doi:10.20944/preprints202103.0477.v1

Article

Phylogenetic diversity of Lhr proteins and biochemical activities of the Thermococcales aLhr2 DNA/RNA helicase

Mirna Hajj[1,2†], Petra Langendijk-Genevaux[1†], Manon Batista[1], Yves Quentin[1], Sébastien Laurent[3], Ziad Abdel Razzak[2], Didier Flament[3], Hala Chamieh[2], Gwennaele Fichant[1*], Béatrice Clouet-d'Orval[1*] and Marie Bouvier[1]

[1] Laboratoire de Microbiologie et de Génétique Moléculaires, UMR5100, Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, Université Paul Sabatier, F-31062 Toulouse and France
[2] Laboratory of Applied Biotechnology, Azm Center for Research in Biotechnology and its application, Lebanese University, Tripoli, Lebanon
[3] Laboratoire de Microbiologie des Environnements Extrêmes, UMR6197, Ifremer, Université de Bretagne Occidentale, CNRS, F-29280 Plouzané, France
† Co-first authors
* Correspondence: Corresponding authors

Abstract Helicase proteins are known use the energy of ATP to unwind nucleic acids and to remodel protein-nucleic acid complexes. They are involved in almost every aspect of the DNA and RNA metabolisms and participate in numerous repair mechanisms that maintain cellular integrity. The archaeal Lhr-type proteins are SF2 helicases that are mostly uncharacterized. They have been proposed to be DNA helicases that act in DNA recombination and repair processes in Sulfolobales and Methanothermobacter. In Thermococcales, a protein annotated as an Lhr2 protein was found in the network of proteins involved in RNA metabolism. To this respect, we performed in-depth phylogenomic analyses to report the classification and taxonomic distribution of Lhr-type proteins in Archaea, and to better understand their relationship with bacterial Lhr. Furthermore, with the goal of envisioning the role(s) of aLhr2 in Thermococcales cells, we deciphered the enzymatic activities of aLhr2 from *Thermococcus barophilus* (*Tbar*). We showed that *Tbar*-aLhr2 is a DNA/RNA helicase with a significant annealing activity that is involved in processes dependent of DNA and RNA transactions.

KEYWORDS: SF2 helicases; aLhr2 helicases; Archaea; RNA metabolism and DNA repair; Thermococcales

1. Introduction

Helicases are proteins that unwind nucleic acids and remodel protein-nucleic acid complexes in a wide spectrum of cellular tasks. DNA helicases are critical in maintaining cellular integrity by playing important roles in DNA replication, recombination and repair. RNA helicases are likewise fundamental by orchestrating transcription, RNA processing, ribosome biogenesis, translation and RNA turnover. Helicases are classified into 6 superfamilies (SF1-6) [1]. The SF1-6 share a common helicase core with a set of helicase signature motifs. The SF2 is the largest and most diverse group of helicases with more than ten families. SF2 members are non-hexameric helicases that share a conserved helicase core with nine characteristic motifs and that often contain N- and/or C-terminal accessory domains involved in the regulation of their activities [2,3]. The core provides the active site for ATP hydrolysis, binds nucleic acid and carries a basal unwinding activity. Although ATP-dependent unwinding of nucleic acid duplexes is their hallmark reaction, not all helicases catalyse unwinding *in vitro*, and disrupt duplexes *in vivo* [4,5]. Among SF2 helicases, the Lhr (Large helicase related) proteins are scarcely characterized. They are found in some Bacteria but are ubiquitous in Archaea [4,5]. To date, no homologs of Lhr proteins have been reported in Eukarya.

In Bacteria, Lhr proteins are mostly prevalent in Proteobacteria and Actinobacteria. Lhr proteins from *Pseudomonas putida* (*Pput*), *Escherichia coli* (*Ecol*) and *Mycobacterium smegmatis* (*Msme*) are among the few helicases that were characterized [6–9]. The 1507 amino acids (aa) *Msme*-Lhr is the founding member of the Lhr helicase family [9]. The crystal structure of *Msme*-Lhr restricted to the first 856 aa was solved and uncovered a specific structural domain organization also referred to as the "Lhr-Core": two RecA domains in tandem (RecA1 and RecA2), a winged-helix (WH) motif and a domain annotated as Domain 4 whose function is still unknown. Interestingly, the WH displays a similar fold to the one observed in Hjm and RecQ (also called Hel308) DNA helicases [9,10]. While *Pput*-Lhr is restricted to the "Lhr-Core", *Msme*-Lhr and *Ecol*-Lhr have an additional C-terminal domain [6–9] (Figure 1).

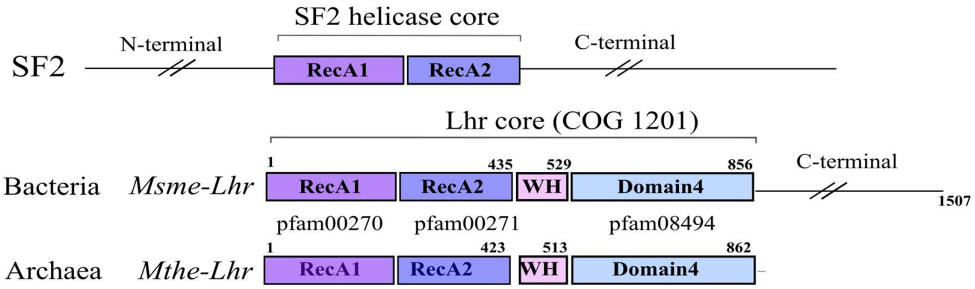

Figure 1. Overall domain organization of SF2 helicase superfamily and Lhr-like subfamily. Lhr proteins of the *Mycobacterium smegmatis* bacterium and the *Methanothermobacter thermautotrophicus* archaeon are shown. The "Lhr-core" (COG1201) is composed of the RecA1 (PF00270) and RecA2 (PF00271) domains of the SF2 helicase core, the Winged-Helix domain (WH) and the Domain4 (PF084494) of unknown function that is specific to the Lhr proteins.

The studies characterizing the biochemical activities and the functions of bacterial Lhr proteins have mainly revealed a role of Lhr helicases in DNA repair. Nonetheless, some of their properties suggest that Lhr may also participate in RNA processing. *In vivo*, the gene encoding *Msme*-Lhr was shown to be upregulated when cells are exposed to DNA damaging agents [11,12]. Regarding *Ecol*-Lhr, though its deletion does not increase cell-sensitivity to UV or H_2O_2 [8], Cooper *et al.* demonstrated a synthetic genetic interaction with RadA, a RecA-related protein involved in the processing of recombination intermediates [13]. *In vitro*, *Pput*-Lhr and *Msme*-Lhr helicases were shown to have DNA-dependent ATPase and ATP-dependent 3′-to-5′ translocase activities. While *Pput*-Lhr exhibits no preference for DNA:DNA or DNA:RNA duplex [7], *Msme*-Lhr prefers to unwind DNA:RNA duplex in which the displaced strand is RNA [6]. Finally, the importance of *Ecol*-Lhr rose from its occurrence in a cluster with the gene encoding RNase T, a ribonuclease involved in the maturation of stable RNAs, as well as in DNA repair pathways [8]. This interaction occurs at the transcriptional level, as the Lhr and RNase T are co-transcribed, but no interaction at the protein level was reported yet either *in vitro* or *in vivo*.

In Archaea, some genome annotations record two types of Lhr proteins, called here, aLhr1 and aLhr2. The aLhr1 and aLhr2 exhibit a "Lhr-Core" domain organization [14]. aLhr1 has an additional cysteine-rich motif at its C-terminal end. Only few Lhr from Sulfolobales (TACK) and Methanobacteriales (Euryarchaea) have been studied [15,16]. Lhr of *Sulfolobus islandicus* (SiRe_1605) was found to be important for the transcription of genes in nucleotide metabolism and DNA repair [17]. Monomeric Lhr of *Sulfolobus solfataricus*, also known as Hel112, was characterized *in vitro* as an ATP-dependent DNA helicase with a 3′-5′ polarity and a preference for forked DNA substrates [16]. Lhr of *Sulfolobus acidocaldarius* (saci_1500, also named RecQ-like helicase) was found to be important for DNA repair after UV-induced stress [18]. *In vitro*, Lhr of *Methanothermobacter thermautotrophicus* (*Mthe*) was also found to have a 3′-5′ directional DNA translocase activity and to act on forked DNA structures. In a genetic assay, its expression gave a phenotype identical to the DNA helicases Hel308 and RecQ involved in replication-coupled DNA repair [15]. In addition, aLhr2 of *Pyrococcus abyssi* (*Paby*) was detected in the interaction network of proteins implicated in DNA replication and repair [19]. Recently, we also spotted *Paby*-aLhr2 as a partner of players in RNA metabolism. Indeed, *Paby*-aLhr2 was identified in the interaction network of the RNA helicase ASH-Ski2 together with the 5′-3′ and 3′-5′ RNA degradation machineries, aRNase J and the RNA exosome [20] (Table S1). This questions the role of the aLhr2 proteins in Thermococcales.

In this study, we highlighted archaeal Lhr-type proteins as ubiquitous enzymes by revisiting the Lhr-type proteins landscape, using in-depth phylogenomic analyses. We identified six distinct phylogenetic groups of Lhr proteins, three in Archaea and three in Bacteria. We also defined to which phylogenetic group, each of the experimentally studied Lhr helicases belong to. To go further in understanding the relevance of the archaeal aLhr2 group members in DNA and/or RNA metabolism, we characterized the enzymatic properties of aLhr2 from the Thermococcales *Thermococcus barophilus* (*Tbar*-aLhr2). Our results allowed us to propose that *Tbar*-aLhr2 is a DNA/RNA helicase with significant annealing activity that acts on DNA:RNA hybrids and on RNA:RNA duplexes.

2. Materials & Methods

2.1. Building Lhr-type dataset

Completely sequenced genomes of 286 Archaea and 3769 Bacteria showing a high level of annotation were downloaded from EBI (http://www.ebi.ac.uk/genomes/). The complete genomes of these 4055 strains, their proteomes and EMBL features were managed with an in house MySQL database. Moreover, we have performed the annotation of the protein sequences of these genome against the conserved domain database downloaded from the NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) using the *rpsblast* program [21]. The *hmmscan* program [22] was used to annotate proteins with Pfam (32.0) domains. To avoid redundancies due to multiple repetitions of strains of the same species, we have retained only one strain per species. Conversely, in order to obtain a better coverage of Archaea's diversity, 75 Asgard proteomes were retrieved from UniProt, although they do not have the same sequencing and annotation quality as the other archaeal genomes (35 Lokiarchaeota, 27 Thorarchaeota, 12 Heimdallarchaeota and one Odinarchaeota). An initial sample of 1381 Lhr protein candidates were identified using the COG1201 (Lhr-like helicase) annotation performed by *rpsblast*. In order to eliminate the false positives while keeping the most divergent sequences, we have set an e-value threshold ≤ $1e^{-04}$ associated with an alignment covering of at least 30% of the COG.

To identify Lhr-like families, we performed all-against-all *blastp* comparisons of our initial set of proteins with default parameters, except *-max_target_seqs* which was set to 1381 sequences. The results were filtered to retain only the best bi-directional hits between proteins of different species. Protein relationships were then converted into a graph in which the vertices represent protein sequences, and the edges represent their relationships [23]. The edges were weighted by the average pairwise $-\log_{10}$ *E*-value. The graph was further processed by a graph-partitioning approach based on the Markov Clustering algorithm (MCL, [24]). The inflate factor (IF) value is an important parameter of MCL as it regulates the cluster granularity. We tested several IF values (from 2 to 6) and a partitioning into 9 stable classes was observed starting from an IF ≥ 4. Classes 1 to 9 have a size of 615, 344, 220, 106, 68, 22, three, one, and one sequences, respectively. One protein from *Pseudomonas viridiflava* (*A0A1Y6JKR4_PSEVI)* was not classified and was discarded as false positive since it shares only a small region of similarity with PF00271.

In order to facilitate phylogenetic reconstructions while preserving the diversity of sequences in the original sample, we have represented sequences with more than 70% identity by a single sequence, the medoid. To achieve this, the edges of the previous graph with an identity < 70% were removed. This pruned graph was further processed by MCL to identify groups of closely related sequences (identity ≥ 70). This identified 352 groups (including 27 Asgard clusters) composed of a unique sequence and 111 groups composed of many closely related sequences. For each of the 111 groups, we computed the medoid, *i.e.*, the sequence with the minimal average dissimilarity to all the other proteins in the group. We added the constraint that its length should be close to the median length of all sequences of the group. This resulted in a set of 463 proteins composed of 352 unique sequences and 111 medoids. Eight unique sequences were discarded as they did not have the PF00270 (DEAD) and/or the PF00271 (Helicase_C) domains. The aLhr1 and aLhr2 sequences of *T. barophilus* belong to two groups of closely related sequences. As a result of our sample size reduction process, these two sequences were not selected as medoid. The selected medoids are aLhr1 from *Thermococcus profundus* (TproA01.ASJ02541.1) and aLhr2 from *Methanocaldococcus jannaschii* (MjanA01. AAB98279.1). As aLhr2 from *T. barophilus* is the subject of this experimental study, the *Tbar*-aLhr1 (TbarA01.ADT83607.1) and *Tbar*-aLhr2 (TbarA01.ADT83510.1) protein sequences were added back to our sample. The tree of Figure 2 shows that each sequence of *T. barophilus* has a direct common ancestor with their respective medoid. Our final sample contains 457 sequences (Supplementary Tables S2A & S2B).

Since the Sfth helicases appear to be the closest related family of Lhr helicases, both families sharing a common ancestor [14], we used Sfth sequences to root our Lhr family tree. To build the Sfth sample, we applied a similar strategy as the one described above for Lhr protein identification. Sfth protein candidates were identified using the COG1205 annotation performed by *rpsblast*. Proteins that do not possess the three expected domains (PF00270 DEAD; PF09369 DUF1998; PF00271 Helicase_C) were excluded. Using an identity threshold of 55%, we obtained 24 medoid sequences further used as representatives of this family (Supplementary Table S2C).

2.2. *Alignment of the core helicase domain*

In order to eliminate the variability of the N- and C-term regions of the proteins, we have extracted the central domain of the SF2 helicase core, *i.e.*, the RecA1 and RecA2 regions (Figure1).The coordinates of the alignment of the sequences with the PF00270 and PF00271 domains are used to extract both regions, which are then merged for each sequence. These sequences were aligned with *mafft* [25] (parameters: --reorder –localpair --maxiterate 1000). In order to improve the quality of the alignments and to keep as much information as possible, we used the *divvier* method [26] with the option *-divvy* (full divvying) and *-mincol* 4. This strategy was applied to the dataset containing only Lhr sequences and to the dataset composed of Lhr and Sfth sequences.

2.3. *Protein family and archaeal species trees*

The best-fit amino acid substitution model for the data were selected with *modeltest-ng* [27] and the phylogenetic trees were inferred using the *iq-tree* software [28]. The same best model was selected for both datasets (-m LG4M+I). Branch supports were measured with ultra-fast bootstrap approximation (*-bb* 1000) and single branch test (*-alrt* 1000). The trees were annotated and visualized with the online tool Interactive Tree Of Life (iTOL, https://itol.embl.de) [29].

To construct the archaeal species tree we used the 122 markers that have been identified as reliable for phylogenetic inference [30]. We first thought of including the Asgard species in the tree but as their genomes are mostly partial, too many markers were missing for this to be feasible. Therefore the tree was built by taking into account 219 archaeal species. The set of 122 protein markers is characterized by HMM profiles from the Pfam (v27) and the TIGRFAMs (v15.0) databases. For each genome included in this study, the proteins were identified by using each Pfam entry as query with *hmmsearch* program from the HMMER3 package with the *--cut_tc* (trusted cutoff) parameter (http://hmmer.org/; [22]). The *hmmsearch* output domain file was parsed to extract for each genome and for each HMM profile the best protein hit. Protein alignments with HMM profiles were merged for each marker. We thus obtained 122 sequence alignments. The columns of the alignments that had a high deletion frequency were removed with *trimal* (*-gt* 0.1) [31]. The quality of the alignments was estimated using the *t-coffee* transitive consistency score (TCS) [32]. The analysis of the results obtained on each alignment allowed (i) to eliminate sequences with outlier TCS values and (ii) to discard two alignments (PF04104.9 and PF01990.12) with a low overall TCS value (TCS < 65). The resulting 120 marker alignments were concatenated and the tree was inferred with *fasttree* [33] under the LG+GAMMA model and branch support values were determined using 100 non-parametric bootstrap replicates. The tree was rooted on DPANN Archaea according to [34].

2.4. Genomic context analysis

We extracted the proteins encoded by the genes located at less than 4000 bp upstream and downstream from the predicted aLhr2 genes. To obtain a functional characterization and classification of these proteins, they were annotated by *hmmscan* with TIGR HMM profiles. iTOL was used to associate these gene neighborhoods to the species tree (DATASET_DOMAINS option).

2.5. Expression vectors

The supplementary Table S3 summarizes the oligonucleotides used in this study. All constructions were obtained by assembling PCR fragments using InFusion® cloning kit (Takara). Using an appropriate set of oligonucleotides, pET11b (untagged protein) vector was linearized with the PrimeSTAR Max DNA polymerase (Takara), and the coding sequence of *T. barophilus* aLhr2 (TERMP_00533) and *P. abyssi* Hel308 (PAB_0592) were amplified from genomic DNA with the Phusion High-Fidelity DNA polymerase (ThermoFisherScientific). The pET11b vectors expressing the aLhr2-T215A, aLhr2-W577A and aLhr2-I512A variants were generated by site-directed mutagenesis of their wildtype counterpart with appropriate sets of oligonucleotides using the QuikChange II XL Kit (Stratagene). The pET11b vectors expressing the truncated aLhr2-ΔDom4 and the Domain 4 by itself (aLhr2-Dom4) were constructed by reverse PCR on the pET11b-aLhr2-WT using specific phosphorylated oligonucleotides and by DNA ligation (T4 DNA ligase).

2.6. Purification of Tbar-aLhr2 recombinant proteins.

E. coli BL21-CodonPlus (DE3) cells freshly transformed with pET11b-aLhr2, pET11b-aLhr2-T215A, pET11b-aLhr2-W577A, pET11b-aLhr2-I512A, pET11b-aLhr2-ΔDom4 and pET11b-aLhr2-Dom4 vectors were grown in 400mL of LB medium at 37°C. Protein production was induced at OD_{600nm} 0.8 with 0.2mM IPTG. After 3h of induction at 30°C, the cells were collected, suspended in 10mL of lysis buffer (50mM NaPhosphate, 300mM NaCl, 10mM Imidazole) supplemented with 1mg.mL^{-1} of lysozyme and a mix of EDTA-free protease inhibitor (cOmplete™, Roche), and lysed by sonication (4x[5*10s], 50% cycle, VibraCell Biolock Scientific). The cleared extracts, obtained by centrifuging the crude extracts (20,000g, 4°C, 20min), were treated with a mix of RNase A (20µg.mL^{-1}), RNase T1 (1U.µL^{-1}) and DNase I (20µg.mL^{-1}) containing 10mM of $MgCl_2$ for 30min at 37°C. After a heating step at 70°C for 20min, the extracts were furthered clarified by centrifugation (20,000g, 4°C, 20min). First, the recombinant proteins were purified from the soluble fractions to near homogeneity using FPLC (Fast Protein Liquid Chromatography, Äkta-purifier10, GE-Healthcare) and specific columns (GE Healthcare): for wild type aLhr2, the punctual mutants and aLhr2-Dom4 by a cation exchange chromatography (Hitrap SP HP); for aLhr2-ΔDom4 by a heparin column (Heparin FF) with a linear gradient of NaCl (300mM to 1M). Then, all recombinant proteins were loaded on a size-exclusion HiLoad 16/60 Superdex 200 PG column in 50mM HEPES pH 6.8, 300mM NaCl, 10% glycerol buffer.

2.7. Preparation of radiolabelled nucleic acid substrate

The 26-nt RNA (RNA_{26}) and all the DNA (DNA_{26}, DNA_{31}, DNA_{50} and DNA_{59}) oligonucleotides were synthesized by Eurofins. The 50-nt RNA substrate (RNA_{50}) was obtained by *in vitro* transcription from a PCR fragment where DNA_{50} is fused to the T7

doi:10.20944/preprints202103.0477.v1

promoter using the MEGAscript kit (Ambion). The DNA and RNA substrates were 5'-end radiolabelled using T4 polynucleotide kinase and γ-^{32}P-ATP. To prepare nucleic acid duplexes, the short DNA or RNA oligonucleotide was radiolabelled, mixed with an unlabelled DNA or RNA complementary strand at a 1:1 molar ratio (100nM each), incubated for 5 min at 95°C in 1X SSC buffer, and then slowly cooled at room temperature. The nucleotide sequences of all the substrates used in this study are given in Supplementary Table S4.

2.8. ATPase hydrolysis assay

500nM of recombinant protein were mixed with 5nM of DNA_{50} or DNA_{59}:DNA_{31} substrates in a 50mM Hepes pH 7.5, 50mM KCl, 5mM $MgCl_2$, 2mM DTT buffer and preincubated for 10min at 65°C. 2 mM ATP and 0.85µCi γ-32P-ATP were added at 0 time point. The kinetic was performed at 65°C. At indicated time, aliquots were spotted directly on TLC plate (PEI-cellulose, Nagel). TLC were developed with 0.25M KH_2PO_4. Radioactive signals were measured using a PhosphorImager device (Typhoon Trio) and quantified with MultiGauge software (FujiFilm). The percentage of ATP versus ADP was plotted over time. Identical experiments were performed with 5nM of DNA_{50} or RNA_{50} with a range of ATP concentration (0.025, 0.05, 0.1, 1, and 2mM) in triplicates. The plots were derived using GraphPad Prism 7 software.

2.9. Nucleic acid binding assay

Double filtration binding assays were performed with range of protein concentration from 0 to 350nM and 0.5nM of ^{32}P-labelled RNA or DNA substrate using a Slot blot device (Amersham Biosciences). The protein was preincubated for 10min at 65°C in 25mM Tris-HCl pH 8, 50mM NaAc, 5mM $MgCl_2$, 2.5mM β-Mercaptoethanol. After adding the substrate, the reactions were incubated 15min at 30°C. Free nucleic acids were separated from nucleoprotein complexes on double filtration system using Nylon and Nitrocellulose membranes (Amersham™ Hydond-N and Protran, respectively). Radioactive signals were measured using a PhosphorImager device and quantified with MultiGauge software. The apparent dissociation constant K_D were calculated using GraphPad Prism 7 software.

2.10. Helicase (unwinding) assay

The unwinding assays were done with 250nM of protein, 5nM of α-^{32}P-labeled nucleic acid duplex and 200-fold excess of the unlabelled oligo trap (1µM). The protein was preincubated separately for 5min at 65°C in 25mM Tris-HCl pH 8, 50mM NaAc, 2.5mM β-Mercaptoethanol, 25mM $MgCl_2$, 25mM ATP. After addition of the recombinant protein (250nM), the reaction mixtures were incubated at 65 °C for the indicated times and then quenched with 0.5% SDS, 40mM EDTA, 0.5mg.mL^{-1} Proteinase K, 0.1% Bromophenol blue, and 20% glycerol. The reaction products were separated on a native 8% polyacrylamide gel (1X TBE, 0.1% SDS) by electrophoresis in 1X TBE (200 Volts, 90 min). Radioactive signals were measured using a PhosphorImager device and quantified with MultiGauge software. All assays were repeated at least three times.

2.11. Strand-annealing assay

5nM of radiolabelled substrates and 250nM of recombinant protein were preincubated separately for 5min at 65°C in 25mM Tris-HCl pH 8, 50mM NaAc and 2.5mM β-Mercaptoethanol. The reactions were started by mixing the protein and nucleic acid samples. After incubation at 65°C, samples of 5µL were withdrawn at the indicated time points. The reactions were quenched and analysed as described in section 2.10. All assays were independently repeated at least three times.

3. Results

3.1. Phylogenomic studies of Lhr-type helicases in Archaea & Bacteria

Our initial Lhr library was composed of 1380 proteins that were identified by similarity search against the COG1201 profile of the COG database that covers the "Lhr core" organization of Lhr helicases, *i.e.*, the two conserved RecA1 and RecA2 domains, the winged-helix motif and the Domain 4 (Figure 1). To explore the family organization, the protein relationships were converted into a graph that was further processed with MCL to identify groups of Lhr proteins. Nine groups were obtained including five main classes of 615, 344, 220, 106, 68 sequences respectively. The aLhr2 sequences from *P. abyssi* (SP: Q9UZM4) and *T. barophilus* (SP: F0LJX3, TbarA01.ADT83510.1) belong to the MCL class 1 as well as the characterized Lhr from *S. solfataricus* (SP: P95949, SSO0112), *S. acidocaldarius* (SP: Q4J8R1, saci_1500) and *M. thermautotrophicus* (SP: O27830, MTH_1802). The aLhr1 sequences from *P. abyssi* (SP: Q9V0H2) and *T. barophilus* (SP: F0LKE9, TbarA01.ADT83607.1) as well as the studied sequence of *S. islandicus* (SiRe_1605) are found in the MCL class 3 (Supplementary Table S2A). In our sample, *S. islandicus* is represented by the strain L.S.2.15 (SislA01.ACP36165.1, SP: C3MR20) whereas the REY15A strain is the one that has been functionally studied; aLhr1 from L.S.2.15 presents 99.8% of identity with its REY15A ortholog. For the bacterial Lhr helicases, the proteins from *M. smegmatis* (SP: A0QT91) and *E. coli* (SP: P30015) have been both found in MCL class 1 while the one of *P. putida* (SP: Q88NV1, PP_1103) belongs to MCL class 2 (Supplementary Table S2B).

To go further, we computed two phylogenetic trees. To avoid bias due to the overrepresentation of closely related species in public databases, sequences showing more than 70% of identity were displayed by a representative sequence (see Material and Methods). This facilitates phylogenetic reconstructions while preserving the diversity of sequences in the original sample. Our reduced sample contains 457 proteins. In order to eliminate the variability of the C-terminal regions of the Lhr proteins and to allow the comparison with the Sfth helicases, the multiple alignments were performed with the SF2 helicase core composed of the RecA1/RecA2 domains. A first Lhr family tree was rooted by adding 24 reference sequences of the Sfth helicase family (Supplementary Figure S1) that appears to be the closest related family of the Lhr helicases [14]. Then, we constructed a second tree on the sole Lhr sequences that was rooted by using the most external Lhr subtree identified above (Figure 2). The topologies of the trees are consistent with the MCL classes obtained on complete sequences, with six subtrees clearly identified.

The first colour-coded ring around the trees indicates the bacterial (purple), the archaeal (green) and the Asgard (yellow) genomes, respectively. The second colour-coded ring figures the MCL groups as stated in the figure legend. The colour-coded pie slices indicate the boundaries of each subtree. Two subtrees, one corresponding to MCL class 3 and the other, smaller, corresponding to MCL class 6, contain only archaeal sequences. A third subtree corresponding to MCL class 2 encloses only bacterial sequences. Based on the tree topologies, the MCL class 1 can be clearly subdivided into two subtrees, one containing only bacterial sequences and the other only archaeal sequences. Finally, the last subtree regroups the sequences belonging to the remaining MCL classes. Except for the MCL classes 7, 8 and 9 that contains only few sequences (one to three), we can notice that the MCL classes 4 and 5 correspond to groups of sequences that share a common ancestor. The location of the MCL class 2 subtree is different in the two trees. In the tree rooted with Sfth, it shares a common ancestor with the bacterial subtree of MCL class 1 (Supplementary Figure S1) while in the tree based on Lhr sequences alone (Figure 2), it forms a group external to the MCL class 1 sequences that are under the same ancestor node. Since the branch shows a weak bootstrap support in the Sfth tree (SH-like approximate likelihood ratio test (alrt) < 60) (Supplementary Figure S1), we favoured the topology obtained on the sole Lhr proteins (Figure 2). It can be noticed that MCL class 2 subtree presents an acceleration of the rate of evolution that could be at the origin of the instability of its placement in the different trees.

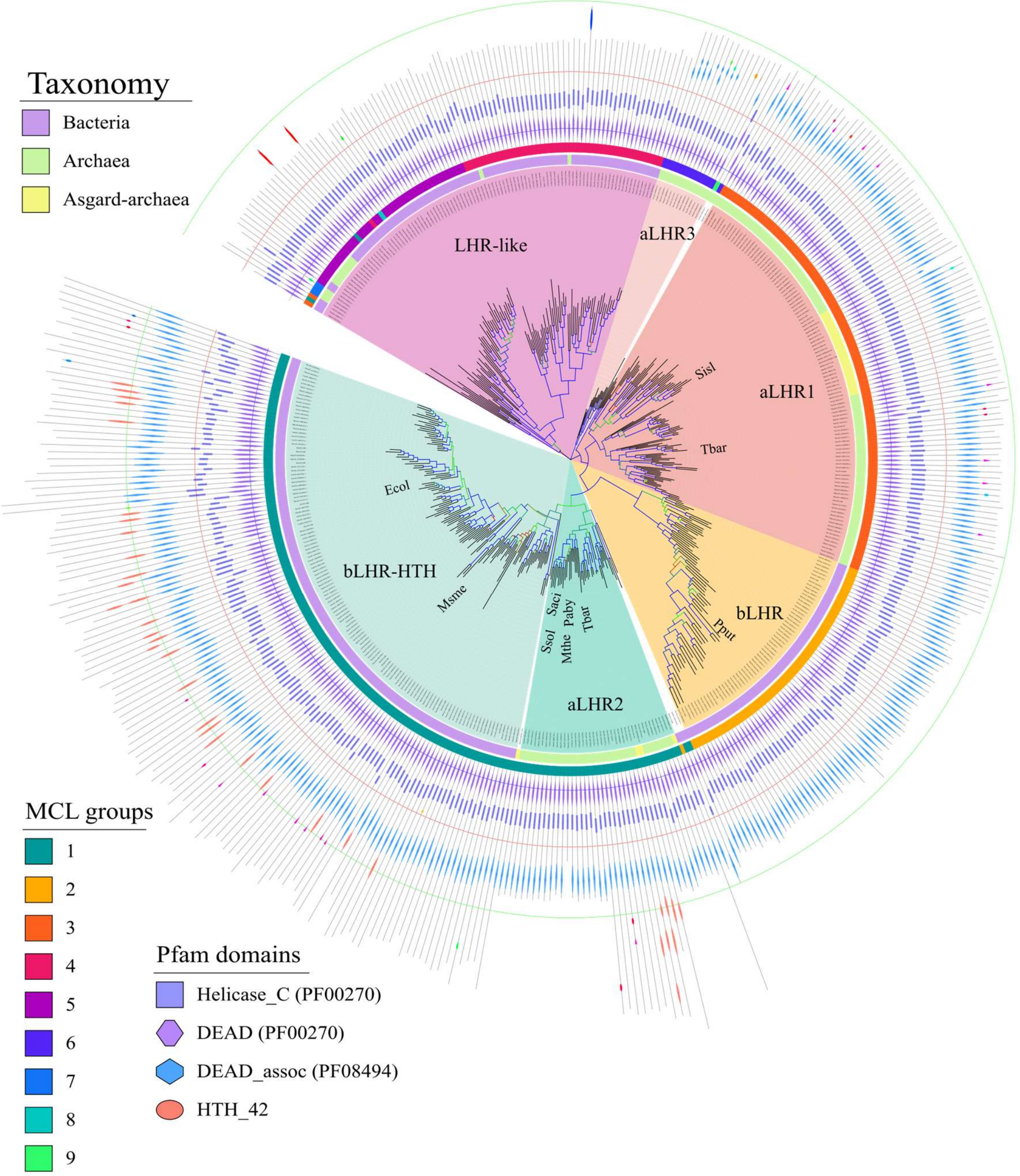

Figure 2. Phylogenetic tree and domain organization of archaeal and bacterial Lhr representative sequences. The tree is rooted to the branch that separates the subtrees Lhr-like and aLhr3. The branches are coloured according to their bootstrap value using a colour gradient from red (bootstrap value of 0) to blue (bootstrap value of 100) with flashy green at midpoint value. The taxonomic origin of the sequences is shown in the first outer ring: purple for Bacteria, light green for all Archaea except Asgard shown in yellow. The MCL subclasses obtained with an IF of 4 are shown on the second outer ring and colour-coded as indicated in the legend panel “MCL groups”. To highlight the different subfamilies, their corresponding subtrees are coloured and the sequences for which the location on the tree and the MCL classification do not correspond are left white. The location of reference sequences are indicated at their respective leaf (Tbar, Sisl) or at the leaf of their representative medoid (Msme, Ecol, Pput, Ssol, Saci, Mthe, Paby). Pfam motif architecture for each member is specified at the circumference of the tree. Their colour code is given in the legend panel “Pfam domains”. The sequence identifiers for each subtree are in Supplementary Tables S2A and S2B. The tree display was obtained with online iTOL [26] (https://itol.embl.de/).

The domain organization based on Pfam profiles of each protein are shown as outer rings on the tree (Figure 2). While they all possess the RecA1 (PF00270 entry) and RecA2 (PF00271 entry) domains that form the helicase core, Domain 4 (PF08494 entry named "DEAD-associated domain") is present in the subtree corresponding to MCL classes 1 to 3 and is missing from the subtree corresponding to the MCL classes 4 to 9. These sequences have a shorter C-terminal region that is not characterized by any conserved domain except for class 6 sequences that have a highly deteriorated Domain 4. The bacterial MCL class 1 sequences have a longer C-terminal region containing an additional HTH_42 domain (PF06224 entry). This is also the case for some archaeal MCL class 1 sequences.

The groups of Lhr helicases that correspond to the different subtrees were named based on the MCL class of the experimentally characterized Lhr proteins and on their domain organization. In Bacteria, we detected three orthologous groups of Lhr proteins that we referred to as: bLhr (MCL class 2) when they are restricted to the "Lhr-core"; bLhr-HTH (bacterial MCL class 1) based on their additional HTH_42 domain at their C-terminal end; and finally Lhr-like as its sequences do not possess Domain 4 (mostly MCL classes 4 and 5). This last group contains some scattered sequences of Archaea that where probably acquired by horizontal gene transfers. To our knowledge, it is the first time that three different groups of bacterial Lhr helicases are reported. In Archaea, we found the already reported aLhr1 (MCL class 3) and aLhr2 (archaeal MCL class 1) groups [14]. We also identified for the first time a third small group that we named aLhr3 (MCL class 6) and that is characterized by a highly deteriorated Domain 4.

To go further the taxonomic distribution of the archaeal Lhr groups was performed (Figure 3). Note that the Asgard with incomplete genomes have not been included in the species tree. aLhr1 members are found in two DPANN genomes out of six, in all TACK except in two Candidatus genomes (*C. nitrosmarinus catalina SPOT01* and *C. nitosopumilus sp. AR2*) for which no Lhr proteins were detected, and in almost all the Euryarchaeota genomes with the exception of the Methanopyraceae, the Methanococcales and the Methanobacteriales. Members of the aLhr2 group are found in the DPANN genomes, in most TACK genomes except in Thaumarchaeota and Thermoproteales and in the majority of Euryarcheota with the exception of the Methanomicrobiales and the Methanosarcinales. It can be noticed that in the genomes of Thermoproteales which do not contain the *alhr2* gene, two paralogous *alhr1* genes are found. Members of the aLhr3 group are only found in Sulfolobales, Desulfurococcales and Acidilobales. Interestingly, these genomes usually also encode a member of the aLhr1 and aLhr2 groups. Finally, two genomes *Candidatus Methanomassiliicoccus intestinalis Issoire-Mx1* (Methanomassiliicoccales) and *Aciduliprofundum sp. MAR08-339* (Aciduliprofundum) encode an Lhr-like protein in addition to aLhr1 and aLhr2.

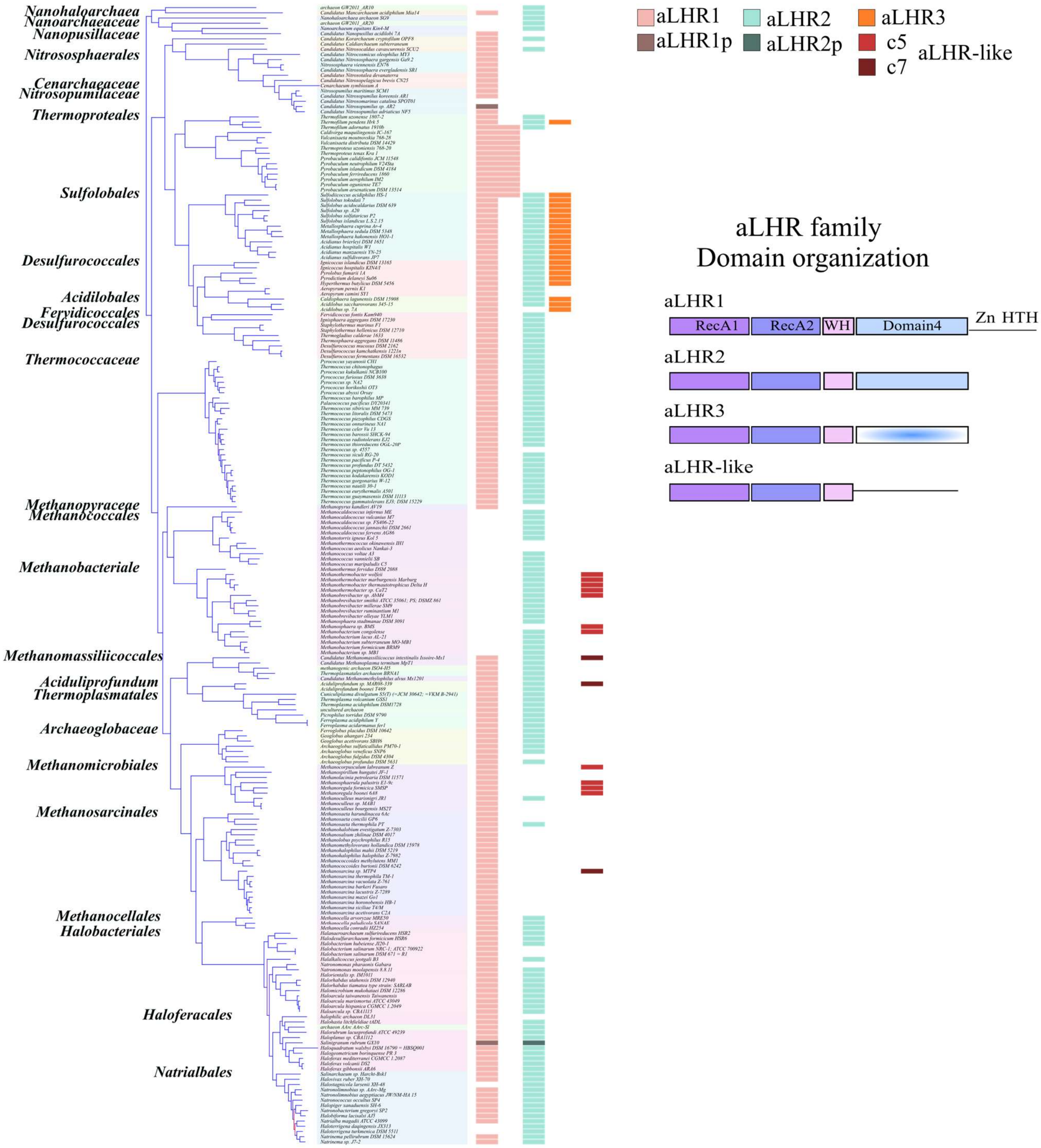

Figure 3. Distribution and domain organization of Lhr families in Archaea. Left panel: species tree of the archaeal genomes was deduced from sets of 120 concatenated sequence alignments with fasttree [30] using LG+GAMMA models and support values determined using 100 non-parametric bootstrap replications (lower bootstrap supports are indicated by red branches). The tree was rooted on DPANN Archaea according to [31]. NCBI taxonomy was reported at the order or family level. The distribution of aLhr families is displayed as bar charts, whose height is proportional to the number of paralogues in each genome (0, 1 or 2). The darker marks in the aLhr1 and aLhr2 column indicate the presence of pseudogenes and are figured as aLhr1p and aLhr2p. The c5 and c7 correspond to archaeal aLhr-like proteins. Right panel: the domain architecture is represented for each archaeal aLhr family. The SF2 helicase core is shown in light and dark violet, WH domain in pink and Domain 4 in light blue. aLhr1 architecture shows an additional C-terminal domain containing a conserved Zn-finger like and HTH motifs; aLhr2 is restricted to the Lhr Core; aLhr3 has a deteriorated Domain 4 (in gradient blue) and aLhr-like misses a Domain 4.

As shown in Figure 2, some aLhr2 proteins are longer than the majority. We therefore analysed more closely their domain organization (Supplementary Figure S2). aLhr2 proteins from Methanomassiliicoccales harbour a longer C-terminal region containing an additional HTH_42 domain (PF06224) also found in bLhr-HTH proteins. An extended C-terminal region is found in other genomes such as the Thermoplasmatales, but with no detected Pfam domain. Interestingly, in few cases, aLhr2 proteins have their RecA2 domain (like in *P. horikoshii* and *Methanocaldococcus sp FS406-22*) or RecA1 domain (like in *Natrialba magadii ATCC 43099*) split by an intein_splicing domain (PF14890) in which a LAGLIDADG domain (PF14528) is inserted.

Concerning the Asgard sequences, we identified 22 aLhr1 proteins and only five aLhr2 proteins distributed as follow (Figure 2). aLhr1 is encoded in 18 Thorarchaeota, in three Heimdallarchaeota and in the only representative of Odinarchaeota. aLhr2 was found in four Heimdallarchaeota and one Thorarchaeota. Since these genomes are incomplete, it is difficult to draw conclusions. However, none of the genomes analysed possesses two Lhr helicases, and in the only complete genome (*Prometheoarchaeum syntrophicum*), we have not identified any Lhr protein. Moreover, we observed that the Asgard aLhr1 proteins form a group that shares a common ancestor node within the aLhr1 subtree (Figure 2). On the other hand, the few Asgard aLhr2 sequences are scattered in the aLhr2 subtree, suggesting acquisition by horizontal transfer. However, more data are needed to confirm these hypotheses.

Finally, we analysed the genomic context of the genes encoding aLhr1 and aLhr2 in order to detect microsynteny since such conservation can give information about function and functional interactions. The products of the genes surrounding *alhr1* and *alhr2* genes were functionally annotated with TIGR HMM profiles. For aLhr1 and aLhr2, no gene conservation has been found across all the archaeal genomes studied (Supplementary Figure S3). Since our experimental work concerns aLhr2 from *T. barophilus*, we looked at the *alhr2* gene neighbourhood more carefully. Conservation of two different neighbour genes has been observed into not closely related genomes. The first one is found either upstream or downstream of the *alhr2* gene in DPANN genomes, in two Thermoproteales genomes and in some Euryarchaeota genomes scattered across the phylogeny (in green for the alhr2 context, Supplementary Figure S3). Its gene product shows similarity with the TIGR0024 profile (putative phosphoesterase), the COG1407 (Predicted ICC-like phosphoesterase) and the cd07391 whose members include archaeal and bacterial proteins homologous to the *Pyrococcus furiosus* PF1019 protein. The domain present in these members belongs to the metallophosphatase (MPP) superfamily. One can notice that such a gene is also found upstream of the *alhr1* gene in six out of nine Methanomicrobiales genomes, in eight out of 20 Methanosarcinales genomes, in most genomes of Halobacteriales and Haloferacales and finally in two genomes of Natrialbales (in yellow on the *alhr1* context, Supplementary Figure S3). However, its presence in the neighbourhood of the *alhr* genes appears to be mutually exclusive, either upstream of alhr1 or in the neighbourhood of alhr2. Interestingly, in Bacteria, a strongly conserved homologous gene called MPE for metallophosphoesterase is also found in the vicinity of *blhr* genes (in 276 genomes out of 319) but not in the neighbourhood of *blhr-HTH* genes that show no apparent conservation. This is in agreement with previous published results [7]. Finally, a second gene is found just upstream of the *alhr2* gene, either in the same or reverse orientation, in most genomes of Sulfolobales, Desulfurococcales and a group of genomes from Thermococcaceae (in yellow for *alhr2* context; Supplementary Figure 3). Its gene product shows a low similarity with the TIGR03937 (poly-beta-1,6 N-acetyl-D-glucosamine synthase) and the COG1215 whose members are described as glycosyltransferases, probably involved in cell wall biogenesis. However, they exhibit very weak similarities that prevent from making any functional prediction.

3.2. The Biochemical properties of aLhr2 of Thermococcus barophilus

In view of previous studies identifying *Paby*-aLhr2 as part of the interaction networks of proteins involved in DNA and RNA transactions [17,18], we focused our attention on characterizing the biochemical properties of Thermococcales aLhr2. We initially chose to study the *Paby* version of aLhr2. However, because the corresponding recombinant protein was toxic when expressed in *E. coli* cells, we decided to perform *in vitro* assays with aLhr2 from *T. barophilus* (*Tbar*-aLhr2). It should be noted that *P. abyssi* and *T. barophilus* are two closely phylogenetically related hyperthermophilic Euryarchaea from the order of Thermococcales that were both identified in the same ecological niche, deep-sea hydrothermal vents [35]. Since *Paby*-aLhr2 and *Tbar*-aLhr2 amino acid sequences share 90% of similarities, we chose to assume that the two orthologous proteins have the same biochemical properties. In addition, we previously showed that antibodies raised against *P. abyssi* proteins recognized their *T. barophilus* counterparts [20]. Interestingly, while the genome of *P. abyssi* cannot be modified with current techniques, *T. barophilus* is now amenable to genetic manipulations, such as gene deletion [18,33].

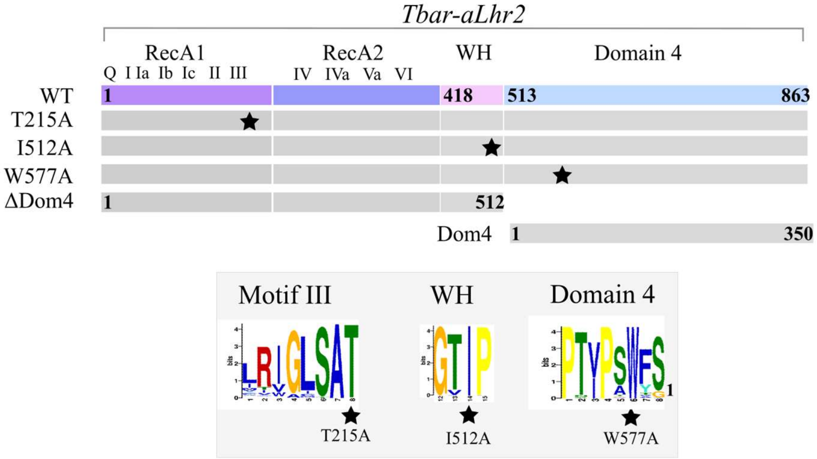

Figure 4. Domain organisation of wild type *Tbar*-aLhr2 and derivatives. The colour code is as in Figure 3. Punctual mutations are represented by a star. The weblogos at the site of mutation are shown below.

Using purified untagged *Tbar*-aLhr2 recombinant protein that we have shown to be monomeric (Supplementary Figure S4), we performed a series of assays to determine properties inherent to helicase enzymes. In the following, we report the capacity of *Tbar*-aLhr2 to hydrolyse ATP, to bind nucleic acids, and to form and unwind duplexes. To do so, we designed a panel of basic RNA and DNA substrates (26 to 50 nucleotides long; Supplementary Table S4) with sequences based on the first study reporting the *in vitro* activity of the archaeal DNA helicase Hel308 of *M. thermautotrophicus* (*Mthe*) [35]. The *in vitro* assays were performed with the wild type protein, with proteins harbouring substitutions in the signature motifs of aLhr2 proteins (T215A, I512A and W577A), and with protein deleted of or restricted to Domain 4 (ΔDom4 and Dom4, respectively) (Figure 4 & Supplementary Figure S4). The residue T215 of Motif III of the SF2 core was predicted to be important for coordination of ATP hydrolysis and nucleic acid binding [36]. The residue W577 of the Domain 4 is highly conserved and was shown to be important in the coupling of ATP hydrolysis and DNA translocation in *M. smegmatis* [7,9].

3.2.1. Tbar-aLhr2 is a nucleic-acid dependent ATPase

To measure the capacity of *Tbar*-aLhr2 to hydrolyse ATP, we performed an ATPase assay with an excess of ATP, and in absence or presence of a single-stranded DNA$_{50}$ molecule or a DNA$_{59}$:DNA$_{31}$ hybrid (Supplementary Table S4). The release of inorganic phosphate was followed over time. Our results show that both DNA$_{50}$ and DNA$_{59}$:DNA$_{31}$ stimulate the *Tbar*-aLhr2 ATPase activity (Supplementary Figure S5A). In absence of nucleic acids, only residual ATP hydrolysis was observed. Altogether, these results show

doi:10.20944/preprints202103.0477.v1

that *Tbar*-aLhr2 is a nucleic acid-dependent ATPase. As observed for other SF2 helicases, ATP hydrolysis can be inactivated by mutating the active site formed by the two RecA1 and RecA2 domains [37,38]. The *Tbar*-aLhr2-T215A protein mutated in the conserved motif III has only a residual ATPase activity in presence of DNA_{50} (Supplementary Figure S5A). Alone, *Tbar*-aLhr2-Dom4 exhibits no ATPase activity. On the other hand, the truncated *Tbar*-aLhr2-ΔDom4, restricted to the RecA1/RecA2 domains and the WH motif, conserved its capacity to hydrolyse ATP but with less efficiency (Supplementary Figure S5A). Altogether, these results suggest that Domain 4 is critical for the optimal ATPase activity of *Tbar*-aLhr2.

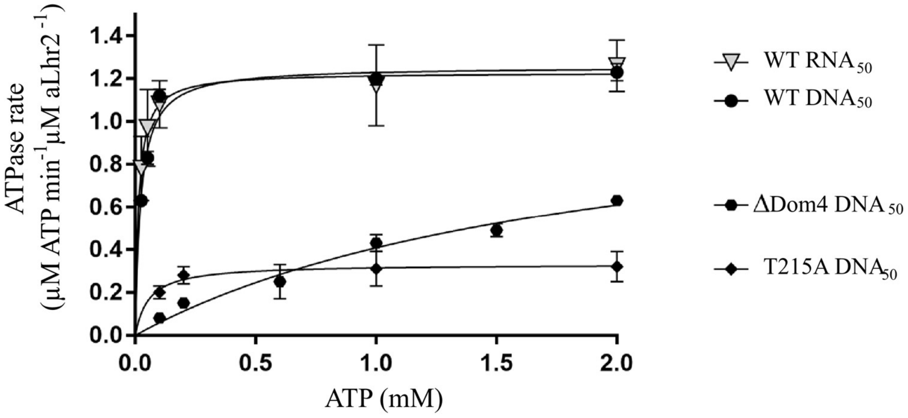

Figure 5. ATPase activity of wild type *Tbar*-aLhr2 and of ΔDom4 and T215A derivatives in presence of nucleic acid molecules. The apparent Michaelis dissociation constant (Km) of WT-*Tbar*-aLhr2 for the DNA_{50} and RNA_{50} molecules are 23±1µM and 13±1µM, respectively. See also Supplementary Figure S5.

Moreover, by assessing ATPase activity rate, we showed that *Tbar*-aLhr2 has no preference for DNA or RNA molecules (Figure 5). To do so, kinetics of ATP hydrolysis were measured at different concentration of ATP in presence of 50-nt long DNA_{50} or RNA_{50} molecules (Supplementary Figure S5B). The calculated ATPase rates of *Tbar*-aLhr2 are identical. We also confirmed that the ATPase rates of the T215A and ΔDom4 mutants are greatly affected (Figure 5; Supplementary Figure S5C).

3.2.2. Tbar-aLhr2 is a nucleic acid binding protein with similar affinities for RNA and DNA molecules

We used a nitrocellulose-filter binding assay to test the capacity of wild type *Tbar*-aLhr2 to bind single-stranded nucleic acid molecules (Figure 6A, left panel) or homoduplex substrates (Supplementary Figure S6). Briefly, an increased concentration of protein was incubated with 5nM of substrates (Supplementary Table S4). The nucleoprotein complexes were separated from free nucleic acids by double filtration on nitrocellulose and nylon membranes. The percentage of bound fraction retained by the nitrocellulose membrane was plotted against the protein concentration (Figure 6A, left panel). The binding curves show a sigmoidal shape with a Hill coefficient superior to 1 (S-shape curves), indicating positive binding cooperativity or multiple binding sites. Most likely, more than one molecule of protein binds to multiple sites on a single molecule of nucleic acids.

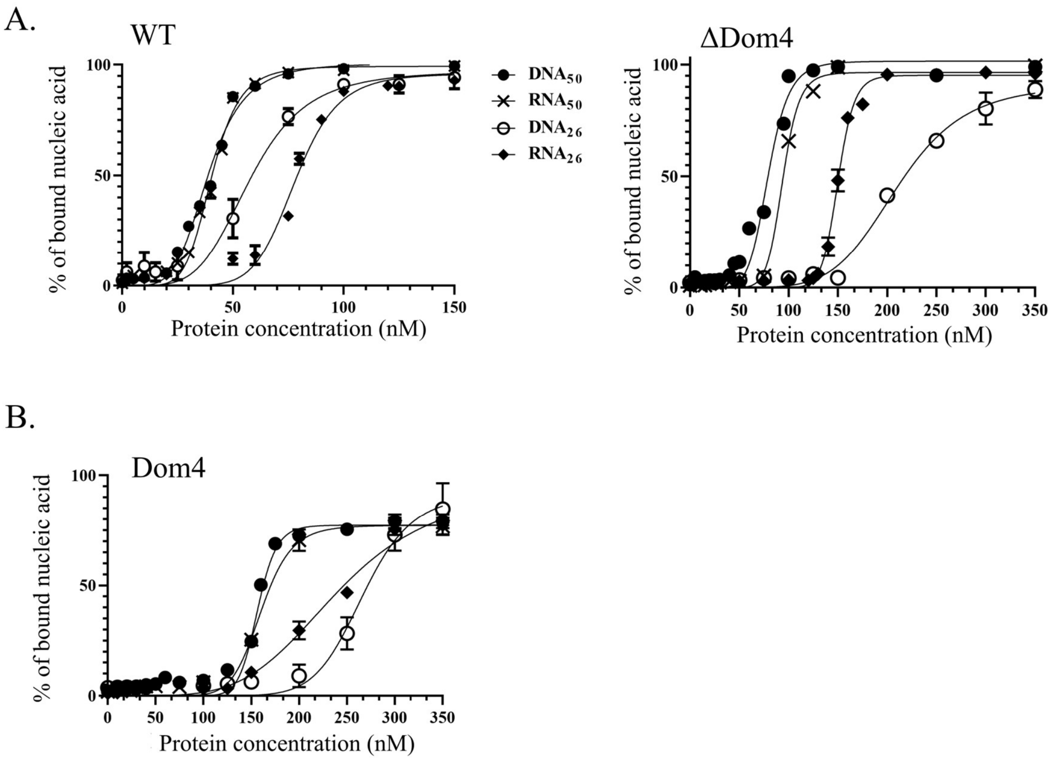

Figure 6. Binding affinities of wild type (WT) *Tbar*-aLhr2 and derivatives for single-stranded nucleic acids. (A) Using WT and ΔDom4 *Tbar*-aLhr2, the percentage of nucleoprotein complex formed after 15 min of incubation was plotted versus the protein concentrations. The experiments were carried with RNA$_{50}$, RNA$_{26}$, DNA$_{50}$ and DNA$_{26}$ substrates (Supplementary Table S4). Three independent experiments were performed in each condition. (B) Identical assays were performed with a recombinant protein corresponding to Domain 4 alone.

The apparent dissociation constant Kd determined for each complex show that *Tbar*-aLhr2-WT binds single-stranded RNA and DNA substrates with the same affinity. A slightly higher affinity was observed for 50nt-long substrates when compared to their 26nt-long counterparts, but with less than 2-fold differences in Kd values it is not significative (Figure 6A, left panel; Table 1). Comparable affinities are also observed for RNA$_{50}$:RNA$_{26}$ and DNA$_{59}$:DNA$_{31}$ homoduplexes (Supplementary Figure S6).

Table 1. Apparent dissociation constant (Kd in nM) derived from data shown in Figure 6 (n.d. means not determined).

Substrates	WT	ΔDom4	Dom4	Hel308
DNA$_{50}$	39±1	54±1	162±5	32±2
DNA$_{26}$	58±2	211±9	277±16	n.d.
RNA$_{50}$	41±1	69±1	161±5	38±1
RNA$_{26}$	80±2	150±1	241±17	n.d.
3’DNA$_{59}$:DNA$_{31}$	41±2	n.d.	n.d.	23±1
3’RNA$_{50}$:RNA$_{26}$	39±2	n.d.	n.d.	n.d.

Tbar-aLhr2-ΔDom4 mutant has similar affinities for 50nt-long DNA and RNA substrates as *Tbar*-aLhr2-WT, but has a lower affinity for the shorter 26nt-long substrates (Figure 6A, right panel; Table 1). While the difference in K$_d$ values is 2.2-fold for the RNA substrates, the difference in Kd values for the DNA substrates is almost 4-fold. Moreover, we observed that Domain 4 of *Tbar*-aLhr2 by itself has the capacity to bind nucleic acids with a lower affinity (Figure 6B; Table 1). Altogether, these results suggest that Domain 4 significantly enhances the capacity of *Tbar*-aLhr2 to bind small nucleic acid substrates. This is consistent with the structure of *Msme*-Lhr (restricted to the first 1-

doi:10.20944/preprints202103.0477.v1

856aa) in complex with AMP-PNP and a single-stranded 16-mer DNA molecule (PDB: 5V9X) showing the RecA2 and Domain 4 forming a clamp around the ssDNA [9].

3.2.3. Tbar-aLhr2 displaces single-stranded RNA from RNA:DNA or RNA:RNA hybrids with a preferred 3′ to 5′ unfolding directionality

To assess the helicase activity of *Tbar*-aLhr2, we tested its capacity to unwind nucleic acid duplexes. The assays were performed with homo- (DNA:DNA or RNA:RNA) or hetero- (DNA:RNA) duplexes with 5′ or 3′ overhangs (Supplementary Table S4). Note that to obtain stable DNA homoduplex in our experimental conditions, we used longer ssDNA molecules (DNA$_{59}$:DNA$_{31}$) than for the heteroduplex (DNA$_{50}$:RNA$_{26}$) or RNA homoduplex (RNA$_{50}$:RNA$_{26}$). The shorter strand was radiolabelled at its 5′ end. Duplex unwinding was followed over time at 65°C in presence of 250nM of *Tbar*-aLhr2. Unlabelled Trap oligo was added in excess to prevent new rounds of duplex formation. A protein-free control experiment was done to assess temperature-dependent unwinding (Supplementary Figure 7A).

The percentage of newly formed single strands was plotted over time (Figure 7A, left panel). It takes 90 min for wild type *Tbar*-aLhr2 to unwound more than 60 % of 3′ overhang 3′RNA$_{50}$:RNA$_{26}$ and 3′DNA$_{50}$:RNA$_{26}$ duplexes. These unwinding activities that are rather slow when compared to other helicases are even more less efficient for a 5′ overhang 5′RNA$_{50}$:RNA$_{26}$ duplex (Figure 7A, left panel). This indicates that *Tbar*-aLhr2 has a slow helicase activity and a preference for 3′ overhang duplexes. In addition, we observed that *Tbar*-aLhr2 is not able to unwind a 3′ overhang 3′DNA$_{59}$:DNA$_{31}$ duplex. Indeed, almost no unwinding is observed as for the control without protein (Figure 7A, left panel). This suggests that *Tbar*-aLhr2 displaces only ssRNA molecules and not ssDNA from homo- or hetero-duplexes.

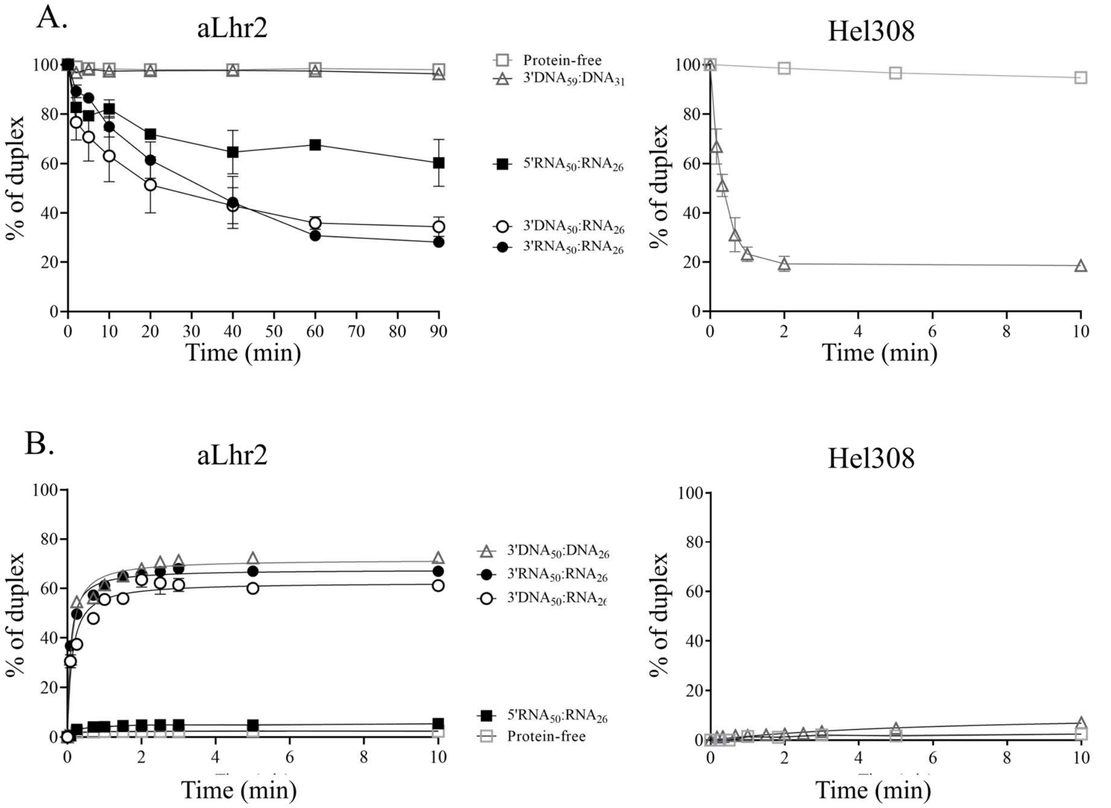

Figure 7. Unwinding and annealing activities of wild type *Tbar*-aLhr2. **(A)** Left panel: Kinetics of strand dissociation by *Tbar*-aLhr2 in presence of ATP are shown for 3′ overhang duplexes

doi:10.20944/preprints202103.0477.v1

(3′DNA$_{59}$:DNA$_{31}$, 3′DNA$_{50}$:RNA$_{26}$ and 3′RNA$_{50}$:RNA$_{26}$) and 5′ overhang duplex (5′RNA$_{50}$:RNA$_{26}$); right panel: Kinetics of strand dissociation by *Paby*-Hel308 are shown for 3′DNA$_{59}$:DNA$_{31}$ duplexes; **(B)** Kinetics of strand association in absence of ATP by *Tbar*-aLhr2 are shown for 3′ overhang duplexes (3′DNA$_{50}$:DNA$_{26}$, 3′DNA$_{50}$:RNA$_{26}$ and 3′RNA$_{50}$:RNA$_{26}$) and 5′ overhang duplex (5′RNA$_{50}$:RNA$_{26}$); Three independent experiments were performed in each condition.

To go further in understanding the unwinding activity of *Tbar*-aLhr2, we compared it to the one of *Paby*-Hel308, which was reported to be a helicase with DNA unwinding activity [39]. Conversely to *Tbar*-aLhr2, *Paby*-Hel308 was able to rapidly unwind 80% of 3′ overhang 3′DNA$_{59}$:DNA$_{31}$ duplexes after only 2 minutes of incubation (Figure 7A, right panel). While, both proteins have similar binding affinities for nucleic acids and do not show any specificity for DNA$_{50}$ substrates (Table 1), their activities differ because of their respective substrates and unwinding velocities. Therefore, Hel308 and aLhr2 that were previously reported to be involved in DNA repair [15,40] most likely operate to perform different tasks in DNA transactions.

All the previous experiments were performed in presence of ATP. We also performed unwinding experiments of 3′ overhang 3′RNA$_{50}$:RNA$_{26}$ duplexes in absence of ATP or in presence of non-hydrolysable ATP analogues. Unexpectedly, we observed that in the absence of ATP, the unwinding activity is comparable to the one obtained in presence of ATP (Supplementary Figure S8A). In our experimental conditions, ATP binding and hydrolysis seem to not be required for slow-rate unwinding activity. It is possible that the energy required for duplex separation is provided by the binding of the protein to the nucleic acid substrate. Moreover, the addition of AMP-PNP or ATPγS abolished completely the observed unwinding activity (Supplementary Figure S8A). The binding of ATP analogues might somehow trap *Tbar*-aLhr2 in an inactive state.

3.2.4. Tbar-aLhr2 forms 3′ overhang duplexes with no preference for RNA or DNA molecules

To test the capacity of wild type *Tbar*-aLhr2 to anneal nucleic acid strands, we did the reverse experiment and followed the formation of homo- and hetero-duplexes from complementary single-stranded DNA or RNA substrates (Supplementary Table S4). Strands annealing was followed overtime at 65°C in presence of 250nM of *Tbar*-aLhr2 (Supplementary Figure S7B).The percentage of newly formed duplexes was plotted overtime (Figure 7B). A protein-free control experiment was done to assess temperature dependent annealing (Supplementary Figure S7B). In absence of ATP, *Tbar*-aLhr2 is able to rapidly anneal nucleic acid single strands to form 3′ overhang 3′DNA$_{50}$:RNA$_{26}$, 3′RNA$_{50}$:RNA$_{26}$ and 3′DNA$_{50}$:DNA$_{26}$ duplexes with no major differences (Figure 7B, left panel). After 10 min of reaction, duplex formation plateaued at 60%, 70% and 75%, respectively. On the other hand, the annealing velocity is drastically reduced when the single strands form 5′ overhang 5′RNA$_{50}$:RNA$_{26}$ duplexes. At 10 min, almost no duplexes are formed as for the control without protein. It seems that in our experimental conditions, *Tbar*-aLhr2 mainly adopts an annealing-competent state.

Again, we compared the activity of *Tbar*-aLhr2 and *Paby*-Hel308. In the same experimental conditions, we showed that *Paby*-Hel308 is unable to rapidly form 3′DNA$_{50}$:DNA$_{26}$ duplexes (Figure 7B, right panel). This result is consistent with Hjm/Hel308 from *Sulfolobus tokodaii* that only exhibits structure specific ssDNA annealing [39]. This also confirms that *in vitro* characteristics of *Tbar*-aLhr2 and *Paby*-Hel308 differ with distinct range of activities.

In presence of ATP, *Tbar*-aLhr2 annealing capacity is reduced with only 20% of 3′RNA$_{50}$:RNA$_{26}$ duplexes formed after 10min (Supplementary Figure S8B). It is possible that the protein conformation switch to an inactive state upon the binding of ATP. This effect is even more drastic in presence of non-hydrolysable ATP analogues (AMP-PNP and ATPγS) (Supplementary Figure S8B). While, it could mean that ATP hydrolysis switches *Tbar*-aLhr2 in an unwinding conformation, it seems unlikely since ATP does not seem to stimulate its unwinding activity that is rather slow (Supplementary Figure

S8A). More likely, the annealing competent state of *Tbar*-aLhr2 is sensitive to the presence of ATP or non-hydrolysable ATP analogues.

3.2.5. Domain 4 is essential for the unwinding and annealing activities of *Tbar*-aLhr2

We revealed that Domain 4 stimulates the ATPase activity of *Tbar*-aLhr2 and by itself has the capacity to bind nucleic acids (Figure 5 & Figure 6). Here, we investigated its role in terms of unwinding and strand-annealing activities (Figure 8A & 8B, respectively). The protein *Tbar*-aLhr2-ΔDom4 which is deprived of Domain 4 has only residual activities. Consistently, *Tbar*-aLhr2-W577A, in which the highly conserved tryptophan at position 577 of Domain 4 is mutated (Figure 4), is also defective for both reactions (Figure 8). We concluded that Domain 4 is essential for the formation of either an active unwinding or annealing-competent state of *Tbar*-aLhr2. The protein variant *Tbar*-aLhr2-I512A with a mutated residue in the WH domain showed similar activities as the wild-type suggesting that this highly conserved residue is not important for the *in vitro* helicase activity in these experimental conditions (Figure 8).

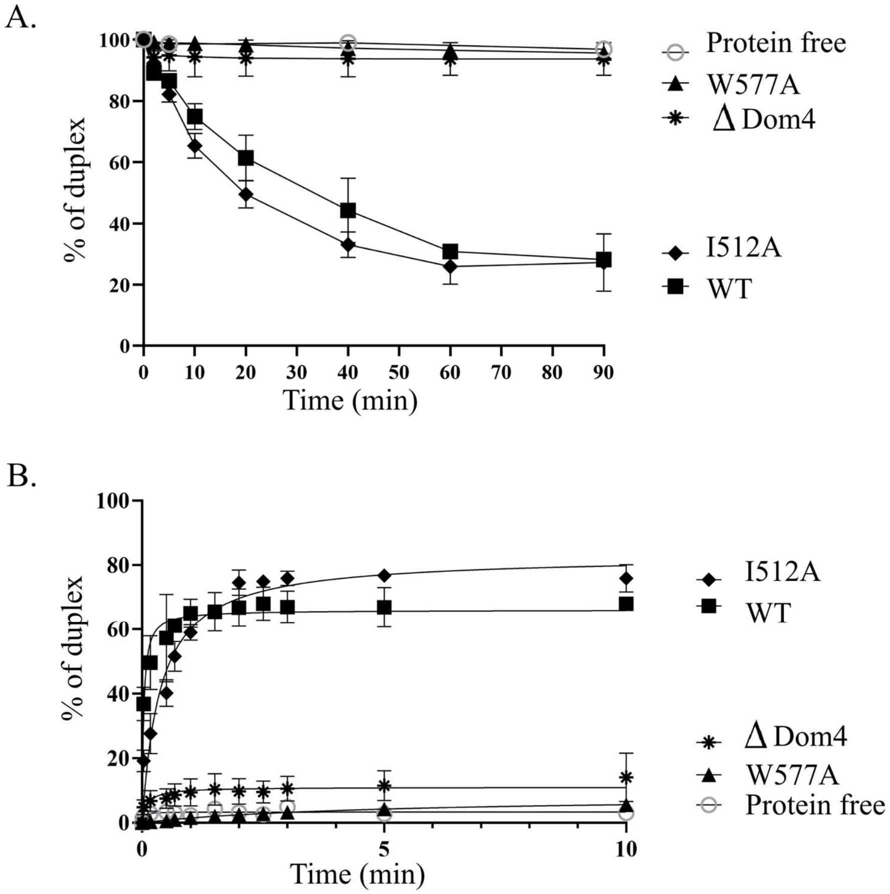

Figure 8. Domain 4 is critical for Tbar-aLhr2 unwinding and strand-annealing activities. **(A)** Kinetics of unwinding reaction (performed as in Figure 7A) of Tbar-aLhr2-ΔDom4, Tbar-aLhr2-W577A and Tbar-aLhr2-I512A using the 3′RNA$_{50}$:RNA$_{26}$ substrate are shown; **(B)** Kinetics of strand-annealing reaction (performed as in Figure 7B) of Tbar-aLhr2-ΔDom4, Tbar-aLhr2-W577A and Tbar-aLhr2-I512A to form 3′RNA$_{50}$:RNA$_{26}$ duplexes are shown.

4. Discussion

Helicases are key enzymes involved in processes that depend on DNA and RNA transactions. They are known to use the energy of ATP to unwind nucleic acids and remodel protein-nucleic acid complexes. Here, we focused on Thermococcales SF2 helicase aLhr2. *Paby*-aLhr2 was found in the network of proteins involved in DNA replication and repair [19]. But, our recent observation also found that *Paby*-aLhr2 was part of the interaction network of proteins involved in RNA processing [20]. This questioned the function(s) of Thermococcales aLhr2. Since *Paby*-aLhr2 is toxic when expressed in *E. coli*, we investigated the *in vitro* activity of its orthologue in *T. barophilus* (*Tbar*-aLhr2). We determined that it is a monomeric DNA/RNA helicase able to process DNA:RNA and RNA:RNA duplexes. Moreover, for other archaeal aLhr helicases that were proposed to be DNA helicases involved in DNA repair and recombination [15,41], it was unclear if they belong to the same aLhr2 group. It also interrogated the relationship that exists between the archaeal and bacterial Lhr proteins. Thus, we performed extensive phylogenomic analyses to elucidate the evolutionary links between Lhr proteins in Archaea and in Bacteria.

The Lhr-type proteins are defined by their unique domain organization [9]. The “Lhr core” is composed of a SF2 helicase core, a winged-helix motif and an Lhr-specific Domain 4 (Figure 1). After recovering and annotating the archaeal and bacterial Lhr-type proteins based on the domain organization, we established their partition in MCL groups and computed a protein family tree with the conserved SF2 core region (Figure 2). We could distinct six groups sharing a common origin: three groups include only Lhr proteins from Archaea: aLhr1, aLhr2 and aLhr3, two groups include only Lhr proteins from Bacteria: bLhr and bLhr-HTH, and the last group while dominated by Bacteria includes few archaeal proteins (Lhr-like). The few archaeal sequences belonging to the Lhr-like group are scattered on the trees and should have been acquired from Bacteria through horizontal gene transfers. The Lhr groups were named based on previous studies [14] and domain organization. According to the tree topology (Figure 2), the sequences of the bLhr-HTH and aLhr2 groups share a common ancestor that predates the divergence of Bacteria and Archaea, suggesting that these sequences are orthologous and may have conserved similar roles in these genomes. The sequences of the bLhr group are characterized by significantly longer branches than those of the other groups. This reflects an acceleration of their rate of evolution and could be responsible for the instability of the anchoring of this group between the Sfth rooted (Supplementary Figure S1) and Lhr-like rooted (Figure 2) trees. Despite this difference in localization, the bLhr sequences share, on both trees, a common hypothetical ancestor with the bLhr-HTH and aLhr2 helicases. On the other hand, the aLhr1 group does not appear directly related to bacterial sequences.

While initial work identified two groups of archaeal Lhr proteins, aLhr1 and aLhr2 [14], we identified a third group aLhr3 that has a highly deteriorated Domain 4 and that seems to be limited to the Sulfolobales, Desulfurococcales and Acidobales. On the other hand, aLhr1 and aLhr2 are widespread in archaeal genomes and often present together. Their absence in some genomes would most likely result from different independent events of gene loss. Interestingly, only four out of 219 have lost both genes. We defined that the characterized Lhr proteins of *S. solfataricus*, *S. acidocaldarius*, and *M. thermautotrophicus* [15,16] belong to the aLhr2 group. To date, no aLhr1 were characterized. Genetic studies on a strain of *S. islandicus* deleted for the gene encoding aLhr1 [17] suggest that *Sisl*-aLhr1 has a role in DNA repair, as shown for *Ssol*-, *Saci*- and *Mthe*-aLhr2, but its biochemical properties are unknown. More work is needed to identify the common and/or specific functions of aLhr1 and aLhr2 in archaeal cells. In particular, since aLhr1 differs from aLhr2 by an additional cysteine-rich motif at its C-terminal end, the presence or not of an additional putative Zinc-finger might differentiate the proteins activities and protein partners.

Interestingly, in eleven genomes the RecA2 (10/11) or RecA1 (1/11) domain of aLhr2 is spliced by an intein_splicing domain. Proteins containing this integration are dispersed on the tree suggesting independent acquisitions. Interestingly, ATPase domains were shown to be hot-spots for inteins integration, with 70% of all inteins residing in ATPase-containing proteins, and at many different integration sites [42]. While they are

doi:10.20944/preprints202103.0477.v1

generally considered as being selfish parasites, intein splicing has recently been shown to be regulated by external stimuli such as temperature, pH, salt and DNA damage [42]. Thus, some aLhr2 proteins might be regulated at the post-translational level and activated upon stress.

In Bacteria, for the first time, we identified three groups of Lhr proteins: bLhr, bLhr-HTH and Lhr-like. While bLhr proteins are restricted to the "Lhr core", the group of bLhr-HTH has an additional HTH_42 domain at its C-terminal end. As before, the presence of an extra domain interrogates its role in the protein activities and interaction with partners. To this day, one bLhr from *P. putida* and two bLhr-HTH from *M. smegmatis* and *E. coli* were studied, but the role of the HTH_42 domain was not investigated. Interestingly, some aLhr2 from Methanomassilicoccales also possess an additional HTH_42 extension. Finally, the Lhr-like were defined as Lhr-type proteins but while they possess a C-terminal region no Domain 4 can be defined. Nonetheless, structure prediction of Lhr-like from *Streptomyces coelicolor* showed that its C-terminal domain adopts a structure that is similar to the predicted structure of Domain 4 (Supplementary Figure S9).

In this study, we also report the *in vitro* activity of aLhr2 from *T. barophilus*. First, we showed that the ATPase activity of *Tbar*-aLhr2 is consistent with the one measured for the bacterial *Msme*-bLhr-HTH and *Pput*-bLhr proteins [6,43]. We showed that *Tbar*-aLhr2 is a nucleic acid-dependent ATPase with no preference for DNA or RNA molecules (Figure 5). Only, the archaeal *Ssol*-aLhr2 was shown to be able to hydrolyse ATP in absence of nucleic acids [41]. *Ssol*-aLhr2 also differs from the monomeric *Tbar*-aLhr2 and bacterial Lhr proteins by its low affinity for single-stranded DNA and by its oligomeric state that can be both monomeric and dimeric; monomers and dimers having specific biochemical activities.

We also characterized *Tbar*-aLhr2 as a monomeric DNA/RNA helicase able to process DNA:RNA and RNA:RNA duplexes (Figure S4 & Figure 7, left panels). Interestingly, we highlighted that the *in vitro* unwinding and annealing activities of *Tbar*-aLhr2 differs drastically from the ones of Hel308, described as a DNA helicase involved in DNA repair [35,44]. Indeed, we showed that while *Tbar*-aLhr2 is more prone to anneal nucleic strands than to unwind them, *Paby*-Hel308 unwinds 3'DNA:DNA homoduplexes but do not form them from ssDNA molecules (Figure 7). Altogether, these results clearly indicate that while both proteins are proposed to be involved in DNA repairs [15,40], they most likely operate to perform different tasks in DNA transactions.

Among Lhr proteins, the capacity to process DNA:RNA hybrids is not specific of *Tbar*-aLhr2. Indeed, while the bacterial *Msme*-bLhr-HTH and *Pput*-bLhr, and the archaeal *Mthe*-aLhr2 were shown to process other substrates (forked DNA or Holliday junctions), they are also able to unwind DNA:RNA hybrids [6,7,9,15]. *Msme*-bLhr-HTH even prefers 3'-tailed RNA:DNA hybrids over DNA:DNA duplex and was also described as an RNA/DNA helicase [9]. The capacity of *Tbar*-aLhr2 to unwind more efficiently *in vitro* hybrids with a 3' overhang strand that is indicative of a 3' to 5' polarity is also consistent with the polarity previously observed for its archaeal and bacterial counterparts [6,7,9,15,16]

Tbar-aLhr2 has a significant ability to anneal single-stranded nucleic acid substrates to form DNA:DNA, RNA:RNA or RNA:DNA duplexes with no apparent preferences (Figure 7). Both the monomeric and dimeric *Ssol*-aLhr2 were also shown to have DNA strand annealing activities that are comparable to the one of *Tbar*-aLhr2 [16]. Intriguingly, we found that *Tbar*-aLhr2 is less prone to unwind duplexes (Figure 7A) than to anneal nucleic strands (Figure 7B). Indeed, we noted that the *in vitro* unwinding activity of *Tbar*-aLhr2 is slow. This is also the case of *Msme*-bLhr-HTH which was shown to have a slow capacity to dissociate nucleic acid strands (observed after 30min of incubation) [6]. This can be relevant for the cellular functions of *Tbar*-aLhr2 but it can also mean that our experimental conditions are not optimal.

ATP binding is proposed to act as a molecular switch from a strand-annealing to an unwinding mode by changing the protein conformation. Here, we showed that *Tbar*-aLhr2 unwinding activity is independent of the presence of ATP. While it might seem surprising, ATP-independent unwinding activities were previously reported for the human SF2 NS3h helicase and the bacterial SF1 RecBCD helicase [45,46]. It was proposed that the energy required for duplex separation is provided by nucleic acid binding and not by ATP binding and hydrolysis. *In vivo*, it is possible that the balance between unwinding and annealing states is displaced upon interaction with specific protein partners.

We also investigated the role of Domain 4, a novel structural domain specific of Lhr proteins, and demonstrated that it is essential for *Tbar*-aLhr2 to adopt active conformations. First, we showed that while the ATPase activity is carried by the SF2 catalytic core composed of the RecA1 and RecA2 domains, Domain 4 stimulates ATP hydrolysis (Supplementary Figure S4). Finally, we found that Domain 4 is essential for *Tbar*-aLhr2 annealing and the unwinding activities. The substitution of the highly conserved tryptophan at position 577 in *Tbar*-aLhr2 Domain 4 is sufficient to abolish these reactions (Figure 8). These results are in agreement with those obtained for bacterial *Msme*-bLhr-HTH restricted to the Lhr core [9].

Our results underline the high capacity of *Tbar*-aLhr2 to perform nucleic acid strand annealing. This property could be of high importance in hyperthermophile organisms, such as *T. barophilus*, to maintain nucleic acid duplexes at high temperature. While there is much that remains to be uncovered with respect to the cellular functions of aLhr2 *in vivo*, we can propose the following roles for Thermococcales aLhr2. First, the detection of *Paby*-aLhr2 in the interaction network of the replication protein A complex (RPA) [19] is consistent with studies proposing that aLhr2 helicases are involved in DNA recombination and repair in *S. solfataricus* and *M. thermautotrophicus* [15,41]. Indeed, RPA that binds ssDNA is crucial for both DNA replication and DNA damage response [47]. This is also coherent with an involvement of Thermococcales aLhr2 in RNA transactions. In *P. abyssi*, RPA was also shown to enhance transcription [19] and to be part of the interaction network of 5′-3′ exoribonuclease aRNase J [20], questioning its involvement in RNA metabolism. The involvement of Thermococcales aLhr2 in RNA metabolism is also supported by our initial observation that *Paby*-aLhr2 was found to be in the protein network of ASH-Ski2 with a high specificity index (Supplementary Table S1); ASH-Ski2 is an archaeal specific Ski2-like helicase that forms a complex with aRNase J [20]. Interestingly, RPA is also found in the interaction network of ASH-Ski2 (Supplementary Table S1).

Furthermore, we showed that *Tbar*-aLhr2 is a DNA/RNA helicase able to process DNA:RNA duplexes. The ability of Lhr proteins to process such hybrids was also identified for bacterial *Msme*-bLhr-HTH and *Pput*-bLhr, and for archaeal *Mthe*-aLhr2 helicases [6,7,9,15]. In the cells, DNA:RNA hybrids are often found in R-loop, a three-stranded structure that harbours a DNA:RNA hybrid and a displaced single-stranded DNA. Controlling R-loop formation and suppression is critical for many cellular processes. While R-loops are often associated with genome instability, DNA damage and transcription elongation defect, mounting evidence suggest that R-loops promote DNA transactions including DNA recombination and repair [48]. Interestingly, RPA was also recently revealed to act as a sensor of R-loops and to regulate RNase H1 in human cells [49]. Moreover, defect in mRNA processing was recently associated with R-loop-dependent genome instability in Eukaryotes [48]. Further physiological and mechanical studies are necessary to determine the function(s) of aLhr2 in Thermococcales cells.

Supplementary Materials: Figure S1: Phylogenetic tree of Lhr sequences rooted with Sfth representative protein, Figure S2 : Domain architecture of aLhr2 proteins ,Figure S3: The neighbourhood of aLhr1 and aLhr2 encoding genes, Figure S4: *Tbar*-aLhr2 recombinant proteins used in this study, Figure S5: ATPase activity of WT and derivative*Tbar*-aLhr2, Figure S6 Binding affinity of *Tbar*-aLhr2 and P*aby*-Hel308 for 3′overhang DNA and RNA duplexes, Figure S7: Wild type *Tbar*-aLhr2 and variants unwinding and annealing activities Figure S8: Unwinding and strand-annealing activities of *Tbar*-aLhr2WT in presence of ATP, AMP.PNP and ATPγS (ATP analogues), Figure S9: Structure model of *M. smegmatis* bLhr-HTH Lhr-core and *S. coelicolor* Lhr-like, Table S1: (His$_6$)-*Paby*-ASH-Ski2 list of protein partners extracted from [20], Table S2: The Lhr and Sfth protein sequence identifiers used in Figures 2 and Supplementary Figure S1, with the related organisms, Uniprot accession numbers and locus-tags of the achaeal Lhr (Excel file Table S2A), bacterial

Lhr (Excel file Table S2B) and Sfth (Excel file Table S2C), Table S3: Sequences of synthetic oligonucleotides used in this study. Table S4: Nucleotide sequences of the single-stranded and duplex DNA and RNA substrates used in this study.

Author Contributions: "Conceptualization, H.C., G.F., M.Bo. and B.C.O.; Funding acquisition, H.C., Z.A.R, G.F., B.C.O.; Methodology, M.H., P.L.G., M.Ba., Y.Q., S.L., G.F., M.Bo., and B.C.O.; Validation, M.H., M.Ba., P.L.G., Y.Q., G.F. M.Bo., and B.C.O.; Formal analysis, M.H., M.Ba., P.L.G, and Y.Q. ; Investigation, M.H., P.L.G., M.Ba., Y.Q., S.L.; Writing - original draft preparation, G.F., M.Bo. and B.C.O.; Writing - review and editing, M.H., H.C., Z.A.R., D.F., S.L., Y.Q., G.F., M.Bo. and B.C.O.; Supervision, G.F., M.Bo and B.C.O.; All authors have read and agreed to the published version of the manuscript."

Funding: This work was supported by CNRS, University Paul Sabatier, French "Agence Nationale pour la Recherche" [ANR-16-CE12-0016-01 to B.C.O], Lebanese University and AZM & SAADÉ association (PhD fellowship to M.H)

Acknowledgments: We thank D.K. Phung for initial work on *P.abyssi* aLhr2, M. Kwapisz for helpful discussion and Julien Anglade for technical help.

Conflicts of Interest the authors declare no conflict of interest

References

1. Singleton, M. R.; Dillingham, M. S.; Wigley, D. B. Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.* **2007**, *76*, 23–50.
2. Fairman-Williams, M. E.; Guenther, U.-P.; Jankowsky, E. SF1 and SF2 helicases: family matters. *Curr. Opin. Struct. Biol.* **2010**, *20*, 313–324.
3. Jankowsky, E.; Fairman, M. E. RNA helicases--one fold for many functions. *Curr. Opin. Struct. Biol.* **2007**, *17*, 316–324.
4. Fairman, M. E.; Maroney, P. A.; Wang, W.; Bowers, H. A.; Gollnick, P.; Nilsen, T. W.; Jankowsky, E. Protein displacement by DExH/D "RNA helicases" without duplex unwinding. *Science* **2004**, *304*, 730–734.
5. Jankowsky, E.; Bowers, H. Remodeling of ribonucleoprotein complexes with DExH/D RNA helicases. *Nucleic Acids Res.* **2006**, *34*, 4181–4188.
6. Ordonez, H.; Shuman, S. Mycobacterium smegmatis Lhr Is a DNA-dependent ATPase and a 3'-to-5' DNA translocase and helicase that prefers to unwind 3'-tailed RNA:DNA hybrids. *J. Biol. Chem.* **2013**, *288*, 14125–14134.
7. Ejaz, A.; Shuman, S. Characterization of Lhr-Core DNA helicase and manganese- dependent DNA nuclease components of a bacterial gene cluster encoding nucleic acid repair enzymes. *J. Biol. Chem.* **2018**, *293*, 17491–17504.
8. Reuven, N. B.; Koonin, E. V.; Rudd, K. E.; Deutscher, M. P. The gene for the longest known Escherichia coli protein is a member of helicase superfamily II. *J. Bacteriol.* **1995**, *177*, 5393–5400.
9. Ejaz, A.; Ordonez, H.; Jacewicz, A.; Ferrao, R.; Shuman, S. Structure of mycobacterial 3'-to-5' RNA:DNA helicase Lhr bound to a ssDNA tracking strand highlights distinctive features of a novel family of bacterial helicases. *Nucleic Acids Res.* **2018**, *46*, 442–455.
10. Büttner, K.; Nehring, S.; Hopfner, K.-P. Structural basis for DNA duplex separation by a superfamily-2 helicase. *Nat. Struct. Mol. Biol.* **2007**, *14*, 647–652.
11. Rand, L.; Hinds, J.; Springer, B.; Sander, P.; Buxton, R. S.; Davis, E. O. The majority of inducible DNA repair genes in Mycobacterium tuberculosis are induced independently of RecA. *Mol. Microbiol.* **2003**, *50*, 1031–1042.
12. Boshoff, H. I. M.; Reed, M. B.; Barry, C. E.; Mizrahi, V. DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in Mycobacterium tuberculosis. *Cell* **2003**, *113*, 183–193.
13. Cooper, D. L.; Boyle, D. C.; Lovett, S. T. Genetic analysis of Escherichia coli RadA: functional motifs and genetic interactions. *Mol. Microbiol.* **2015**, *95*, 769–779.
14. Chamieh, H.; Ibrahim, H.; Kozah, J. Genome-wide identification of SF1 and SF2 helicases from archaea. *Gene* **2016**, *576*, 214–228.
15. Buckley, R. J.; Kramm, K.; Cooper, C. D. O.; Grohmann, D.; Bolt, E. L. Mechanistic insights into Lhr helicase function in DNA repair. *Biochem. J.* **2020**, *477*, 2935–2947.
16. De Felice, M.; Aria, V.; Esposito, L.; De Falco, M.; Pucci, B.; Rossi, M.; Pisani, F. M. A novel DNA helicase with strand-annealing activity from the crenarchaeon Sulfolobus solfataricus. *Biochem. J.* **2007**, *408*, 87–95.
17. Song, X.; Huang, Q.; Ni, J.; Yu, Y.; Shen, Y. Knockout and functional analysis of two DExD/H-box family helicase genes in Sulfolobus islandicus REY15A. *Extremophiles* **2016**, *20*, 537–546.

doi:10.20944/preprints202103.0477.v1

18. van Wolferen, M.; Ma, X.; Albers, S.-V. DNA Processing Proteins Involved in the UV-Induced Stress Response of Sulfolobales. *J. Bacteriol.* **2015**, *197*, 2941–2951.

19. Pluchon, P.-F.; Fouqueau, T.; Crezé, C.; Laurent, S.; Briffotaux, J.; Hogrel, G.; Palud, A.; Henneke, G.; Godfroy, A.; Hausner, W.; Thomm, M.; Nicolas, J.; Flament, D. An extended network of genomic maintenance in the archaeon Pyrococcus abyssi highlights unexpected associations between eucaryotic homologs. *PLoS One* **2013**, *8*, e79707.

20. Phung, D. K.; Etienne, C.; Batista, M.; Langendijk-Genevaux, P.; Moalic, Y.; Laurent, S.; Liuu, S.; Morales, V.; Jebbar, M.; Fichant, G.; Bouvier, M.; Flament, D.; Clouet-d'Orval, B. RNA processing machineries in Archaea: the 5'-3' exoribonuclease aRNase J of the β-CASP family is engaged specifically with the helicase ASH-Ski2 and the 3'-5' exoribonucleolytic RNA exosome machinery. *Nucleic Acids Res.* **2020**, *48*, 3832–3847.

21. Yang, M.; Derbyshire, M. K.; Yamashita, R. A.; Marchler-Bauer, A. Ncbi's conserved domain database and tools for protein domain analysis. *Curr Protoc Bioinformatics* **2020**, *69*, e90.

22. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195.

23. Enright, A. J.; Van Dongen, S.; Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30*, 1575–1584.

24. van Dongen, S.; Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* **2012**, *804*, 281–295.

25. Katoh, K.; Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780.

26. Ali, R. H.; Bogusz, M.; Whelan, S. Identifying clusters of high confidence homologies in multiple sequence alignments. *Mol. Biol. Evol.* **2019**, *36*, 2340–2351.

27. Darriba, D.; Posada, D.; Kozlov, A. M.; Stamatakis, A.; Morel, B.; Flouri, T. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.* **2020**, *37*, 291–294.

28. Minh, B. Q.; Schmidt, H. A.; Chernomor, O.; Schrempf, D.; Woodhams, M. D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534.

29. Letunic, I.; Bork, P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **2019**, *47*, W256–W259.

30. Parks, D. H.; Rinke, C.; Chuvochina, M.; Chaumeil, P.-A.; Woodcroft, B. J.; Evans, P. N.; Hugenholtz, P.; Tyson, G. W. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2017**, *2*, 1533–1542.

31. Capella-Gutiérrez, S.; Silla-Martínez, J. M.; Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009**, *25*, 1972–1973.

32. Chang, J.-M.; Di Tommaso, P.; Notredame, C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.* **2014**, *31*, 1625–1637.

33. Price, M. N.; Dehal, P. S.; Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **2009**, *26*, 1641–1650.

34. Williams, T. A.; Szöllősi, G. J.; Spang, A.; Foster, P. G.; Heaps, S. E.; Boussau, B.; Ettema, T. J. G.; Embley, T. M. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E4602–E4611.

35. Guy, C. P.; Bolt, E. L. Archaeal Hel308 helicase targets replication forks in vivo and in vitro and unwinds lagging strands. *Nucleic Acids Res.* **2005**, *33*, 3678–3690.

36. Zhang, X.-P.; Janke, R.; Kingsley, J.; Luo, J.; Fasching, C.; Ehmsen, K. T.; Heyer, W.-D. A conserved sequence extending motif III of the motor domain in the Snf2-family DNA translocase Rad54 is critical for ATPase activity. *PLoS One* **2013**, *8*, e82184.

37. Banroques, J.; Doère, M.; Dreyfus, M.; Linder, P.; Tanner, N. K. Motif III in superfamily 2 "helicases" helps convert the binding energy of ATP into a high-affinity RNA binding site in the yeast DEAD-box protein Ded1. *J. Mol. Biol.* **2010**, *396*, 949–966.

38. Mackeldanz, P.; Alves, J.; Möncke-Buchner, E.; Wyszomirski, K. H.; Krüger, D. H.; Reuter, M. Functional consequences of mutating conserved SF2 helicase motifs in the Type III restriction endonuclease EcoP15I translocase domain. *Biochimie* **2013**, *95*, 817–823.

39. Li, Z.; Lu, S.; Hou, G.; Ma, X.; Sheng, D.; Ni, J.; Shen, Y. Hjm/Hel308A DNA helicase from Sulfolobus tokodaii promotes replication fork regression and interacts with Hjc endonuclease in vitro. *J. Bacteriol.* **2008**, *190*, 3006–3017.

doi:10.20944/preprints202103.0477.v1

40. Johnson, S. J.; Jackson, R. N. Ski2-like RNA helicase structures: common themes and complex assemblies. *RNA Biol.* **2013**, *10*, 33–43.

41. De Falco, M.; Massa, F.; Rossi, M.; De Felice, M. The Sulfolobus solfataricus RecQ-like DNA helicase Hel112 inhibits the NurA/HerA complex exonuclease activity. *Extremophiles* **2018**, *22*, 581–589.

42. Belfort, M. Mobile self-splicing introns and inteins as environmental sensors. *Curr. Opin. Microbiol.* **2017**, *38*, 51–58.

43. Ejaz, A.; Goldgur, Y.; Shuman, S. Activity and structure of Pseudomonas putida MPE, a manganese-dependent single-strand DNA endonuclease encoded in a nucleic acid repair gene cluster. *J. Biol. Chem.* **2019**, *294*, 7931–7941.

44. Zhai, B.; DuPrez, K.; Han, X.; Yuan, Z.; Ahmad, S.; Xu, C.; Gu, L.; Ni, J.; Fan, L.; Shen, Y. The archaeal ATPase PINA interacts with the helicase Hjm via its carboxyl terminal KH domain remodeling and processing replication fork and Holliday junction. *Nucleic Acids Res.* **2018**, *46*, 6627–6641.

45. Reynolds, K. A.; Cameron, C. E.; Raney, K. D. Melting of Duplex DNA in the Absence of ATP by the NS3 Helicase Domain through Specific Interaction with a Single-Strand/Double-Strand Junction. *Biochemistry* **2015**, *54*, 4248–4258.

46. Lohman, T. M.; Fazio, N. T. How Does a Helicase Unwind DNA? Insights from RecBCD Helicase. *Bioessays* **2018**, *40*, e1800009.

47. Maréchal, A.; Zou, L. RPA-coated single-stranded DNA as a platform for post-translational modifications in the DNA damage response. *Cell Res.* **2015**, *25*, 9–23.

48. Crossley, M. P.; Bocek, M.; Cimprich, K. A. R-Loops as Cellular Regulators and Genomic Threats. *Mol. Cell* **2019**, *73*, 398–411.

49. Nguyen, H. D.; Yadav, T.; Giri, S.; Saez, B.; Graubert, T. A.; Zou, L. Functions of replication protein A as a sensor of R loops and a regulator of rnaseh1. *Mol. Cell* **2017**, *65*, 832–847.e4.