTIN, I. G., KOVALSKY, O. I. & BUDOWSKY, E. I. 1989. Magnesium-dependent supercoiling-induced transition in (dG)n.(dC)n stretches and formation of a new G-structure by (dG)n strand. *Nucleic Acids Research,* 17**,** 8257-8271.

PARKINSON, G. N., LEE, M. P. H. & NEIDLE, S. 2002. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature,* 417**,** 876-880.

PARVATHANENI, S., LU, X., CHAUDHARY, R., LAL, A., MADHUSUDAN, S. & SHARMA, S. 2017. RECQ1 expression is upregulated in response to DNA damage and in a p53-dependent manner. *Oncotarget,* 8**,** 75924-75942.

PATEL, S. S. & DONMEZ, I. 2006. Mechanisms of helicases. *Journal of Biological Chemistry,* 281**,** 18265-18268.

PHAN, A. T. & PATEL, D. J. 2003. Two-repeat human telomeric d(TAGGGTTAGGGT) sequence forms interconverting parallel and antiparallel G-quadruplexes in solution:

distinct topologies, thermodynamic properties, and folding/unfolding kinetics. *Journal of the American Chemical Society,* 125**,** 15021-15027.

PIERCE, A. J., HU, P., HAN, M., ELLIS, N. & JASIN, M. 2001. Ku DNA end-binding protein modulates homologous repair of double-strand breaks in mammalian cells. *Genes and Development,* 15**,** 3237-3242.

PIKE, J. E., BURGERS, P. M. J., CAMPBELL, J. L. & BAMBARA, R. A. 2009. Pif1 helicase lengthens some okazaki fragment flaps necessitating Dna2 nuclease/helicase action in the two-nuclease processing pathway. *Journal of Biological Chemistry,* 284**,** 25170-25180.

POPURI, V., BACHRATI, C. Z., MUZZOLINI, L., MOSEDALE, G., COSTANTINI, S., GIACOMINI, E., HICKSON, I. D. & VINDIGNI, A. 2008. The human RecQ helicases, BLM and RECQ1, display distinct DNA substrate specificities. *Journal of Biological Chemistry,* 283**,** 17766-17776.

RAMANAGOUDR-BHOJAPPA, R., BYRD, A. K., DAHL, C. & RANEY, K. D. 2014. Yeast Pif1 accelerates annealing of complementary DNA strands. *Biochemistry,* 53**,** 7659-7669.

RANDELL, J. C. W., BOWERS, J. L., RODRIGUEZ, H. K. & BELL, S. P. 2006. Sequential ATP hydrolysis by Cdc6 and ORC directs loading of the Mcm2-7 helicase. *Molecular Cell,* 21**,** 29-39.

REES, D. C., CONGREVE, M., MURRAY, C. W. & CARR, R. 2004. Fragment-based lead discovery. *Nature Reviews Drug Discovery,* 3**,** 660-672.

REZLER, E. M., BEARSS, D. J. & HURLEY, L. H. 2002. Telomeres and telomerases as drug targets. *Current Opinion in Pharmacology,* 2**,** 415-423.

RIBEYRE, C., LOPES, J., BOULÉ, J.-B., PIAZZA, A., GUÉDIN, A., ZAKIAN, V. A., MERGNY, J.-L. & NICOLAS, A. 2009. The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences *in vivo*. *PLoS genetics,* 5**,** e1000475-e1000475.

RODRIGUEZ, R., MILLER, K. M., FORMENT, J. V., BRADSHAW, C. R., NIKAN, M., BRITTON, S., OELSCHLAEGEL, T., XHEMALCE, B., BALASUBRAMANIAN, S. & JACKSON, S. P. 2012. Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nature chemical biology,* 8**,** 301-310.

ROMAN, L. J. & KOWALCZYKOWSKI, S. C. 1989. Characterization of the helicase activity of the Escherichia coli recBCD enzyme using a novel helicase assay. *Biochemistry,* 28**,** 2863-2873.

ROSENTHAL, A. S., DEXHEIMER, T. S., GILEADI, O., NGUYEN, G. H., CHU, W. K., HICKSON, I. D., JADHAV, A., SIMEONOV, A. & MALONEY, D. J. 2013. Synthesis and SAR studies of 5-(pyridin-4-yl)-1,3,4-thiadiazol-2-amine derivatives as potent inhibitors of Bloom helicase. *Bioorganic & Medicinal Chemistry Letters,* 23**,** 5660-5666.

SAALAU-BETHELL, S. M., WOODHEAD, A. J., CHESSARI, G., CARR, M. G., COYLE, J., GRAHAM, B., HISCOCK, S. D., MURRAY, C. W., PATHURI, P., RICH, S. J., RICHARDSON, C. J., WILLIAMS, P. A. & JHOTI, H. 2012. Discovery of an allosteric mechanism for the regulation of HCV NS3 protein function. *Nature chemical biology,* 8**,** 920-925.

SALLMYR, A., TOMKINSON, A. E. & RASSOOL, F. V. 2008. Up-regulation of WRN and DNA ligase III alpha in chronic myeloid leukernia: consequences for the repair of DNA double-strand breaks. *Blood,* 112**,** 1413-1423.

SAMBROOK, J. & RUSSELL, D. W. 2001. Molecular cloning: A laboratory manual. *Molecular cloning: A laboratory manual*. Ed Cold Spring Harbor Laboratory press.

SANDERS, C. M. 2010. Human Pif1 helicase is a G-quadruplex DNA-binding protein with G-quadruplex DNA-unwinding activity. *Biochemical Journal,* 430**,** 119-128.

SANDERS, C. M., KOVALEVSKIY, O. V., SIZOV, D., LEBEDEV, A. A., ISUPOV, M. N. & ANTSON, A. A. 2007. Papillomavirus E1 helicase assembly maintains an asymmetric

state in the absence of DNA and nucleotide cofactors. *Nucleic Acids Research,* 35**,** 6451-6457.

SCHROTH, G. P., CHOU, P. J. & HO, P. S. 1992. Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J Biol Chem,* 267**,** 11846- 11855.

SCHROTH, G. P. & HO, P. S. 1995. Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic acids research,* 23**,** 1977-1983.

SCHULZ, V. P. & ZAKIAN, V. A. 1994. The *saccharomyces* PIF1 DNA helicase inhibits telomere elongation and de novo telomere formation. *Cell,* 76**,** 145-55.

SEO, Y. S., MÜLLER, F., LUSKY, M. & HURWITZ, J. 1993. Bovine papilloma virus (BPV)-encoded E1 protein contains multiple activities required for BPV DNA replication. *Proceedings of the National Academy of Sciences of the United States of America,* 90**,** 702-706.

SHAH, F., LOGSDON, D., MESSMANN, R. A., FEHRENBACHER, J. C., FISHEL, M. L. & KELLEY, M. R. 2017. Exploiting the Ref-1-APE1 node in cancer signaling and other diseases: from bench to clinic. *NPJ precision oncology,* 1**,** 19.

SHARBEEN, G., MCCARROLL, J., GOLDSTEIN, D. & PHILLIPS, P. A. 2015. Exploiting base excision repair to improve therapeutic approaches for pancreatic cancer. *Frontiers in nutrition,* 2**,** 10-10.

SHARMA, S. & BROSH, R. M., JR. 2007. Human RECQ1 is a DNA damage responsive protein required for genotoxic stress resistance and suppression of sister chromatid exchanges. *PloS one,* 2**,** e1297-e1297.

SHARMA, S., STUMPO, D. J., BALAJEE, A. S., BOCK, C. B., LANSDORP, P. M., BROSH JR, R. M. & BLACKSHEAR, P. J. 2007. RECQL, a member of the RecQ family of DNA helicases, suppresses chromosomal instability. *Molecular and Cellular Biology,* 27**,** 1784-1794.

SHELDRICK, G. M. 2008. A short history of SHELX. *Acta Crystallographica Section A,* 64**,** 112-122.

SIDDIQUI-JAIN, A., GRAND, C. L., BEARSS, D. J. & HURLEY, L. H. 2002. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 11593-11598.

SIITONEN, A. H., SOTKASIIRA, J., BIERVLIET, M., BENMANSOUR, A., CAPRI, Y., CORMIER-DAIRE, V., CRANDALL, B., HANNULA-JOUPPI, K., HENNEKAM, R., HERZOG, D., KEYMOLEN, K., LIPSANEN-NYMAN, M., MINY, P., PLON, S. E., RIEDL, S., SARKAR, A., VARGAS, F. R., VERLOES, A., WANG, L. L., KÄÄRIÄINEN, H. & KESTILÄ, M. 2009. The mutation spectrum in RECQL4 diseases. *European Journal of Human Genetics,* 17**,** 151-158.

SIITONEN, H. A., KOPRA, O., KÄÄRIÄINEN, H., HARAVUORI, H., WINTER, R. M., SÄÄMÄNEN, A. M., PELTONEN, L. & KESTILÄ, M. 2003. Molecular defect of RAPADILINO syndrome expands the phenotype spectrum of RECQL diseases. *Human Molecular Genetics,* 12**,** 2837-2844.

SILVESTRE, H. L., BLUNDELL, T. L., ABELL, C. & CIULLI, A. 2013. Integrated biophysical approach to fragment screening and validation for fragment-based lead discovery. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 12984-12989.

SINDEN, R. R. & PETTIJOHN, D. E. 1984. Cruciform transitions in DNA. *Journal of Biological Chemistry,* 259**,** 6593-6600.

SINDEN, R. R., PYTLOS-SINDEN, M. J. & POTAMAN, V. N. 2007. Slipped strand DNA structures. *Frontiers in Bioscience,* 12**,** 4788-4799.

SINGH, M. K., SHWETA, H. & SEN, S. 2016. Dispersed dynamics of solvation in G-quadruplex DNA: comparison of dynamic stokes shifts of probes in parallel and antiparallel quadruplex structures. *Methods and Applications in Fluorescence,* 4**,** 034009.

SINGLETON, C. K., KLYSIK, J., STIRDIVANT, S. M. & WELLS, R. D. 1982. Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. *Nature,* 299**,** 312-316.

SINGLETON, M. R., DILLINGHAM, M. S. & WIGLEY, D. B. 2007. Structure and mechanism of helicases and nucleic acid translocases. *Annual Review of Biochemistry,* 76**,** 23-50.

SKORDALAKES, E. & BERGER, J. M. 2003. Structure of the Rho transcription terminator: Mechanism of mRNA recognition and helicase loading. *Cell,* 114**,** 135-146.

SLEDZ, P., SILVESTRE, H. L., HUNG, A. W., CIULLI, A., BLUNDELL, T. L. & ABELL, C. 2010. Optimization of the interligand overhauser effect for fragment linking: Application to inhibitor discovery against *Mycobacterium tuberculosis* pantothenate synthetase. *Journal of the American Chemical Society,* 132**,** 4544-4545.

SNOW, B. E., MATEYAK, M., PADEROVA, J., WAKEHAM, A., IORIO, C., ZAKIAN, V., SQUIRE, J. & HARRINGTON, L. 2007. Murine Pif1 interacts with telomerase and is dispensable for telomere function *in vivo*. *Molecular and Cellular Biology,* 27**,** 1017-1026.

STALLINGS, R. L., TORNEY, D. C., HILDEBRAND, C. E., LONGMIRE, J. L., DEAVEN, L. L., JETT, J. H., DOGGETT, N. A. & MOYZIS, R. K. 1990. Physical mapping of human-chromosomes by repetitive sequence fingerprinting. *Proceedings of the National Academy of Sciences of the United States of America,* 87**,** 6218-6222.

SUGIMURA, T. & MIWA, M. 1994. Poly(ADP-ribose): Historical perspective. *Molecular and Cellular Biochemistry,* 138**,** 5-12.

SULTANA, R., MCNEILL, D. R., ABBOTTS, R., MOHAMMED, M. Z., ZDZIENICKA, M. Z., QUTOB, H., SEEDHOUSE, C., LAUGHTON, C. A., FISCHER, P. M., PATEL, P. M., WILSON III, D. M. & MADHUSUDAN, S. 2012. Synthetic lethal targeting of DNA double-strand break repair deficient cells by human apurinic/apyrimidinic endonuclease inhibitors. *International Journal of Cancer,* 131**,** 2433-2444.

SYED, S., PANDEY, M., PATEL, SMITA S. & HA, T. 2014. Single-molecule fluorescence reveals the unwinding stepping mechanism of replicative helicase. *Cell Reports,* 6**,** 1037-1045.

SYMINGTON, L. S. & GAUTIER, J. 2011. Double-strand break end resection and repair pathway choice. *Annual Review of Genetics,* 45**,** 247-271.

TADOKORO, T., RAMAMOORTHY, M., POPURI, V., MAY, A., TIAN, J., SYKORA, P., RYBANSKA, I., WILSON III, D. M., CROTEAU, D. L. & BOHR, V. A. 2012. Human RECQL5 participates in the removal of endogenous DNA damage. *Molecular Biology of the Cell,* 23**,** 4273-4285.

TAKEBAYASHI, Y., POURQUIER, P., ZIMONJIC, D. B., NAKAYAMA, K., EMMERT, S., UEDA, T., URASAKI, Y., KANZAKI, A., AKIYAMA, S. I., POPESCU, N., KRAEMER, K. H. & POMMIER, Y. 2001. Antiproliferative activity of ecteinascidin 743 is dependent upon transcription-coupled nucleotide-excision repair. *Nat Med,* 7**,** 961-966.

TANAKA, S. & DIFFLEY, J. F. 2002. Interdependent nuclear accumulation of budding yeast Cdt1 and Mcm2-7 during G1 phase. *Nat Cell Biol,* 4**,** 198-207.

THOMSEN, N. D. & BERGER, J. M. 2009. Running in reverse: the structural basis for translocation polarity in hexameric helicases. *Cell,* 139**,** 523-534.

TODD, A. K., JOHNSTON, M. & NEIDLE, S. 2005. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Research,* 33**,** 2901-2907.

TOTH, E. A., LI, Y., SAWAYA, M. R., CHENG, Y. F. & ELLENBERGER, T. 2003. The crystal structure of the bifunctional primase-helicase of bacteriophage T7. *Molecular Cell,* 12**,** 1113-1123.

TRAN, P. L. T., POHL, T. J., CHEN, C.-F., CHAN, A., POTT, S. & ZAKIAN, V. A. 2017. PIF1 family DNA helicases suppress R-loop mediated genome instability at tRNA genes. *Nature communications,* 8**,** 15025-15025.

TREISMAN, R. & AMMERER, G. 1992. The SRF and MCM1 transcription factors. *Current opinion in genetics & development,* 2**,** 221-226.

TUTEJA, N. & TUTEJA, R. 2004. Prokaryotic and eukaryotic DNA helicases. *European Journal of Biochemistry,* 271**,** 1835-1848.

UCHIUMI, F., SEKI, M. & FURUICHI, Y. 2015. Helicases and human diseases. *Frontiers in Genetics,* 6**,** 39.

UMATE, P., TUTEJA, N. & TUTEJA, R. 2011. Genome-wide comprehensive analysis of human helicases. *Communicative & integrative biology,* 4**,** 118-37.

VAGIN, A. & TEPLYAKOV, A. 1997. MOLREP: an automated program for molecular replacement. *Journal of Applied Crystallography,* 30**,** 1022-1025.

VARE, D., GROTH, P., CARLSSON, R., JOHANSSON, F., ERIXON, K. & JENSSEN, D. 2012. DNA interstrand crosslinks induce a potent replication block followed by formation and repair of double strand breaks in intact mammalian cells. *DNA Repair,* 11**,** 976-985.

WANG, G. L., CHRISTENSEN, L. A. & VASQUEZ, K. M. 2006. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 2677-2682.

WANG, G. L. & VASQUEZ, K. M. 2004. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 13448-13453.

WATSON, J., HAYS, F. A. & HO, P. S. 2004. Definitions and analysis of DNA Holliday junction geometry. *Nucleic Acids Research,* 32**,** 3017-3027.

WATSON, J. D. & CRICK, F. H. C. 1953. Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. *Nature,* 171**,** 737-738.

WILLIAMSON, J. R., RAGHURAMAN, M. K. & CECH, T. R. 1989. Monovalent cation-induced structure of telomeric DNA: The G-quartet model. *Cell,* 59**,** 871-880.

WONG, B., CHEN, S., KWON, J.-A. & RICH, A. 2007. Characterization of Z-DNA as a nucleosome-boundary element in yeast *Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 2229-2234.

WONG, I. N., SAYERS, J. R. & SANDERS, C. M. 2013. Characterization of an unusual bipolar helicase encoded by bacteriophage T5. *Nucleic Acids Research,* 41**,** 4587-4600.

WU, C. G. & SPIES, M. 2016. G-quadruplex recognition and remodeling by the FANCJ helicase. *Nucleic Acids Research,* 44**,** 8742-8753.

WU, Y., SHIN-YA, K. & BROSH, R. M., JR. 2008. FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Molecular and Cellular Biology,* 28**,** 4116-4128.

WU, Y., SUHASINI, A. N. & BROSH, R. M. 2009. Welcome the family of FANCJ-like helicases to the block of genome stability maintenance proteins. *Cellular and Molecular Life Sciences,* 66**,** 1209-1222.

WYATT, P. G., WOODHEAD, A. J., BERDINI, V., BOULSTRIDGE, J. A., CARR, M. G., CROSS, D. M., DAVIS, D. J., DEVINE, L. A., EARLY, T. R., FELTELL, R. E., LEWIS, E. J., MCMENAMIN, R. L., NAVARRO, E. F., O'BRIEN, M. A., O'REILLY, M., REULE, M., SAXTY, G., SEAVERS, L. C. A., SMITH, D. M., SQUIRES, M. S., TREWARTHA, G., WALKER, M. T. & WOOLFORD, A. J. A. 2008. Identification of N-(4-piperidinyl)-4-(2,6-dichlorobenzoylamino)-1H- pyrazole-3-carboxamide (AT7519), a novel cyclin dependent kinase inhibitor using fragment-based X-ray crystallography and structure based drug design. *Journal of Medicinal Chemistry,* 51**,** 4986-4999.

WYRICK, J. J., APARICIO, J. G., CHEN, T., BARNETT, J. D., JENNINGS, E. G., YOUNG, R. A., BELL, S. P. & APARICIO, O. M. 2001. Genome-wide distribution of ORC and MCM proteins in *S-cerevisiae*: High-resolution mapping of replication origins. *Science,* 294**,** 2357-2360.

XU, H. Q., DEPREZ, E., ZHANG, A. H., TAUC, P., LADJIMI, M. M., BROCHON, J. C., AUCLAIR, C. & XI, X. G. 2003. The *Escherichia coli* RecQ helicase functions as a monomer. *Journal of Biological Chemistry,* 278**,** 34925-34933.

YI, C. Q. & HE, C. 2013. DNA repair by reversal of DNA damage. *Cold Spring Harbor Perspectives in Biology,* 5**,** a012575.

ZEGA, A. 2017. NMR methods for identification of false positives in biochemical screens. *Journal of Medicinal Chemistry,* 60**,** 9437-9447.

ZHANG, D. H., ZHOU, B., HUANG, Y., XU, L. X. & ZHOU, J. Q. 2006. The human Pif1 helicase, a potential *Escherichia coli* RecD homologue, inhibits telomerase activity. *Nucleic Acids Research,* 34**,** 1393-1404.

ZHAO, J., BACOLLA, A., WANG, G. & VASQUEZ, K. M. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences,* 67**,** 43-62.

ZHAO, X., LIU, B., YAN, J., YUAN, Y., AN, L. & GUAN, Y. 2014. Structure variations of TBA G-quadruplex induced by 2'-O-methyl nucleotide in K^+ and Ca^{2+} environments. *Acta Biochimica Et Biophysica Sinica,* 46**,** 837-850.

ZHOU, J. Q., MONSON, E. K., TENG, S. C., SCHULZ, V. P. & ZAKIAN, V. A. 2000. Pif1p helicase, a catalytic inhibitor of telomerase in yeast. *Science,* 289**,** 771-774.

ZHOU, R., ZHANG, J., BOCHMAN, M. L., ZAKIAN, V. A. & HA, T. 2014. Periodic DNA patrolling underlies diverse functions of Pif1 on R-loops and G-rich DNA. *eLife,* 3**,** e02190-e02190.

ZHOU, X., REN, W., BHARATH, SAKSHIBEEDU R., TANG, X., HE, Y., CHEN, C., LIU, Z., LI, D. & SONG, H. 2016. Structural and functional insights into the unwinding mechanism of *bacteroides sp*Pif1. *Cell Reports,* 14**,** 2030-2039.