Published in final edited form as:

Methods Enzymol. 2017 ; 592: 103–121. doi:10.1016/bs.mie.2017.03.026.

Expression and Structural Analyses of Human DNA polymerase θ (POLQ)

Andrew W. Malaby¹, Sara K. Martin², Richard D. Wood², and Sylvie Doublié^{1,*}

¹Department of Microbiology and Molecular Genetics, University of Vermont, 89 Beaumont Ave, Burlington, VT 05405, USA

²Department of Epigenetics & Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center P.O. Box 389, Smithville, TX, 78957, USA and The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

Abstract

DNA polymerase theta (pol θ) is an evolutionarily conserved protein encoded by the *POLQ* gene in mammalian genomes. Pol θ is the defining enzyme for a pathway of DSB repair termed “alternative end-joining” (altEJ) or “theta-mediated end-joining”. This pathway contributes significantly to the radiation resistance of mammalian cells. It also modulates accuracy in repair of breaks that occur at stalled DNA replication forks, during diversification steps of the mammalian immune system, during repair of CRISPR-Cas9, and in many DNA integration events. Pol θ is a potentially important clinical target, particularly for cancers deficient in other break repair strategies. The enzyme is uniquely able to mediate joining of single-stranded 3' ends. Because of these unusual biochemical properties and its therapeutic importance, it is essential to study structures of pol θ bound to DNA. However, challenges for expression and purification are presented by the large size of pol θ (2,590 residues in humans) and unusual juxtaposition of domains (a helicase-like domain and distinct DNA polymerase, separated by a region predicted to be largely disordered). Here we summarize work on the expression and purification of the full-length protein, and then focus on the design, expression and purification of an active C-terminal polymerase fragment. The generation of this active construct was non-trivial and time-consuming. Almost all published biochemical work to date has been performed with this domain fragment. Strategies to obtain and improve crystals of a ternary pol θ complex (enzyme:DNA:nucleotide) are also presented, along with key elements of the structure.

Keywords

DNA polymerase; alternative end-joining; MMEJ; DNA double strand breaks; DNA synthesis; synthetic lethality

*To whom correspondence should be addressed: sdoublié@uvm.edu, tel (802) 656-9531, fax (802) 656-8749.

The human pol θ construct (aa 1792-2590) and the three deletion constructs (Ins1, Ins2, and Ins3) were deposited in the Addgene plasmid repository (Addgene.org) and are available to investigators from academia under a Materials Transfer Agreement.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no financial, personal or professional competing interests that could be construed to have influenced this paper.

1. Introduction

DNA polymerase theta (pol θ) is the defining enzyme for a pathway of DNA end-joining that is important for double-strand break repair during diverse genomic transactions. Pol θ -dependent end joining is distinct from the more widely studied “classical” nonhomologous end-joining (cNHEJ) repair initiated by the Ku proteins. In mammalian cells, pol θ is encoded by the *POLQ* gene. Pol θ has a characteristic C-terminal DNA polymerase domain, linked via a central region to an N-terminal DNA helicase-like domain (Seki, Marini, & Wood, 2003; Seki, Masutani, Yang, Schuffert, Iwai, Bahar et al., 2004; Yousefzadeh & Wood, 2013) (Fig. 1). Genes encoding proteins with this arrangement are present throughout nature in many eukaryotes, though not in fungi (Beagan & McVey, 2016; Wood & Doublié, 2016; Yousefzadeh & Wood, 2013).

Pol θ -dependent end-joining is currently referred to by several names including “alternative end-joining” (altEJ) and “theta-mediated end-joining” (TMEJ). It is a pathway that can join substrates with 3'-single-stranded DNA tails generated by enzymatic resection (Fig. 1) (Yousefzadeh, Wyatt, Takata, Mu, Hensley, Tomida et al., 2014; Wyatt, Feng, Conlin, Yousefzadeh, Roberts, Mieczkowski et al., 2016). This can occur, for example, when a DNA break cannot be efficiently repaired by Ku-dependent nonhomologous end-joining. Inactivation of *POLQ* increases the sensitivity of cells to ionizing radiation and to some double strand break-inducing drugs, and so suppression of *POLQ* might be a useful adjuvant to some DNA damaging therapies (Goff et al. 2009; Higgins, Prevo, Lee, Helleday, Muschel, Taylor et al., 2010; Lemée, Bergoglio, Fernandez-Vidal, Machado-Silva, Pillaire, Bieth et al., 2010; Yousefzadeh et al., 2014).

Pol θ -dependent DNA repair is critical when other pathways of double-strand break repair are compromised. Pol θ -dependent repair has been shown to be particularly important in the absence of ATM, FANCD2, BRCA1, BRCA2, and Ku70 (Ceccaldi, Liu, Amunugama, Hajdu, Primack, Petalcorin et al., 2015; Mateos-Gomez, Gong, Nair, Miller, Lazzerini-Denchi, & Sfeir, 2015; Shima, Munroe, & Schimenti, 2004; Wyatt et al., 2016). As a result, homologous-recombination defective cells are more sensitive to *POLQ* ablation. This raises the prospect of improving treatment of homologous recombination defective tumors (such as BRCA1- and BRCA2 defective breast and ovarian cancers) by inhibition of *POLQ*. Pol θ has at least two enzymatic activities (polymerase and ATPase) that may be targeted by small molecule inhibitors. Cells of at least some cancers may be particularly susceptible because they appear to depend on higher levels of pol θ expression.

The action of altEJ/TMEJ leaves characteristic DNA sequence “signatures” at the site of break repair. One of these signatures is the presence of “microhomologies” (1–2 bp) that are identical at the ends of the joined breaks (Fig. 1) (Chan, Yu, & McVey, 2010; Wyatt et al., 2016). A second molecular trace is the insertion of additional DNA bases at some sites of DSB joining. These insertions arise by templating from the adjoining single-stranded DNA tails, or by template switching from other DNA ends in the cell (Yousefzadeh et al., 2014), even from other chromosomes (Wyatt et al., 2016). Often several cycles of template switching are apparent. Such events have been observed for example during *POLQ*-dependent repair of directed double-strand breaks in *Drosophila* (Chan et al., 2010), *C.*

elegans (Koole, van Schendel, Karambelas, van Heteren, Okihara, & Tijsterman, 2014), and the mouse (Wyatt et al., 2016; Yousefzadeh et al., 2014).

The ability of pol θ to switch templates and prime from microhomologies or minimally-primed templates is also manifested by the observation that it can extend some single-stranded DNA substrates (Hogg, Sauer-Eriksson, & Johansson, 2012; Kent, Mateos-Gomez, Sfeir, & Pomerantz, 2016; Yousefzadeh et al., 2014). The fact that pol θ can extend minimally primed DNA is consistent with its unusual efficiency in extending from mismatched DNA termini (Seki et al., 2004; Seki & Wood, 2008), and its tendency towards primer-template slippage (Arana, Seki, Wood, Rogozin, & Kunkel, 2008). With physiological concentration of the catalytic Mg^{2+} cation, pol θ does not have template-independent terminal transferase activity (Hogg et al., 2012), whereas high ratios of Mn^{2+} over Mg^{2+} appear to trigger template-independent extension (Kent, Chandramouly, McDevitt, Ozdemir, & Pomerantz, 2015; Kent, Mateos-Gomez, Sfeir, & Pomerantz, 2016). Mn^{2+} has been shown to relax template specificity in many polymerases (Vashishtha, Wang, & Konigsberg, 2016). The ability of pol θ to extend single-stranded DNA could thus be largely explained by cycles of templating with other DNA molecules, slippage, and retemplating (Yousefzadeh et al., 2014; Wyatt et al., 2016).

Pol θ -dependent end-joining and its attendant signatures have now been observed in numerous biological processes other than repair of breaks in damaged DNA. Examples include DNA joining during retrohoming of linear group II intron RNAs in *Drosophila* (White & Lambowitz, 2012), T-element insertions in plants (van Kregten, de Pater, Romeijn, van Schendel, Hooykaas, & Tijsterman, 2016) and a fraction of joining during class switch recombination (CSR) of immunoglobulin genes (Yousefzadeh et al., 2014). Many insertions observed upon joining of CRISPR-Cas9 induced breaks are also POLQ-dependent (Mateos-Gomez et al., 2015; van Schendel, Roerink, Portegijs, van den Heuvel, & Tijsterman, 2015). In *C. elegans*, pol θ limits large catastrophic deletions at DNA replication fork barriers, but generates small indels, templated by DNA adjacent to the excision site (Koole et al., 2014; Roerink, van Schendel, & Tijsterman, 2014). In mammalian cells, unprotected telomeres are normally joined by cNHEJ, but can be joined by altEJ as a backup in a POLQ-dependent manner (Mateos-Gomez et al., 2015).

Chromosome translocations can arise in cells when either altEJ or cNHEJ are inactivated, indicating that both processes can promote joining of broken ends of different chromosomes (Wei, Chang, Kao, Du, Meyers, Alt et al., 2016). The Myc-IgH translocation in mice (a model for the oncogenic Burkitt lymphoma translocation) is increased by ~4-fold in *Polq*-defective mice (Yousefzadeh et al., 2014), showing that POLQ protects against translocations in this instance. On the other hand, translocations initiated by CRISPR-Cas9 cleavage of the *Rosa26* locus on mouse chr 6 and the *H3f3b* locus on Chr 11, were decreased in frequency by ~4 fold in *Polq*-defective mouse pluripotent stem cells (Mateos-Gomez et al., 2015). The same translocation in MEF cells showed no significant change in frequency in *Polq*^{-/-} single mutants, and a ~3 fold enhancement in a *Polq*^{-/-} *Ku70*^{-/-} double mutant (Wyatt et al., 2016).

2. Purification and Structure Determination of Human Pol θ

2.1 Full-length Pol θ purified from insect cells

Several reviews have covered the discovery and molecular cloning of *POLQ* orthologs, studies on their biochemical properties, and the consequences of *POLQ* disruption (Beagan et al., 2016; Wood et al., 2016; Yousefzadeh et al., 2013). To produce full-length pol θ protein, a complete *POLQ* ORF was assembled, verified, and cloned into the baculovirus vector pFastBacHTc for protein expression in Sf9 insect cells as described (Seki et al., 2003). Initial experiments with a construct containing an N-terminal 6X His tag showed that the protein was easily degraded during extraction. A FLAG tag was added at the C-terminus to facilitate purification of intact protein. Pol θ was purified by lysis of baculovirus-infected cells in the presence of protease inhibitors. Two methods were successful for recovery of pol θ following lysis. In the first, cells were lysed with low ionic strength buffer containing 0.5% NP40. The nuclear pellet was then back-extracted with buffer containing 0.6 M ammonium sulfate (Seki et al., 2003). In a second approach, cells were directly lysed in buffer containing 0.6 M ammonium sulfate and 0.5% NP-40 (Seki et al., 2004). DNA was then removed by precipitation with polyethyleneimine, and the clarified supernatant was used for sequential purification on FLAG resin and Ni²⁺ resin. The purified protein could be stored in aliquots at -80°C in the Ni²⁺ column elution buffer containing imidazole. Both procedures gave yields of about 100 μg of full-length pol θ per liter of Sf9 culture. The protein was active as a DNA polymerase and DNA-dependent ATPase in biochemical assays (Arana et al., 2008; Seki et al., 2003; Seki et al., 2004; Seki et al., 2008).

2.2 Polymerase Domain Identification and Expression System

The purification schemes from insect cells offered the ability to measure enzymatic activities in the context of the full-length, post-translationally modified protein. For biochemical and structural studies, however, it was necessary to produce milligram amounts of the isolated polymerase domain fold. Determining the catalytic domain of any uncharacterized enzyme involves iterations of construct design, expression tests, and activity screening. This task can be greatly aided by analysis of sequence or structural homologs and *de novo* secondary structure and disorder predictions. Due to the largely unknown characteristics of pol θ , choosing an appropriate construct for expression trials was nontrivial. Pol θ belongs to the A family of DNA polymerases, which share sequence and structural homology with *E. coli* DNA polymerase I (Braithwaite & Ito, 1993; Delarue, Poch, Tordo, Moras, & Argos, 1990). These enzymes include an N-terminal exonuclease (exo) domain with 3'-5' proofreading activity. Early analysis of the pol θ sequence identified the conserved polymerase fold starting at residue 2060 (Seki et al., 2003) and a possible exo domain beginning at residue 1900 (Fig. 2a). Sequence alignments revealed three sequence inserts unique to vertebrates. With limited knowledge of the properties of these large inserts, conservation analysis and structural modeling based on bacterial polymerases such as *E. coli* and *Taq* pol I was used to assign inserts P1, P2, and P3 to residues 2149–2170, 2264–2315, and 2497–2529, respectively, for the human enzyme (Seki et al., 2004) (Fig. 2b). It was speculated that these inserts were involved in the unusual biochemical characteristics of mammalian pol θ . The importance of the exonuclease domain was unclear due to poor sequence conservation and mutation of all catalytic carboxylate residues, which partly explains the unusually low

fidelity of pol θ compared to other A-family polymerases (Arana et al., 2008; Seki et al., 2004).

In preliminary experiments, a minimal construct beginning at residue 2076 near the start of the predicted polymerase fold was not readily expressed. Subsequent work therefore involved large-scale expression of longer constructs, a decision bolstered by the homology between human pol θ and *Drosophila* POLQ/Mus308 extending a further >200 residues (Seki et al., 2003). Disorder predictions using the server PONDR (Dunker, Lawson, Brown, Williams, Romero, Oh et al., 2001; Garner, Romero, Dunker, Brown, & Obradovic, 1999) and secondary structure prediction programs identified the segment spanning residues 1633–2590 as a suitable starting construct for expression and screening trials (Fig. 2c).

Overexpression in *E. coli* was chosen as a logical and cheaper alternative to insect cells for scalable expression of soluble protein absent of eukaryotic modifications that can sometimes interfere with crystallization (Structural Genomics, China Structural Genomics, Northeast Structural Genomics, Graslund, Nordlund, Weigelt et al., 2008). Typically, bacterial expression involves short-term induction of protein expression with isopropyl- β -D-thiogalactoside (IPTG) and the inclusion of affinity tags, generally at the N-terminus, such as His₆, maltose binding protein (MBP), or glutathione S-transferase (GST) tags. Although this strategy is very successful for some proteins, several aspects of the pol θ polymerase domain made it a challenging target. First, the larger size and rare codons characteristic of eukaryotic transcripts often overwhelm the translational machinery of *E. coli*, and proteins containing disordered regions often prove insoluble due to a lack of proper chaperones in bacteria. Second, for proteins >60 kDa, IPTG induction can be inconsistent, depending highly on growth conditions (Studier, 2005). Finally, although affinity tags are often essential for facile purification of recombinant proteins, they can introduce other problems. For example, a GST tag can lead to non-physiological oligomerization of the target protein and an MBP tag can confer solubility but the protein can revert to an insoluble form when the tag is cleaved (Jenny, Mann, & Lundblad, 2003; Sachdev & Chirgwin, 2000; Tudyka & Skerra, 1997; Zanier, Nomine, Charbonnier, Ruhlmann, Schultz, Schweizer et al., 2007).

For these reasons, it was necessary to iteratively assess protein expression in a variety of *E. coli* strains under many growth and induction conditions using plasmids encoding POLQ and various affinity tags. Initial experiments used IPTG expression of pol θ 1633–2590 construct in several *E. coli* host strains, including BL21(DE3)CodonPlus, ArcticExpress(DE3) RIL, and Rosetta2(DE3)pLysS. These strains harbor rare codon tRNAs on a separate, antibiotic selectable plasmid, allowing for protein expression over long time scales at high cell densities. Rosetta2 cells offered the best expression and led to testing of various N-terminal tags including 6xHis, GST, MBP, His-MBP, TRX, protein G B1 domain, NusA, and SUMO. From these trials, SUMO-tagged constructs using slow induction and low IPTG concentration produced some soluble protein, and allowed for further troubleshooting. SUMO is a relatively small (11 kDa) tag often suitable in expression of targets for structural analysis. Specific cleavage by the protease Ulp1 allows easy removal in subsequent purification steps (Mossesso & Lima, 2000; Satakarni & Curtis, 2011). Autoinduction experiments were also carried out in Rosetta2(DE3)pLysS cells (Joachimik, 2009; Studier, 2005), testing variables of media components, aeration, and temperature

control. Greatly improved expression was achieved through autoinduction compared to IPTG induction. After inoculation, 1 L cultures in 2.8 L flasks were grown for 60 hr at 20°C, with saturated cultures reaching a final OD_{600nm} between 6 and 10.

2.3 Expression and Activity screening

Disorder predictions with the PONDR server (Dunker et al., 2001; Garner et al., 1999) as well as secondary structure predictions led to the design of protein constructs beginning at residues 1633, 1792, 1852, 1999, and 2076 (Fig. 2c). These constructs were expressed then purified using a three column purification protocol. Cell pellets were resuspended at a ratio of 4 ml per gram of cell pellet in lysis buffer (40 mM Tris-HCl pH 8.0, 300 mM NaCl, 10% (v/v) glycerol, 20 mM imidazole pH 8.0, 0.01% NP-40, and 5 mM 2-mercaptoethanol) and sonicated. Following clarification by centrifugation, the supernatant was applied to Ni-NTA resin in either column or batch purification. Regardless of method, the column was washed with 10–30 volumes of lysis buffer and eluted with lysis buffer containing 250 mM imidazole pH 8.0. At this point, the protein was <50% pure, with contaminants identified by mass spectrometry as *E. coli* translational and sugar metabolism proteins. Attempts to improve efficiency of this step using doubly N- and C-terminal His tags, shallow imidazole gradients, or washes with Mg²⁺/ATP to release chaperones did not improve protein purity and only decreased overall protein yield.

To efficiently purify low-expressing pol θ samples from abundant contaminants, selection approaches beyond ion exchange chromatography were necessary. Attempts at hydrophobic interaction chromatography proved successful, but were inconsistent with regards to protein solubility or column binding in high ionic strength buffer. Heparin affinity chromatography however greatly improved yields. Eluents from Ni-NTA chromatography were diluted at least 2-fold or dialyzed into buffer (40 mM Tris-HCl pH 8.0, 10% (v/v) Glycerol, and 5 mM 2-mercaptoethanol) containing 0.2 M NaCl, applied to the equilibrated column, washed with 2 column volumes of the same buffer, and eluted with a gradient to 1 M NaCl in this buffer over 20–30 column volumes. Fractions containing pol θ were cleaved using 6x His-tagged Ulp1 for 2 hr to overnight at 4°C, and flowed through Ni²⁺ affinity resin equilibrated with low salt buffer to separate cleaved from uncleaved protein and protease. A final purification step was size exclusion chromatography using a Superdex-200 10/30 column (GE Healthcare) in storage buffer (40 mM Tris-HCl pH 8.0, 150 mM KCl, 150 mM ammonium acetate, 2% (v/v) Glycerol, and 1 mM TCEP). Protein was then concentrated, and frozen rapidly in a liquid nitrogen bath prior to storage at –80°C. Alternatively, Ulp1 protease cleavage and subsequent Ni-NTA separation of cleaved protein can be performed prior to heparin chromatography.

Truncation constructs outlined in Fig. 2c were expressed, purified, and assayed for activity using a polymerase primer extension activity assay (Hogg, Seki, Wood, Doublié, & Wallace, 2011). Through this analysis, constructs 1792–2590 (QM1), and 1633–2590 were found to be active, whereas shorter constructs lacking most of the predicted exonuclease domain had low expression or no activity. Additionally, constructs lacking one of the three insertion elements (Ins1, Ins2, and Ins3) were assayed for extension past undamaged, abasic, or thymine glycol sites. As expected based on its location at the tip of the thumb subdomain,

deletion of insert P1 caused a processivity defect. Deletion of insert P2 or P3 generated protein constructs which exhibited difficulties in extending undamaged DNA and were unable to extend past lesions (Hogg et al., 2011). More generally, the deletion of insert P2 or P3 caused a significant drop in protein yields to below 1 mg per L of culture (compared to 2–5 mg for wild type).

3. Crystallization and Structure Determination of Human Pol θ

3.1 Crystal structure of Pol θ 1792–2590 Bound to Damaged DNA and Nucleotide

Most polymerases undergo a large conformational change upon binding a DNA oligonucleotide and an appropriate incoming nucleotide (Doubl, Sawaya, & Ellenberger, 1999). The protein is said to be in an open conformation when it is unliganded or in a binary complex with either DNA or nucleotide. In a ternary complex, the polymerase finger subdomain is closed upon binding DNA and the correct incoming nucleotide. It is often beneficial to trap ternary polymerase:DNA:nucleotide complexes to improve crystallization odds. Short DNA substrates decrease flexibility observed in the apoprotein state while limiting flexible DNA outside the protein fold, (di)deoxynucleotides allow for closure of the fingers domain, and the addition of Ca^{2+} instead of Mg^{2+} hinders nucleotidyl transfer. The DNA length is a critical factor for success, as varied sequences will alter protein:DNA binding dynamics and contacts involved in crystal packing (Hollis, 2007).

Exploratory work was necessary to ensure crystallization of a homogeneous protein:DNA complex. Although isolation of a protein:nucleic acid complex through gel filtration chromatography prior to crystallization is ideal, the high salt conditions (300 mM total) required for protein solubility at concentrations high enough for crystallization prevented this. To drive complex formation the protein was thus incubated with excess DNA, at a 1:2 molar ratio, without removal of free DNA. Reaction conditions for generating the ternary protein:DNA:nucleotide complex were optimized for incorporation of (di)deoxynucleotides across an abasic site or a normal base, visualizing the products in primer extension assays at varied incubation times, temperatures, and reagent ratios by polyacrylamide gel electrophoresis.

Initial crystallization screening was performed using a sparse matrix representing previously published crystallization conditions of polymerase ternary complexes, along with grid screening from commercial kits optimized for protein:nucleic acid complexes (such as the protein-nucleic acid screen from Kerafast (Pryor, Wozniak, & Hollis, 2012) or Natrix screen from Hampton Research) with varied DNA and nucleotide combinations. These experiments produced no *de novo* crystals but the most promising precipitate was observed with a 13/18mer duplex. Multiple rounds of optimization yielded ordered, microcrystalline precipitate, which was later improved by small molecule additive screens (Hampton Research) yielding very small needle clusters with a variety of DNA duplex substrates. Spermine tetrahydrochloride and sucrose monolaurate were included in the DNA:protein:nucleotide reactions in order to enhance solubility and DNA order during crystallization. After final optimization, thin rectangular parallelepipeds of about 400 x 60 x 60 μm were grown in Tris-HCl pH 8.5 with 50 mM CaCl_2 , 150 mM KCl, 2% MPD, and 9.5% PEG 2000 MME (Zahn, 2013) (Fig. 3a). Cryoprotection was achieved by increasing

the KCl, MPD, and PEG to 260 mM, 25%, and 25%, respectively. Further optimization including streak seeding to increase crystal size was unsuccessful.

These crystals were small and failed to diffract on a copper K- α X-ray source. Attempts to collect data on these crystals using synchrotron radiation presented further challenges. Diffraction data collected at the Diamond synchrotron was limited by significant radiation damage. To overcome this problem, data were collected at the Advanced Photon Source (APS) synchrotron at beamlines 23ID-B and 23ID-D by rastering the beam spherically along the long axis of the crystal, moving to new areas of the crystal after radiation damage. Merging of data from several crystals was sometimes necessary as some single crystals were not robust enough for full data collection. Phases were obtained through molecular replacement with a *Taq* polymerase model (PDBID 1QSY) (Li, Mitaxov, & Waksman, 1999; Zahn, Averill, Aller, Wood, & Doublié, 2015).

3.2 Crystal Structure of Pol θ 1819–2590 with Undamaged Substrate

Although the polymerase domain construct used in crystallization experiments included residues 1792–2590, density was absent for the N-terminal thirty residues of the protein. With this in mind, an attempt was made to improve diffraction of pol θ crystals by N-terminal truncation at residue 1819. Additionally, an undamaged double-stranded oligonucleotide with a template strand lacking a 3' overhang and a primer containing two dCMPs at the 3' end was used to glean more information regarding differentiation between damaged and undamaged substrates, and to promote crystal contacts. With these conditions, crystals were obtained in a manner similar to the longer construct, using an undamaged DNA oligonucleotide, with ddGTP as the incoming nucleotide and incubating with 1 mM MgCl₂. Crystals of selenomethionyl-substituted pol θ were also grown and diffracted to similar resolution as the native protein (Doublié, 2007). Although the selenomethionyl data did not provide experimental phases *per se*, the anomalous signal did prove very useful in verifying the trace of the original model, especially in the exonuclease domain, which shares low sequence identity with the bacterial search model.

4. Conclusions

4.1 Identification of the Exonuclease-like Domain and Flexible loops

Biochemical assays of the pol domain identified the requirement for inclusion of a significant region of a few hundred amino acids beginning at residue 1792, not conserved with bacterial A family polymerases but largely conserved in pol θ from other organisms. Still, secondary structure predictions, conservation analysis, and disorder predictions left doubt as to whether this sequence folded into a *bona fide* structural subdomain. The crystal structure offered the first glimpse into a vestigial Exo-like domain, and pointed to a role for this region in protein:protein interactions at the DNA scaffold. Two structural inserts, labeled exo1 and exo2, which were not readily identified in sequence alignments were revealed by the crystal structure (Fig. 3b). Although the exonuclease domain active site residues are catalytically inactive, loop exo1 appears to provide an additional barrier to DNA entering the inactivated active site. Finally, the globular portion of the Exo-like domain proximal to the inserts makes key dimer contacts (see Fig. 4c). Conservation analysis using EPPIC (Duarte,

Srebniak, Scharer, & Capitani, 2012) and other considerations discussed below predict that such dimerization is biologically relevant.

The sequence inserts assigned in initial sequence alignments were established structurally, and their roles in protein expression and activity could be examined through this structural lens (Seki et al., 2004; Zahn et al., 2015). Insert P1 previously implicated in enzyme processivity was confirmed to project from the thumb, bounded as predicted at residues 2144 and 2177 (Hogg et al., 2011). Although disordered in our crystal structures, this segment lies near the bound oligonucleotide and is thus likely to contribute to DNA binding (Fig. 3c). Having the crystal structure allowed us to refine the beginning and end of each insert. For example, the structural model revealed that a significant portion of Insert P3 that was deleted based on the original sequence alignments (Hogg et al., 2011; Seki et al., 2004) forms a helix at the junction of the palm and exo-like subdomains (Fig. 3b). This observation not only explains the requirement for this insert for expression and overall activity, but also presents a potential for a role of insert P3 in modulating interactions with the exo-like domain.

The base of Insert P2 was ordered in the crystal structure and contributes to one of the most enlightening findings – the unique ability of the thumb and insert P2 to coordinate the extension past DNA lesions (Zahn et al., 2015). Three positively charged residues in the thumb (K2181, R2201, and R2202) and two other residues, R2254 and R2315, located in and proximal to Insert P2 grip the backbone of the primer strand (Fig. 4a). While interactions at the -2 and -5 nt positions of the DNA primer by residues R2315 and R2201, respectively, are conserved in other A-family DNA polymerases, interactions at the -1, -3, and -6 positions are unique to pol θ and explain its enhanced ability to process substrates with oxidized lesions (Zahn et al., 2015). Finally, the presence of the large insert P2 projecting away from the thumb subdomain creates a hole for processing of looped or bulged primer strands, as proposed with pol ν (Beard & Wilson, 2015; Lee, Gao, & Yang, 2015) (Fig. 4b).

4.2 Indications of Functional Dimerization and Future Prospects

Obtaining a construct of pol θ that expresses milligram amounts of active enzyme necessitated sustained effort. Careful analysis of sequence alignments, systematic variation of expression and purification conditions, and attention to subtle changes in crystallization space were all essential for a successful outcome. The structure provides a foundation for understanding recognition of diverse DNA lesions by pol θ as well as the ability of this polymerase to bring together DNA ends in microhomology-mediated end joining. Future research will aim at connecting the individual properties of the polymerase domain with other portions of the protein or other proteins in DNA repair pathways. Of particular interest is the role of intra- and/or interdomain multimerization in both repair and protective roles.

The crystal structures of the individual pol and helicase-like domains show the potential for dimerization of pol θ . Both the Ca^{2+} and Mg^{2+} crystal structures of the Pol domain suggest two potentially biologically relevant assemblies (Fig. 4c). In one of these, two protein chains are related by a two-fold-symmetry axis adjacent to the 5'-template DNA ends. A second two-fold-symmetry axis involving insert 3 relates two pol θ molecules. The helicase-like

domain (HLD) is a tetramer (a dimer of dimers) in solution and *in crystallo*, suggesting that at minimum the enzyme acts as a dimer (Newman, Cooper, Aitkenhead, & Gileadi, 2015). The potential for oligomerization of pol θ is of interest in considering possible mechanisms of end-joining involving pol θ . One hypothesis is that two molecules of pol θ operate on either side of a DNA break (Fig. 1). The HLD may participate in the microhomology annealing step, preparing the annealed substrate for further processing by the polymerase domain (Newman et al., 2015).

In a manner reminiscent of the polymerase domain, the HLD is adorned with several insertion elements, one of which is involved in tetramerization (Newman et al., 2015) (Fig 2a). The HLD relies on unique helices in insert H4 that contribute to the formation of a tetrameric interface, a sharp departure from the monomeric archaeal SF2 family helicases (Newman et al., 2015). The HLD exhibits DNA-dependent ATPase function, but no helicase activity (Seki et al., 2003; Newman et al., 2015). Disruption of the ATPase activity in HLD did not overtly alter the correcting function of pol θ addition to knockout cells (Yousefzadeh et al., 2014). Two recent studies, however, have suggested that the ATPase function of pol θ alters the outcome of repair products. When POLQ-deficient MEF cells expressing wild-type human pol θ are supplied with a DNA template designed with 3' single-stranded overhangs containing microhomologies, many repair products are formed using internal rather than terminal microhomologies (Wyatt et al., 2016). However, usage of internal microhomologies is significantly reduced in POLQ-deficient MEF expressing ATPase-dead human pol θ (Wyatt et al., 2016). This observation suggests that the HLD may promote a versatile homology search in single-stranded DNA. Recent work in *Drosophila* indicates that while the ATPase function is not necessary for repair of a transposase-induced double-strand break, insertions at the repair junction are more dependent on function of the ATPase domain of pol θ (Beagan and McVey, In Press).

Finally, a major gap in our understanding of pol θ relates to a possible function of the central region, which is substantially shorter in invertebrates (Yousefzadeh et al., 2013). A large portion of this ~900 residue region is predicted to be disordered in the human protein. However, given the importance of other flexible regions in both the helicase-like and polymerase domains it is reasonable to expect some role for this segment. The region may bind proteins involved in alternate repair pathways to modulate the balance between homologous recombination and alternate end-joining pathways. Alternatively, it could contribute to protein organization at sites of DNA damage or chromosome ends or act as a hinge to link helicase-like and pol domains working in tandem across large or specialized DNA substrates.

Acknowledgments

We acknowledge the contributions of Karl Zahn, Matt Hogg, April Averill, Pierre Aller, Susan Wallace, Mineaki Seki, Federica Marini, Kei-ichi Takata, Lee Wei Yang, and Ivet Bahar to the work described here. We appreciate the helpful comments on the manuscript provided by Karl Zahn. These studies were funded by National Institutes of Health grants R01 CA052040 (S.D.) and CA097175 (R.D.W.) and grant RP130297 from the Cancer Prevention and Research Institute of Texas (R.D.W.) and the Grady F. Saunders Ph.D. Distinguished Research Professorship (R.D.W.).

Abbreviations used

DSB	double-strand break
IR	ionizing radiation
altEJ	alternative end-joining
NHEJ	non-homologous end-joining
MN	micronuclei
pol θ	DNA polymerase theta

References

- Arana ME, Seki M, Wood RD, Rogozin IB, Kunkel TA. Low-fidelity DNA synthesis by human DNA polymerase θ . *Nucleic Acids Res.* 2008; 36:3847–3856. [PubMed: 18503084]
- Beagan K, McVey M. Linking DNA polymerase theta structure and function in health and disease. *Cell Mol Life Sci.* 2016; 73:603–615. [PubMed: 26514729]
- Beard WA, Wilson SH. Structures of human DNA polymerases [nu] and [θ] expose their end game. *Nat Struct Mol Biol.* 2015; 22:273–275. [PubMed: 25837870]
- Braithwaite DK, Ito J. Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res.* 1993; 21:787–802. [PubMed: 8451181]
- Ceccaldi R, Liu JC, Amunugama R, Hajdu I, Primack B, Petalcorin MI, et al. Homologous-recombination-deficient tumours are dependent on Poltheta-mediated repair. *Nature.* 2015; 518:258–262. [PubMed: 25642963]
- Chan SH, Yu AM, McVey M. Dual Roles for DNA Polymerase θ in Alternative End-Joining Repair of Double-Strand Breaks in *Drosophila*. *PLoS Genet.* 2010; 6:e1001005. [PubMed: 20617203]
- Delarue M, Poch O, Tordo N, Moras D, Argos P. An attempt to unify the structure of polymerases. *Protein Eng.* 1990; 3:461–467. [PubMed: 2196557]
- Doublé S. Production of selenomethionyl proteins in prokaryotic and eukaryotic expression systems. *Methods Mol Biol.* 2007; 363:91–108. [PubMed: 17272838]
- Doublé S, Sawaya MR, Ellenberger T. An open and closed case for all polymerases. *Structure.* 1999; 7:R31–35. [PubMed: 10368292]
- Duarte JM, Srebniak A, Scharer MA, Capitani G. Protein interface classification by evolutionary analysis. *BMC Bioinformatics.* 2012; 13:334. [PubMed: 23259833]
- Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. *J Mol Graph Model.* 2001; 19:26–59. [PubMed: 11381529]
- Garner E, Romero P, Dunker AK, Brown C, Obradovic Z. Predicting Binding Regions within Disordered Proteins. *Genome Inform Ser Workshop Genome Inform.* 1999; 10:41–50.
- Goff JP, Shields DS, Seki M, Choi S, Epperly MW, Dixon T, Wang H, Bakkenist CJ, Dertinger SD, Torous DK, Wittschieben J, Wood RD, Greenberger JS. Lack of DNA polymerase θ (POLQ) radiosensitizes bone marrow stromal cells *in vitro* and increases reticulocyte micronuclei after total-body irradiation. *Radiat Res.* 2009; 172:165–74. [PubMed: 19630521]
- Higgins GS, Prevo R, Lee YF, Helleday T, Muschel RJ, Taylor S, et al. A small interfering RNA screen of genes involved in DNA repair identifies tumor-specific radiosensitization by POLQ knockdown. *Cancer Res.* 2010; 70:2984–2993. [PubMed: 20233878]
- Hogg M, Sauer-Eriksson AE, Johansson E. Promiscuous DNA synthesis by human DNA polymerase θ . *Nucleic Acids Res.* 2012; 40:2611–2622. [PubMed: 22135286]
- Hogg M, Seki M, Wood RD, Doublé S, Wallace SS. Lesion bypass activity of DNA polymerase theta (POLQ) is an intrinsic property of the pol domain and depends on unique sequence inserts. *J Mol Biol.* 2011; 405:642–652. [PubMed: 21050863]

- Hollis T. Crystallization of protein-DNA complexes. *Methods Mol Biol.* 2007; 363:225–237. [PubMed: 17272844]
- Jenny RJ, Mann KG, Lundblad RL. A critical review of the methods for cleavage of fusion proteins with thrombin and factor Xa. *Protein Expr Purif.* 2003; 31:1–11. [PubMed: 12963335]
- Joachimiak A. High-throughput crystallography for structural genomics. *Curr Opin Struct Biol.* 2009; 19:573–584. [PubMed: 19765976]
- Kent T, Chandramouly G, McDevitt SM, Ozdemir AY, Pomerantz RT. Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase theta. *Nat Struct Mol Biol.* 2015; 22(3):230–237. [PubMed: 25643323]
- Kent T, Mateos-Gomez PA, Sfeir A, Pomerantz RT. Polymerase theta is a robust terminal transferase that oscillates between three different mechanisms during end-joining. *Elife.* 2016:5.
- Koole W, van Schendel R, Karambelas AE, van Heteren JT, Okihara KL, Tijsterman M. A Polymerase Theta-dependent repair pathway suppresses extensive genomic instability at endogenous G4 DNA sites. *Nat Commun.* 2014; 5:3216. [PubMed: 24496117]
- Lee YS, Gao Y, Yang W. How a homolog of high-fidelity replicases conducts mutagenic DNA synthesis. *Nat Struct Mol Biol.* 2015; 22:298–303. [PubMed: 25775266]
- Lemée F, Bergoglio V, Fernandez-Vidal A, Machado-Silva A, Pillaire MJ, Bieth A, et al. DNA polymerase theta up-regulation is associated with poor survival in breast cancer, perturbs DNA replication, and promotes genetic instability. *Proc Natl Acad Sci (USA).* 2010; 107:13390–13395. [PubMed: 20624954]
- Li Y, Mitaxov V, Waksman G. Structure-based design of Taq DNA polymerases with improved properties of dideoxynucleotide incorporation. *Proc Natl Acad Sci U S A.* 1999; 96:9491–9496. [PubMed: 10449720]
- Mateos-Gomez PA, Gong F, Nair N, Miller KM, Lazzerini-Denchi E, Sfeir A. Mammalian polymerase θ promotes alternative NHEJ and suppresses recombination. *Nature.* 2015; 518:254–257. [PubMed: 25642960]
- McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.* 2013; 41:W597–600. [PubMed: 23671338]
- Mossessova E, Lima CD. Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol Cell.* 2000; 5:865–876. [PubMed: 10882122]
- Newman JA, Cooper CD, Aitkenhead H, Gileadi O. Structure of the Helicase Domain of DNA Polymerase Theta Reveals a Possible Role in the Microhomology-Mediated End-Joining Pathway. *Structure.* 2015; 23:2319–2330. [PubMed: 26636256]
- Pryor EE Jr, Wozniak DJ, Hollis T. Crystallization of *Pseudomonas aeruginosa* AmrZ protein: development of a comprehensive method for obtaining and optimization of protein-DNA crystals. *Acta Crystallogr Sect F Struct Biol Cryst Commun.* 2012; 68:985–993.
- Roerink SF, van Schendel R, Tijsterman M. Polymerase theta-mediated end joining of replication-associated DNA breaks in *C. elegans*. *Genome Res.* 2014; 24:954–962. [PubMed: 24614976]
- Sachdev D, Chirgwin JM. Fusions to maltose-binding protein: control of folding and solubility in protein purification. *Methods Enzymol.* 2000; 326:312–321. [PubMed: 11036650]
- Satakarni M, Curtis R. Production of recombinant peptides as fusions with SUMO. *Protein Expr Purif.* 2011; 78:113–119. [PubMed: 21586326]
- Seki M, Marini F, Wood RD. POLQ (Pol θ), a DNA polymerase and DNA-dependent ATPase in human cells. *Nucleic Acids Res.* 2003; 31:6117–6126. [PubMed: 14576298]
- Seki M, Masutani C, Yang LW, Schuffert A, Iwai S, Bahar I, et al. High-efficiency bypass of DNA damage by human DNA polymerase Q. *EMBO J.* 2004; 23:4484–4494. [PubMed: 15496986]
- Seki M, Wood RD. DNA polymerase θ (POLQ) can extend from mismatches and from bases opposite a (6–4) photoproduct. *DNA Repair (Amst).* 2008; 7:119–127. [PubMed: 17920341]
- Shima N, Munroe RJ, Schimenti JC. The mouse genomic instability mutation *chaos1* is an allele of *Polq* that exhibits genetic interaction with *Atm*. *Mol Cell Biol.* 2004; 24:10381–10389. [PubMed: 15542845]

- Graslund S, Nordlund P, Weigelt J, et al. Structural Genomics C, China Structural Genomics C, Northeast Structural Genomics C. Protein production and purification. *Nat Methods*. 2008; 5:135–146. [PubMed: 18235434]
- Studier FW. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif*. 2005; 41:207–234. [PubMed: 15915565]
- Tudyka T, Skerra A. Glutathione S-transferase can be used as a C-terminal, enzymatically active dimerization module for a recombinant protease inhibitor, and functionally secreted into the periplasm of *Escherichia coli*. *Protein Sci*. 1997; 6:2180–2187. [PubMed: 9336840]
- van Kregten M, de Pater S, Romeijn R, van Schendel R, Hooykaas PJ, Tijsterman M. T-DNA integration in plants results from polymerase-theta-mediated DNA repair. *Nat Plants*. 2016; 2:16164. [PubMed: 27797358]
- van Schendel R, Roerink SF, Portegijs V, van den Heuvel S, Tijsterman M. Polymerase θ is a key driver of genome evolution and of CRISPR/Cas9-mediated mutagenesis. *Nat Commun*. 2015; 6:7394. [PubMed: 26077599]
- Vashishtha AK, Wang J, Konigsberg WH. Different Divalent Cations Alter the Kinetics and Fidelity of DNA Polymerases. *J Biol Chem*. 2016; 291(40):20869–20875. [PubMed: 27462081]
- Wei PC, Chang AN, Kao J, Du Z, Meyers RM, Alt FW, et al. Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. *Cell*. 2016; 164:644–655. [PubMed: 26871630]
- White TB, Lambowitz AM. The retrohoming of linear group II intron RNAs in *Drosophila melanogaster* occurs by both DNA ligase 4-dependent and -independent mechanisms. *PLoS Genet*. 2012; 8:e1002534. [PubMed: 22359518]
- Wood RD, Doublé S. DNA polymerase theta (POLQ), double-strand break repair, and cancer. *DNA Repair (Amst)*. 2016; 44:22–32. [PubMed: 27264557]
- Wyatt DW, Feng W, Conlin MP, Yousefzadeh MJ, Roberts SA, Mieczkowski P, et al. Essential Roles for Polymerase theta-Mediated End Joining in the Repair of Chromosome Breaks. *Mol Cell*. 2016; 63:662–673. [PubMed: 27453047]
- Yousefzadeh MJ, Wood RD. DNA polymerase POLQ and cellular defense against DNA damage. *DNA Repair (Amst)*. 2013; 12:1–9. [PubMed: 23219161]
- Yousefzadeh MJ, Wyatt DW, Takata K, Mu Y, Hensley SC, Tomida J, et al. Mechanism of suppression of chromosomal instability by DNA polymerase POLQ. *PLoS Genet*. 2014; 10:e1004654. [PubMed: 25275444]
- Zahn, KE. Structural studies of DNA polymerases (Unpublished doctoral dissertation). University of Vermont; Burlington, VT: 2013.
- Zahn KE, Averill AM, Aller P, Wood RD, Doublé S. Human DNA polymerase θ grasps the primer terminus to mediate DNA repair. *Nat Struct Mol Biol*. 2015; 22:304–311. [PubMed: 25775267]
- Zanier K, Nomine Y, Charbonnier S, Ruhlmann C, Schultz P, Schweizer J, et al. Formation of well-defined soluble aggregates upon fusion to MBP is a generic property of E6 proteins from various human papillomavirus species. *Protein Expr Purif*. 2007; 51:59–70. [PubMed: 17055740]

Figure 1.

Pol θ mediated double-strand break repair. Double-strand breaks can arise from multiple sources including ionizing radiation, break-inducing chemicals or broken replication forks (A). Pol θ plays a role in one route of repair of double-strand breaks. 5' to 3' resection of DNA ends by nucleases (B) produces DNA tails with 3' termini. The pol θ helicase-like domain (HLD) and polymerase (POL) domains are connected by a central region (CEN) (C). The HLD is a DNA-dependent ATPase that likely acts at the junction of single stranded and double stranded DNA. The HLD and pol domain each have surface elements that promote dimerization or tetramerization, which may facilitate the association of the two ends of an enzymatically processed DNA double-strand break. Pol θ-mediated repair of double-strand breaks has signature outcomes. Pol θ facilitates association of regions of microhomologies (MH, shown here as oblong shapes), and synthesizes DNA to fill in gaps. If either microhomology is internal, DNA flap cleavage leads to a short deletion (D). In some cases, pol θ mediated repair of a double-strand break also produces insertions (E). The insertions arise by pol θ templated synthesis from DNA in the tail or from a more distant template

(shown as a green strand). The break can then be joined via annealing of *de novo* microhomologies between the newly synthesized DNA and the other end of the break. The repair is completed by flap cleavage and ligation by Ligase I or III.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2. Conserved sequence elements and expression constructs of human Pol θ. (a) Representation of the human pol θ protein with relevant subdomains and conserved sequence inserts labeled. WH: winged helix, Hel: helical domain, HhH: Helix-hairpin-Helix motif, Exo: exonuclease-like domain, P: Palm, Fg: Fingers. Numbers designate start and end points of protein sequences of crystal structures for HLD and pol domains. (b) CLUSTAL OMEGA alignment of select pol θ sequences surrounding Insert P1 (top), Insert P2 (middle) and Insert P3 (bottom). Insert boundaries as defined prior to the presented crystal structure are depicted below each alignment. Similarity groups and conservation limits for shading are in CLUSTALW format (McWilliam et al., 2013). (c) Summary for results of expression trials and biochemical analysis of human pol θ polymerase domain constructs.

Figure 3.

Crystallization and structure of the pol θ polymerase domain: (a) Collage illustrating optimization of Pol θ crystals ranging from initial hits at top left to final rod shaped crystals moving counterclockwise. i,ii: microcrystalline clusters; iii: needle clusters; iv: needles nucleated on a fiber; v: rod-shaped crystals with a growth defect; vi: diffraction-quality parallelepiped crystals (b) Pol θ polymerase domain crystal structure with zoomed view of pol θ exo loops and Insert P3. c) Pol θ pol domain crystal structure showing DNA binding site. Inserts P1 and P2 are in the forefront whereas Insert P3 is in the back of the figure, between the palm and exonuclease subdomains.

Figure 4.

Unique residue contacts with DNA substrate. (a) View of basic residues in the thumb subdomain making unique contacts with the primer strand of duplex DNA. (b) Surface representation of pol θ polymerase domain showing a hole (arrow) that could potentially accommodate bulged substrates, analogous to that in pol ν (Lee, Gao, & Yang, 2015) (c) Dimer interfaces of the pol domain showing a dimer formed with chains C and D, and a DNA-mediated dimer between chains A and B. Interfacial residues as indicated by the server EPPIC (Duarte et al., 2012) are colored orange.