Structural basis for a novel mechanism of DNA bridging and alignment in eukaryotic DSB DNA repair

Jérôme Gouge^{1,†}, Sandrine Rosario¹, Félix Romain¹, Frédéric Poitevin², Pierre Béguin³ & Marc Delarue^{1,*}

Abstract

Eukaryotic DNA polymerase mu of the PolX family can promote the association of the two 3'-protruding ends of a DNA double-strand break (DSB) being repaired (DNA synapsis) even in the absence of the core non-homologous end-joining (NHEJ) machinery. Here, we show that terminal deoxynucleotidyltransferase (Tdt), a closely related PolX involved in V(D)J recombination, has the same property. We solved its crystal structure with an annealed DNA synapsis containing one micro-homology (MH) base pair and one nascent base pair. This structure reveals how the N-terminal domain and Loop 1 of Tdt cooperate for bridging the two DNA ends, providing a templating base *in trans* and limiting the MH search region to only two base pairs. A network of ordered water molecules is proposed to assist the incorporation of any nucleotide independently of the *in trans* templating base. These data are consistent with a recent model that explains the statistics of sequences synthesized *in vivo* by Tdt based solely on this dinucleotide step. Site-directed mutagenesis and functional tests suggest that this structural model is also valid for Pol mu during NHEJ.

Keywords DNA repair; DNA synapsis; micro-homology base pair; non-homologous end-joining; X-ray crystallography

Subject Categories DNA Replication, Repair & Recombination; Structural Biology

DOI 10.15252/embj.201489643 | Received 29 July 2014 | Revised 8 January 2015 | Accepted 9 January 2015 | Published online 11 March 2015

The EMBO Journal (2015) 34: 1126–1142

Introduction

Double-strand breaks (DSB) in DNA must be repaired efficiently and rapidly to prevent genomic instability. The non-homologous end-joining (NHEJ) repair system is the predominant DNA repair pathway system in higher eukaryotes (for a recent review see Waters

et al, 2014). NHEJ requires numerous proteins including Ku 70-80, DNA-PKc, a nuclease (Artemis, Metnase), a ligase (Ligase IV) and a polymerase (Pol mu or Pol lambda). All the components of this system must be flexible enough to deal with the different possible substrates arising at a DNA DSB (Lieber, 2008; Lieber *et al*, 2008).

Here, we focus on the polymerases of the eukaryotic NHEJ machinery, which belong to the PolX polymerase family (Moon *et al*, 2007). The crystal structures of the four different eukaryotic PolX have been solved: Pol beta (Sawaya *et al*, 1997), Pol lambda (Garcia-Diaz *et al*, 2004), Pol mu (Moon *et al*, 2007) and Tdt (Delarue *et al*, 2002). While Pol beta is mainly involved in single-stranded break (SSB) DNA repair, the others are involved in DSB repair. Pol lambda (Bebenek *et al*, 2014) and Pol mu directly intervene in NHEJ (Aoufouchi *et al*, 2000; Domínguez *et al*, 2000; Chayot *et al*, 2010, 2012), and Pol mu has been shown to have a gradient of different activities (Nick McElhinny *et al*, 2005) with more or less tolerance/efficiency with respect to the various substrates it encounters at a DNA synapsis. Tdt, which shares 42% sequence identity with Pol mu, is located at the extreme end of the spectrum of possible Pol mu activities, because it is completely template independent. Tdt was one of the first discovered eukaryotic DNA polymerases, purified in 1960 by F.J. Bollum in calf thymus cells and subsequently fully characterized biochemically (Kato *et al*, 1967; Bollum, 1978) and cloned (Peterson *et al*, 1984). *In vivo*, Tdt is involved in V(D)J recombination, where its biological role is to generate junctional diversity at the so-called N regions at VD or DJ junctions in immunoglobulin heavy chains and T-cell receptors (Landau *et al*, 1987; Benedict *et al*, 2000). During V(D)J recombination, Tdt is part of a larger macromolecular complex that contains the same partners as in NHEJ: Ku 70-80 (Mahajan *et al*, 1999), DNA-PKc (Mickelsen *et al*, 1999), Artemis nuclease, XRCC4, XLF and Ligase IV (Malu *et al*, 2012).

Recent biochemical and biophysical experiments have suggested that Pol mu directly promotes the physical alignment of a DNA synapsis (Martin *et al*, 2012). However, there is currently no structural model of such an association. Indeed, despite recent progress in

¹ Unité de Dynamique Structurale des Macromolécules, Institut Pasteur, UMR 3528 du C.N.R.S., Paris, France

² Institut de Physique Théorique, CEA-Saclay, CNRS URA 2306, Gif-sur-Yvette, France

³ Unité de Biologie Moléculaire du Gène chez les Extrémophiles, Institut Pasteur, Paris, France

*Corresponding author. Tel: +33 1 45 68 86 05; E-mail: marc.delarue@pasteur.fr

[†]Present address: Institute of Cancer Research, London, UK

the field of Pol mu structure and function (Moon *et al*, 2014), the only available structures of Pol mu bound to a DNA substrate involve the so-called gap-filling complex, where the DNA is not a DSB substrate and where Loop1 is seen to be completely disordered (Moon *et al*, 2007, 2014). This is incompatible with the finding that Loop1 is important for the formation of the DNA synapsis complex (Juárez *et al*, 2006; Esteban *et al*, 2013). Elucidating the structural details of the interaction of the two DNA ends of a DNA synapsis during NHEJ, especially on a PolX or PolX-related polymerase, is necessary to provide a detailed understanding of the molecular mechanisms governing eukaryotic NHEJ.

Structural information on how this close association between a polymerase and a DNA synapsis takes place in the NHEJ bacterial system was recently obtained by A.J. Doherty and colleagues through structural studies of LigD (Brissett *et al*, 2007, 2011, 2013), a prokaryotic NHEJ polymerase. The proposed model implies the dimerization of LigD at the synapsis. However, it is not known whether this structural model is valid for eukaryotic systems, especially because LigD is not a PolX but rather a member of the structurally unrelated archaea-eukaryotic primase family.

Up to now, it was believed that the principal DNA substrate of Tdt was a single DNA molecule with a 3'-overhang, with no templating base. Here, we describe that Tdt also promotes a close physical association of the two DNA ends of a DNA synapsis, both with a 3'-protruding end, using its Loop1 and N-terminal (8 kDa) domain as a mold. We also directly solve the crystal structures of Tdt with various DNA substrates mimicking a DNA synapsis with a templating base *in trans*. These structural data imply that, similar to Pol mu (Martin *et al*, 2012), Tdt has the intrinsic capacity to direct DNA synapsis pre-assembly and alignment.

We then provide further evidence of the relatedness of Tdt and Pol mu by comparing the effect of single point mutations in the two proteins, with particular emphasis on the wedge induced by Tdt in the DNA by the two side chains L398 and F405, that insert between the 3'-end last and penultimate bases of the primer. Indeed, mutating these residues in mouse Pol mu (M384A and F391A, respectively) strongly suggests that the structural model of a DNA synapsis bound to Tdt is also valid for Pol mu.

Results

Crystal structures of wild-type Tdt bound to a DNA synapsis

We crystallized Tdt in the presence of a primer strand A₅, an incoming nucleotide ddCTP and a 'downstream' DNA duplex with a

3'-protruding end *in trans* (Fig 1A). After one round of DNA synthesis, the primer becomes A₅C with no 3'-OH group, preventing the reaction from proceeding any further. Since the 3'-end of the downstream template strand ends with two overhanging G, a so-called micro-homology base pair (MH-bp) can be formed *in trans*. Indeed, we observed formation of the MH-bp and the nascent one (Fig 1B) and very clear electron density for all partners of the complex (Fig 1C). The incoming ddCTP is engaged in a Watson–Crick base pair with an *in trans* templating G that comes from the downstream DNA duplex. These two base pairs form a continuous double helix with clear electron density (Figs 1C and 2A), but with a helical axis different from both the upstream primer strand and the downstream DNA duplex, which closely follows the path of the DNA seen in DNA Pol beta (Sawaya *et al*, 1997), or Pol lambda (Garcia-Diaz *et al*, 2004, 2006) or Pol mu (Moon *et al*, 2007, 2014) gap-filling complexes. Moreover, we see a break in the helical path of the 3'-side of the primer strand just before the MH-bp, in a manner very similar to what was recently described for Tdt pre-catalytic ternary complex (Gouge *et al*, 2013), but obviously different from what is seen in the Pol mu gap-filling complex (Moon *et al*, 2007, 2014). The two side chains of L398 and F405 in Loop1 are responsible for creating this wedge in the primer strand (Figs 1D and 2C).

Below the nascent ddCTP-G base pair, the side chains of R454 and R458 become more ordered and change rotamers compared to previously known Tdt structures, blocking one side of the nascent base pair (Figs 1D and 2B). This is consistent with the role of these two conserved side chains for catalysis that has been underlined by Molecular Dynamics simulations (Li & Schlick, 2013). These side chains, along with L398 and F405, are literally isolating two base pairs (micro-homology and nascent) from the rest of the DNA substrate.

As shown in Fig 2B and Supplementary Fig S1, both bases participating in the nascent base pair (C-G) are recognized in the minor groove by a network of hydrogen bonds involving strictly conserved side chains located in the previously identified (Romain *et al*, 2009) important regions called SD1 and SD2, short for Substrate Specificity Sequence Determinant (Supplementary Fig S1): N474 (SD2), R461 (Helix N), whose rotamer changes compared to other known Tdt structures, and the carbonyl atom of G449. N474 itself is hydrogen-bonded to D399 (SD1), which forms a hydrogen bond with the strictly conserved W450 and a salt bridge with K403 (SD1) (Gouge *et al*, 2013).

It is possible to build two slightly different conformations of the MH base coming from the template strand, both stacked with the templating base of the nascent base pair and capped by the Loop1 backbone (carbonyl of residue 397). Thus, the MH-bp is not really held

Figure 1. Overall view of Tdt in the DNA DSB complexes and comparison with Pol mu in the gap-filling complex.

- Schematic views of Tdt (left) in the DSB complex and Pol mu (pdb id: 2IHM; Moon *et al*, 2007) in a gap-filling complex (right). The upstream primer strand is depicted in red, the downstream primer is in dark blue, the template strand is colored in cyan, and the incoming nucleotide is in pink.
- Overall view of Tdt (left) and Pol mu (right) in their respective complex. The downstream double-stranded DNA is hybridized to the upstream primer strand by only one micro-homology (MH) base pair, and the nascent base pair. Loop1 is depicted in purple, the SD1 and the SD2 regions (Gouge *et al*, 2013) in orange and in brown, respectively. In the Pol mu gap-filling complex, the template strand is continuous. DNA coloring as in (A).
- In the Tdt DSB complexes (left), the MH and the nascent base pairs are isolated from the rest of the DNA by Loop1 (in purple). L398 and F405 form a wedge that disrupts the usual stacking of the primer strand. In Pol mu structure (right), Loop1 is disordered. The density of the DNA and the incoming nucleotide is in blue and contoured at 1 σ in Tdt structures. DNA coloring as in (A).
- The interactions between the Tdt (left) or Pol mu (right) and the DNA are depicted as a cartoon, highlighting side chains important for the recognition. Positions ending with N make contact through their backbone atoms (side-chain atoms otherwise). The important ions are in lozenges. Catalytic ones are in green. Metal A is shown in a transparent mode because it is not always seen in Pol mu nor in Tdt in the absence of the 3'-OH group of the primer (Gouge *et al*, 2013). DNA coloring as in (A).

Figure 1.

Figure 2. Base pairing at the level of the MH and nascent base pairs.

A General view of the MH and nascent base pairs. The left panel shows the reorganization of Loop1 in the DSB complex, compared to the binary complex with a ssDNA (in transparent brown, pdb id 4I28), that allows the Tdt to 'cap' the downstream template strand. In the right panel, the electron density of the MH-bp is shown when the base on the template strand is varied. When a C is facing a G, two stacked conformations are observed. When the base pair is not a Watson–Crick, two different conformations (one stacked, one unstacked) are observed. The percentage of stacked conformation(s) is indicated.

B The nascent base pair density is always well defined, and several residues help to stabilize and isolate the nascent and the MH base pairs from the rest of the DNA: R454, R458 and R461 of Helix N.

C Two residues in Loop1 and SD1 region, L398 and F405 create a wedge (represented here in space-filling mode) in the primer strand.

together by specific hydrogen bonds between the bases but rather by interactions involving the SD1 region and the preceding base.

On the DNA duplex side of the complex, Tdt clamps the 5' end of the primer strand of the downstream DNA duplex, using helices $\alpha 2$

and $\alpha 3$ (residues 212, 220, 226 in Tdt, Fig 1D), which would be the equivalent of the RP-lyase site in Pol beta and Pol lambda (García-Díaz *et al*, 2001). There is no specific side chain here to bind a 5'-phosphate group on this strand, as in Pol mu (the equivalent of

Table 1. Diffraction data collection and refinement statistics.

Micro-homology base pair	C-A	C-C	C-G	C-T
Protein	WT	WT	WT	WT
PDB ID	4QZ9	4QZA	4QZ8	4QZB
Space group	C2	C2	C2	C2
Cell parameters	56.94 74.72 126.36 90 100.06 90	56.88 74.78 127.15 90 99.14 90	59.33 74.26 118.97 90 97.98 90	57.11 75.20 127.62 90 99.58 90
Resolution shell	45–2.05	45–2.15	41–2.7	45–2.15
No. of observations*	101,432 (14,668)	131,301 (19,609)	57,630 (8,618)	108,040 (15,925)
No. of unique reflections*	32,370 (4,671)	28,715 (4,188)	14,082 (2,042)	28,987 (4,218)
Multiplicity	3.1	4.6	4.2	3.7
Completeness (%)*	98.9 (97.8)	99.9 (99.9)	99.3 (99.8)	99.8 (100)
Rmerge (%)*	4.0 (50.3)	5.4 (80.9)	6.0 (64.1)	4.9 (60.8)
I/ σ (I)*	15.7 (2.5)	15.5 (2.2)	14.4 (2.3)	14.3 (2.3)
R (%)	18.6	19.4	20.1	19.2
Rfree (%)	20.3	21.5	24.0	22.1
No. of protein atoms	2,782	2,813	2,680	2,805
No. of DNA atoms	408	402	389	385
No. of water molecules	333	267	88	247
B factor overall ($\text{\AA}^2$)	49.6	54.6	88.7	54.1
B factor DNA ($\text{\AA}^2$)	56.4	73.0	95.4	60.1
Ramachandran outliers (#)	0	0	0	0
RMSD bond lengths ($\text{\AA}$)	0.01	0.01	0.01	0.01
RMSD bond angles ($^\circ$)	0.97	0.99	0.97	0.96
Micro-homology base pair	C-A	C-C	C-G	C-T
Protein	F401A	F401A	F401A	F401A
PDB ID	4QZF	4QZG	4QZE	4QZH
Space group	C2	C2	C2	C2
Cell parameters	56.43 74.96 125.63 90 97.78 90	56.99 70.85 114.89 90 96.16 90	58.03 71.61 113.95 90 95.19 90	56.31 75.01 125.61 90 97.64 90
Resolution shell	45–2.6	44–2.75	41–2.25	45–2.6
No. of observations*	52,010 (6,511)	48,416 (7,208)	93,813 (13,826)	62,594 (9,107)
No. of unique reflections*	15,816 (2,231)	11,937 (1,736)	22,175 (3,225)	16,036 (2,318)
Multiplicity	3.3	4.1	4.2	3.9
Completeness*	98.4 (96.4)	99.8 (100)	99.9 (100)	99.8 (99.9)
Rmerge (%)*	5.0 (52.6)	8.3 (78.6)	4.5 (64.0)	5.1 (65.5)
I/ σ (I)*	16.5 (2.3)	11.1 (2.1)	18.6 (2.4)	16.1 (2.1)
R (%)	22.4	19.0	19.4	20.4
Rfree (%)	26.3	21.2	21.8	23.8
No. of protein atoms	2,678	2,601	2,636	2,622
No. of DNA atoms	408	401	389	360
No. of water molecules	181	44	200	107
B factor overall ($\text{\AA}^2$)	72.5	80.7	62.7	84.1
B factor DNA ($\text{\AA}^2$)	79.6	73.0	62.4	60.1
Ramachandran outliers (#)	0	0	0	0
RMSD bond lengths ($\text{\AA}$)	0.01	0.01	0.01	0.01
RMSD bond angles ($^\circ$)	0.99	1.00	0.99	0.98

Table 1. (continued)

Micro-homology base pair	C-G	C-C	A-G
Protein	F405A	F405A	F401A
PDB ID	4QZC	4QZD	4QZI
Space group	C2	C2	C2
Cell parameters	56.40 75.18 125.94 90 99.05 90	55.93 74.92 125.63 90 98.43 90	57.70 71.52 113.04 90 94.17 90
Resolution shell (Å)	45–2.75	44–2.70	44–2.65
No. of observations*	57,255 (8,466)	78,706 (11,313)	54,843 (8,173)
No. of unique reflections*	13,623 (1,981)	14,140 (1,990)	13,390 (1,937)
Multiplicity	4.3	5.6	4.1
Completeness*	99.9 (100)	99.8 (99.5)	99.6 (100)
Rmerge (%)*	6.9 (63.9)	6.9 (83.0)	6.3 (79.5)
I/σ(I)*	14.8 (2.2)	15.3 (2.0)	13.6 (2.1)
R (%)	21.1	22.6	20.9
Rfree (%)	24.4	27.0	22.3
No. of protein atoms	2,729	2,753	2,668
No. of DNA atoms	367	346	367
No. of water molecules	85	36	52
B factor overall (Å ²)	78.6	81.7	75.3
B factor DNA (Å ²)	101.8	95.3	72.8
Ramachandran outliers (#)	0	0	0
RMSD bond lengths (Å)	0.01	0.01	0.01
RMSD bond angles (°)	1.02	0.99	1.09

The asterisk corresponds to the value for the highest resolution shell.

R175 in Pol mu is S187 in Tdt), but there is enough room to accommodate it. The 5′-end nucleobase of the downstream primer strand is stacked under the 186–187 peptide bond, with a characteristic distance of 3.4 Å between them (not shown). Interestingly, the N-terminus of the 8 kDa domain, recently shown to be important for DNA end-bridging in Pol mu (Martin *et al*, 2013), contains residues interacting with the *in trans* DNA duplex involving Q152 and Y153 in Tdt (positions 140–141 in Pol mu) and a DNA phosphate (Fig 1D). Analysis of crystal packing reveals that the downstream DNA duplex forms a continuous double helix (10 bp long) with another DNA duplex molecule in the crystal lattice.

Influence of the base pairing at the MH locus: base stacking and Loop1 interactions

We then varied the nature of the micro-homology base pair (MH-bp), keeping the same incoming ddCTP and templating base but using a DNA duplex that ends with an 3′-overhanging C, T or A (Table 1). In general, one observes very similar geometries in the different complexes.

In addition, we see a network of water molecules checking the minor groove of the MH-mini-helix (MH-mh), as shown in Supplementary Fig S1. It involves a water molecule (W1) bridging the two bases of the nascent-bp, another one (W2) checking the MH-bp and two pairs branching out of W2 (W3a and W4a or W3b and W4b). The stability of this network of water molecules will be investigated in more detail at the end of the Results section.

Because of the good resolution of the diffraction data, it was possible to interpret the electron density of the base in the MH-bp locus in terms of two conformations, either stacked or non-stacked (Fig 2A). In the three complexes with a non-Watson–Crick MH-bp, about 50% of the 3′-base of the template strand is stacked between the templating one and the main chain of Loop1. In the complex with a C-G at the MH-bp level, two stacked conformations are observed and, surprisingly, the water molecule bridging the two bases of the nascent base pair is not unambiguously seen, but this may be due to the relatively lower resolution of this particular diffraction data set.

Loop1 is well ordered in most of the complexes (Supplementary Fig S2). L398 is inserted in the primer strand as previously observed in complexes with the primer strand alone (Gouge *et al*, 2013) and the residues 395–397 of Loop1 ‘cap’ the 3′-end base of the *in trans* template strand, thus diverting the rest of this strand outside of the protein. In the C-C complex, Loop1 can be fully built in the electron density map. Next in the level of ordering of Loop1 comes the C-T complex, then C-G and C-A (not shown). Interestingly, Loop1 conformation is markedly different from the one observed when the DNA substrate is just a single-stranded primer (Fig 2A). This ordering of Loop1 contrasts with the situation in the Pol mu gap-filling complex, where it is completely disordered (Moon *et al*, 2007, 2014). In all cases, F405 interacts closely with L398 to form a wedge in the helical path of the primer strand (Figs 1C and 2C).

Although the role of F405 and L398 side chains has previously been recognized in the interaction of Tdt with a single-stranded primer (Gouge *et al*, 2013) and investigated by side-directed mutagenesis in our previous studies (Romain *et al*, 2009), mutants at

these positions were tested only with DNA substrates with an *in cis* template strand. Here, the activity tests were repeated in the presence of a primer strand alone or a DSB substrate with an *in trans* template strand (Fig 3B and D) and, indeed, we observed that the mutants' activity was very much reduced, in accordance with their role in forming this wedge in the primer strand that isolates the MH-mh from the rest of the primer strand.

Altogether, we conclude that Loop1 region (and especially the main-chain atoms of residues 395–398 and the side-chain atoms of L398 and F405) can be considered as the principal molecular determinant for constraining the very short length of MH search zone (exactly 1 bp) in Tdt, where base stacking interactions play a major role, rather than base–base hydrogen bonding.

Testing the stability in solution of the DNA synopsis complex seen in the crystal state

To test the existence of the complex made by Tdt, primer and downstream dsDNA in solution, we employed a simple test involving cellulose beads coated with dT₂₅ (Fig 4A). In brief, Tdt was incubated with the beads and a 5'-radiolabelled downstream DNA duplex added, with the amount of radioactivity after incubation being directly proportional to the amount of Tdt bound to the ssDNA and the dsDNA. The same test was also performed in the absence of Tdt to measure the background (non-specific binding). From this, it is apparent that the presence of Tdt induces duplex DNA binding, giving rise to a signal clearly above the noise level (Fig 4B). When the experiments were repeated in the presence of 1 mM Co²⁺ and ddNTP, we observed a stronger binding when the nucleotide is complementary to the templating base (Fig 4B).

We then analyzed the effect of the presence or absence of a perfect MH-bp and found little differences in the amount of bound DNA duplex (Fig 4B, lines GGp and AGp), while the presence of a 5'-phosphate group on the downstream DNA primer strand did not really matter. Additionally, we performed the test with Pol mu and observed already known features typical of Pol mu (Martin *et al*, 2012), that is, the binding of the downstream DNA duplex is stronger if its primer strand is 5'-phosphorylated (Fig 4B, lines GG and GGp).

Specificity of ddNTP incorporation by Tdt in the presence of a downstream duplex

To establish whether the presence of a downstream DNA duplex induces *in trans* templated elongation activity in Tdt, we performed elongation tests with ddNTP to detect small differences in the initial steps of the reaction (Supplementary Fig S3). Indeed, the regular elongation assays (i.e. distribution of lengths of products after a given amount of time) did not allow to detect any significant templated activity or a difference of activity compared to the single-strand substrate alone, in the presence of the downstream duplex (Fig 3A). Using different sets of oligonucleotides, we were therefore able to test the influence of a downstream template strand (compare ssDNA and DSB substrate), the importance of a MH-bp (compare MH: C-G and MH: C-A) and the effect of the nature of the last base [either a pyrimidine (C) or a purine (A)].

The template-base instructed character of ddNTP incorporation remains very low in the presence of the downstream duplex. Therefore, the biological function of Tdt is basically unaffected by

the presence of this downstream DNA duplex. In addition, there is an overall faster incorporation of dNTP if the last base of the primer strand is a purine instead of a pyrimidine, consistent with previous observations (Kato *et al*, 1967), but still no bias in dNTP incorporation.

In general, assuming that the functionally relevant conformation of the base in the MH-locus position is the one in a stacked conformation, these functional results are consistent with the structural results, where we see in all cases (cognate or non-cognate MH-bp) at least 50% of the base in a stacked conformation (Fig 2). At the molecular level, we checked that the different structural intermediates in the mechanism recently described for Tdt in the presence of a primer strand alone (Gouge *et al*, 2013) are compatible with the presence of the downstream DNA duplex (Supplementary Fig S3B).

Probing the effect of a disordered Loop1 with F401A Tdt mutant

We then crystallized a Tdt mutant (F401A) which we had previously identified as having an unusual Pol mu-like *in cis* templated activity (Romain *et al*, 2009). The most probable reason for this kind of activity in this Tdt mutant is that Loop1 is disordered and unable to assume its role in excluding the *in cis* template strand. We predicted that this destabilization would lead to an inactive mutant on the DSB substrates and/or a primer strand substrate because Loop1 is needed to grip the primer strand and, indeed, that is what we observed (Fig 3C). The structure of Tdt F401A in presence of substrates similar to those described previously (Template strand T₅GY, where Y = C, A, T, and T₅GGG, see Table 1, Fig 2) revealed that most of Loop1 and specifically the 396–398 region of Loop1 is disordered in all cases (Supplementary Fig S2) despite the good resolution of the diffraction data. In particular, L398 and F405A side chains are not visible in the electron density map and the side chains of K403 and D399 are not well defined. The 3'-end base of the primer is not well defined in the case of a C-G MH-bp and could be built as two non-canonical conformations in a C-C MH-bp; as a consequence, although the 3'-end base of the template strand can be built in the case of a cognate MH-bp, two alternative conformations can be built in the case of a non-cognate MH-bp. These structural data are fully consistent with the functional tests for this mutant (Fig 3C).

To further assess the adaptability of Loop1 to substrates with a longer 3'-end which could possibly form two MH base pairs instead of one, we solved the structure of the F401A mutant in complex with the same kind of annealed DSB but with a longer template strand (+1 base, sequence T₅G₃) in the presence of Zn²⁺. This complex is in the post-catalytic state (i.e. the 3'-end of the primer strand lies in the active site because translocation did not occur; Brissett *et al*, 2013), and the MH-bp is A-G (Fig 5). We only observe one conformation for the base of the template strand engaged in the MH-bp but no unique and clear density for the extra 3'-end base. As before, Loop1 is disordered and cannot be built in the electron density map; however, it is clear that the extra base cannot displace Loop1 to form a supplementary base pair with the primer strand (which would in this case also be a G-A base pair). Additionally, we see another binding site for Zn²⁺ (site C), coordinated by residues from the SD2 region, already described by Hogg *et al* (2014). We had expected that Zn²⁺, H475 and D473 would cooperate to bind the extra 3'-phosphate group in

Figure 3. Functional tests of three different Loop1/SDR1 mutants of Tdt.

A–D Three different substrates were used to test the nucleotidyltransferase activity and both the *in cis* and *in trans* templated polymerase activity of Tdt wild-type (A), L398A (B), F401A (C) and F405A (D) mutants. All tests were made in the presence of 1 mM MgCl₂.

Source data are available online for this figure.

Figure 4. Interaction of Tdt with a substrate mimicking a DNA DSB in solution.

A Principle of the method. The beads are coated with a single-stranded DNA (ssDNA) that acts as the upstream primer (dT₂₅); Tdt (with its surface represented in gray) and a 5'-radiolabeled downstream dsDNA are then added. After several washes, the radioactivity is measured and is proportional to the quantity of dsDNA bound to both the Tdt and the ssDNA. Different oligonucleotides have been used to assess the binding in solution of Tdt on the DSB substrate (depicted on the right).
B Quantity of complexes assembled on the beads for different DNA substrates. The same Tdt batch was used in all cases. An experiment with Pol mu is also presented. The error bars represent the standard deviation.

this new complex, but this was not observed. Indeed, one Mn²⁺ ion was observed in the same place in Pol lambda structure and playing this role (Garcia-Diaz *et al*, 2007), but with the gap-filling complex.

Testing the DNA synopsis model seen in Tdt–DSB complexes by F405A Tdt structures

To highlight the importance of the L398-F405 side chains, we collected diffraction data for the F405A Tdt mutant in complex with an *in trans* template strand containing either a Watson–Crick MH-bp (C–G) or a non-Watson–Crick one (C–C) (Table 1). When a Watson–Crick MH-bp is present, Loop1 is better ordered than with a non-cognate MH-bp, but it is not as well stacked over the 3'-end base of the *in trans* template strand as with wild-type Tdt; the L398 side chain is not visible in the electron density map, D399 is disordered, and K403 side chain is also missing (Supplementary Fig S2). When a

mismatch C–C is present, Loop1 is even more disordered and cannot be built in the density. Also, despite the relatively good resolution, the 3'-end base of the primer strand could not be seen well in the density. This is probably due to the absence of the F405 side chain, which prevents clamping the penultimate base of the primer strand (the side chain of L398 is also missing). These observations are consistent with the functional tests for this Tdt mutant, which showed a greatly reduced activity with the DSB substrate (Fig 3D).

Applicability of the wedge model seen in Tdt–DSB recognition to Pol mu

To test the applicability of the Tdt–DNA synopsis recognition mode in isolating the MH-mhin both Tdt and Pol mu, we investigated the role of L398 and F405 by site-directed mutagenesis in mouse Pol mu. We postulated that if these residues help to stabilize the MH-bp, then a mutation to alanine would impair the mutants' activity in the

Figure 5. Structure of the Tdt–DSB F401A_GA complex in presence of Zn²⁺.

Two metal ions were identified in the anomalous map: a catalytic one in the active site at 7.8 σ and an additional one (site C) underneath Loop1 at 5.8 σ , coordinated by D473 and H475 (SD2 region), both represented with a purple density contoured at 4 σ . A similar secondary site has been observed for Mn²⁺ in Pol lambda. The complex is in the post-catalytic state, as always observed in the presence of a divalent transition metal cation (Gouge et al, 2013). In the complex shown here, the MH-bp is A-G. In the presence of Mg²⁺, with no Zn²⁺, the pre-catalytic complex would be observed (see the F401A_CG structure with a C-G MH-bp). The extra 3'-end template base (G) points toward the solvent. The electronic density in the 2Fo-Fc map is contoured at 1 σ .

presence of a primer strand or a DSB substrate but not with a template strand *in cis*. Indeed, the mutant M384A in mouse Pol mu (equivalent to L398A in Tdt, Supplementary Fig S4A) has a normal DNA synthesis activity with an *in cis* template strand but is inactive with an *in trans* one (Supplementary Fig S4C). For the F391A mutant (F405A in Tdt), we observe a weak template-dependent activity in the presence of a regular duplex as well as a weak 3'–5'-exonuclease activity (Supplementary Fig S4D). However, in the case where the substrate is a single-stranded primer, we see a completely impaired primer extension activity which is reversed to a strong 3'–5'-exonuclease activity. This phenotype is also observed but somewhat attenuated in the presence of the downstream DNA duplex substrate (Supplementary Fig S4D).

We also studied the conservative mutation D385E in mouse Pol mu, in the same SD1 region (equivalent to D399 in Tdt). According to the present Tdt–DSB complex, its role would be to make a salt bridge with residue R389 (equivalent to K403 in mouse Tdt). In this case, we also found a strong 3'–5'-exonuclease phenotype (Supplementary Fig S4E).

Possible atomic mechanism for the random incorporation of nucleotides by Tdt

We investigated the role of water molecules in the stabilization of both the MH and nascent base pairs. Six water molecules were seen in the experimental electron density maps close to either the

nascent or MH base pairs (see Supplementary Fig S1). We used standard minimization techniques to build the hydrogen atoms, find their optimal configuration and to probe the importance of these water molecules. We focused on the structure in which the MH-bp is C-C, as it possesses the highest resolution and all of Loop1 is visible in the electron density map. One water molecule (W1) is consistently found at the level of the nascent base pair (Fig 6C). We found that, for this water molecule to stay in place, it was both necessary to assign a rare tautomeric state (imino form) to one of the cytosines involved in the MH-bp (Fig 6B) and to adjust the tautomeric state of H475. Indeed, if these requirements are not met, W1 drifts away due to other water molecules closer to the MH-bp. On the contrary, when the correct tautomeric state is set, the crystallographic water network is topologically conserved with three water molecules linking the MH-bp to residue R461 and the backbone of residue V397 from Loop 1 (Fig 6B) and two water molecules that point to ddCTP and residue D399 (Fig 6A). W1 stays in place and bridges the base of the

incoming nucleotide to the backbone of G449 (Fig 6C). The side chain of the strictly conserved R461 residue is of utmost importance in stabilizing this network. Indeed, mutating this residue to alanine (R461A) resulted in an inactive mutant indicating that it is as important as the catalytic aspartates (Supplementary Fig S5). G449 and D399 are also strictly conserved among Tdt and Pol μ sequences.

Interestingly, it was possible to replace the base of the incoming nucleotide, effectively changing it from ddCTP to ddGTP, while keeping the water network in place (Fig 6D). We checked that ddGTP could be stabilized here if (and only if) it was in its rare enol form. Thus, resorting to rare tautomers makes it possible to explain why any base can be incorporated with more or less the same efficiency, regardless of the chemical nature of the (instructing) templating base. More detailed studies will be necessary to establish this phenomenon on the quantum mechanical level (QM/MM) or using Density Functional Techniques (DFT).

Figure 6. Stability of the water molecules network as judged by energy minimization and selection of the best base tautomers.

- A View parallel to the helical axis of the micro-homology base pair (MH-bp) and nascent base pair with Loop 1 residues capping them. Two H-bonded water molecules are shown: W3a interacts with one base of the MH-bp and the sugar group of the incoming ddCTP, and W4a is hydrogen-bonded to residue D399 from Loop 1.
- B View down the helical axis of MH-bp highlighting both the rare tautomeric state of the cytosine on the duplex side (in its imino form) and a network of three water molecules that link it to residue R461. One water molecule, W2, is interacting with both bases of the MH-bp. The water molecule in the middle, W3b, is also interacting with the backbone of residue V397 from Loop 1 (not shown for clarity).
- C View down the helical axis of the nascent base pair highlighting the standard Watson–Crick pairing between the incoming ddCTP and the templating G nucleotide on the duplex side. Emphasis is put on the presence of a water molecule, W1, bridging the incoming ddCTP and the backbone of residue G449.
- D Same view as in (C), but where the ddCTP nucleotide has been replaced *in silico* by a ddGTP molecule and found to be more stable in the enol tautomeric form, allowing it to share H-bonds with the standard G on the downstream duplex side and the same water molecule W1. Here, W1 is also stabilized by the N474 side chain.

Discussion

The new crystal structures presented here show a tight association of Tdt with both 3'-end protruding DNA ends of a DNA synapsis sharing one MH base pair. This tight association was also confirmed to occur in solution by employing a simple binding test based on sepharose-dT₂₅ beads. The downstream duplex part of the DSB is essentially bound by the 8 kDa domain of Tdt.

Additionally, functional tests in solution indicate that the presence of a downstream DNA duplex only slightly slows down the kinetics of dNTP incorporation but does not change its lack of template specificity, thereby preserving its biological role in the generation of random N regions at the V(D)J junctions in immunoglobulins and T-cell receptors.

In that sense and contrary to what was previously believed, Tdt is not a 'misguided' polymerase (Motea & Berdis, 2010), but rather an example of a polymerase accepting an *in trans* template across a fragile bridge, checking the presence of an MH-bp but not its cognate nature. The fragility of the substrate might be related to the absence of major tertiary structural change of the enzyme throughout the catalytic cycle.

We can now explain the recognition of the *in trans* templating strand at the level of the 'MH-bp' locus in structural terms. First, the fundamental role of Loop1 is here described in atomic details for the recognition of the MH-bp region and its conformation is identified. In particular, we emphasize the importance of the recently characterized L398 and F405 (Gouge *et al*, 2013) and extend their role with a DNA DSB substrate. Second, the role of a well-ordered Loop1 in stabilizing the MH-bp is assessed by the F401A Tdt mutant, which is only active when the template strand is present *in cis* but not when it is present *in trans*. Third, the MH and nascent base pairs form a separate block, physically distinct from the rest of the primer strand. Loop1 is not acting as a 'pseudo-template' but stabilizes the MH-bp through several direct contacts and a highly structured water network.

Other studies have stressed the importance of this dinucleotide step to explain dNTP incorporation specificity by Tdt (Mora *et al*, 2010; Murugan *et al*, 2012), showing that the variability of the inserted sequences is fully accounted for by just the dinucleotide statistics of these two positions. Indeed, studies involving deep sequencing on T-cell receptors sequences and statistical derivation of the underlying law, using Markov Models, concluded that available sequence data of the N regions can be explained solely in terms of the dinucleotide step involving the 3' last and penultimate bases of the primer (Mora *et al*, 2010; Murugan *et al*, 2012). Our structural model provides a natural molecular explanation for these observations.

It is of interest to compare our results with the bacterial LigD (Brissett *et al*, 2007, 2011, 2013), which performs NHEJ in certain bacteria. In both cases, the NHEJ polymerase could be crystallized bound to an annealed DNA double-strand break without the rest of the NHEJ apparatus and the DNA synapsis is stabilized by surface loops of the polymerase (although Loop1 in Tdt is quite different in length and conformation from Loop1 of LigD). There are, however, major differences: in Tdt, the 3'-hydroxyl of the template strand is not positioned into the active-site pocket of an *in trans* second PolX molecule. Rather, the 3'-end of the template strand of the downstream DNA is taken care of by Loop1, SD1 and SD2 regions/motifs

of the same polymerase molecule. In addition, there is no templating base selection relying on Loop 1 in Tdt but rather an insertion of Loop1 in the primer strand, forming a wedge that isolates the MH-bp from the rest of the DNA substrate. This wedge is not seen in the LigD structure, and the reason for this may be that the MH-bp zone used in the latter structure contains four base pairs rather than one as in this study. Given the structures described here, it is not possible to imagine how a four base pairs MH region would be accommodated by Tdt and its characteristic wedge in the primer strand substrate, except by excluding the last three bases of the downstream template strand.

The results obtained with the mutants M384A and F391A strongly suggest that the DNA-DSB-Tdt model is also valid in Pol mu. While this article was being written, two studies were published where mutagenesis of the SD1 and SD2 motifs was used to probe Pol mu.

First, Martin and Blanco (2014) tested several substrates with different lengths of the MH region and, consistent with our results, it appears that the best substrate has exactly one MH-bp. Position F389 in human Pol mu (equivalent to F405 in mouse Tdt, and F391 in mouse Pol mu) was mutated as F389L instead of F391A in mouse Pol mu (Supplementary Fig S4), so this may explain why the exonuclease phenotype was not observed in the latter case. In addition, the Pol mu SD2 region was swapped with the Tdt SD2 sequence (459 NSH 461 => DNH) and helix N was mutated (R447A) and, again, the results are consistent with ours. We believe that the 'network' of interactions between the SD1 and SD2 regions suggested for Pol mu in Martin and Blanco (Martin & Blanco, 2014) could be very similar to that described in Fig 2B and Supplementary Fig S1, by doing just two mutations, namely N474S and K403R in Tdt. However, this remains to be tested by crystallizing Pol mu with the same kind of DSB substrate shown here.

Second, Moon *et al* (2014) studied in detail the mutants M382A of human Pol mu (M384 in mouse Pol mu) and found the same results as described here for mouse Pol mu (Supplementary Fig S4); they used a full NHEJ test (including ligation by Ligase IV) on a DNA substrate containing two MH base pairs, whereas we studied the end-joining of the Pol mu mutants alone on a DNA substrate that contains only one MH base pair, in order to compare functional and structural data. Our structures suggest that, in a DNA substrate with two potential MH base pairs, the last 3'-base of the downstream DNA is excluded from the protein binding site (Fig 5).

Still, given the high degree of structural similarity between Tdt and Pol mu (Fig 1D), there remains the intriguing problem as to why Pol mu is template dependent and Tdt is not, in comparable conditions where the templating base comes from an *in trans* template strand. Here, we show that bases at both the MH-bp and nascent-bp positions in Tdt are likely to form rare tautomers, as seen from energy minimizations aimed at preserving the water molecule network in the minor groove of the DNA helix in this region. It is well known that the use of base tautomers can preserve the volume of a Watson–Crick base pair even for non-cognate ones: indeed, there are ways to make non-cognate base pairs isosteric with cognate ones (Westhof, 2014). This mechanism would easily explain the incorporation of virtually any base by Tdt. This use of rare tautomers is also in line with a number of recent studies involving either a PolX (Pol lambda), a PolA (Pol I

from *Bacillus stearothermophilus*) or a PolB (from phage RB69). In Pol lambda, the structure of a complex with a DNA substrate containing a non-Watson–Crick (G–T base pair) nascent base pair was solved in the presence of Mn^{2+} (Bebenek *et al*, 2011) and very small structural differences were observed compared to a normal Watson–Crick base pair; the structure is consistent with the presence of a tautomeric form of the base pair. In the active site of a member of the PolA family (Wang *et al*, 2011), an almost perfect Watson–Crick geometry was observed for a C–A mismatch in the presence of Mn^{2+} : it involves a rare tautomer of one of the bases and the authors stress the role of the water molecules network to recognize the nascent base pair in the different known DNA polymerase families. In the PolB family, several new structures of the RB69 polymerase, also in the presence of Mn^{2+} , point to the importance of the water network and rare tautomers to stabilize non-cognate nascent base pairs (Xia & Konigsberg, 2014). Furthermore, the water network of RB69 polymerase in the minor groove of the DNA seems to be conserved in human Pol epsilon

(Hogg *et al*, 2014), also a PolB. We note that a common explanation for these observations would be a strong polarization effect of the divalent transition metal ion (Mn^{2+}) onto the network of water molecules to stabilize rare tautomers, or directly on the nucleobases. Strikingly, it has been known for years that Tdt activity is accelerated in the presence of Mn^{2+} or Co^{2+} or traces of Zn^{2+} (Kato *et al*, 1967) and Pol mu displays a nucleotidyltransferase phenotype in the presence of Mn^{2+} (Romain *et al*, 2009). Site C (Fig 5) should also be taken into account in future studies of transition metal ions effects.

It remains to be seen how the two important features reported here for Tdt, that is the water network and the use of rare tautomers, can explain the fidelity (or lack of) of Pol mu and, in particular, if this water network seen in Tdt in the minor groove of the DNA is conserved in Pol mu. The role of Mn^{2+} (or other divalent transition metal ions) also needs to be investigated in detail, especially their possible role in polarizing nearby water molecules. Based on molecular dynamics simulations, other authors have postulated the

Figure 7. Common features of PolX involved in V(D)J recombination or NHEJ DNA repair.

In Tdt (V(D)J), the MH-bp is loosely checked by base stacking and Loop1 interactions, whereas in Pol mu (NHEJ), the Watson–Crick complementarity of the MH-bp is specifically checked. This is likely due to the slightly different SD1 and SD2 regions/motifs in the two proteins (colored magenta and orange, respectively).

existence of an intermediate state (check point) in the reaction path for Pol mu (Li & Schlick, 2013, 2010). One possible scenario, which is compatible with a number of Pol mu mutants that have a 3′–5′-exonuclease phenotype (Rosario *et al*, unpublished), would be a sequential effect and active role of Loop1 in checking the MH-bp in Pol mu—that would not exist in Tdt. In any case, the differences between the mechanisms of Pol mu and Tdt are bound to be very subtle and will require further investigation.

The wedge mechanism we describe here for Tdt and Pol mu to bind a DSB in DNA is likely to be absent in Pol beta or Pol lambda as their Loop1 is smaller and residues L398, D399 or F405 are not conserved. However, we note that yeast Pol IV does possess a long Loop1 and canonical SD1 and SD2 sequences: the equivalent of SD2 sequence is TQH instead of NSH in Pol mu or DNH in Tdt and the equivalent of SD1 sequence is IKKFY instead of FERSF (Tdt) or FQKCF (Pol mu). We therefore predict that Pol IV should work as Pol mu and Tdt with respect to an interrupted template strand substrate and available experimental data seem to confirm this hypothesis (Daley *et al*, 2005; Daley & Wilson, 2008).

One may wonder why Tdt maintains a downstream duplex with a templating base *in trans*. Indeed, this does not seem to change its dNTP specificity—or lack thereof—and, one may therefore ask what the biological benefit of this would be. A simple explanation for maintaining this ability is that it would be obviously better/safer for the cell to keep the downstream DNA duplex of a DNA synopsis in close proximity to the upstream one, independently of the (un)templated character of the nucleotide addition. In this way, when the core complex made of Ku 70-80 and DNA-PKc relaxes its grip on the DNA synopsis to make way for Tdt, the *in trans* DNA would still be in close physical proximity of the 3′-end being processed. On the evolutionary level, we can hypothesize that Tdt evolved from a proto-Pol mu in a straightforward manner, simply by developing a looser grip and dropping the checkpoint on the MH base pair (Fig 7). This would have allowed several incorporations of dNTPs in a row, independently of the identity of the base at the MH-bp locus generated by the previous incorporation. In this way, all components of the V(D)J recombination apparatus used in the adaptive immune system might have evolved 480 million years ago from existing enzymes, first with Rag1 from the Transib transposase (Kapitonov & Jurka, 2005) and then borrowing the core NHEJ machinery and evolving Tdt from Pol mu.

Materials and Methods

Protein purification and mutant preparation

The mouse Pol mu sequence was inserted in a pET28 vector [as described in (Moon *et al*, 2007)], with a TEV cleavage site inserted between the HisTag and P136. The plasmid was then used to transform *E. coli* BL21-Gold(DE3)pLysS. Bacteria were grown with appropriate antibiotics to OD = 0.6. Pol mu expression was induced with 1 mM IPTG overnight at 16°C. The resuspension buffer contained 500 mM NaCl, 5 mM imidazole and 50 mM Tris pH 8.3. The lysate was loaded on a 5-ml nickel affinity column (HisTrapHP, GE Healthcare) and eluted with a gradient up to 500 mM imidazole. Fractions containing the protein were then pooled, concentrated and

subjected gel filtration on a Superdex 75 16/600 (GE Healthcare). The mouse Tdt clones, inserted in a pET28 vector, were expressed in BL21 Gold(DE3)pLysS after an overnight induction at 16 h with 1 mM IPTG. The purification was described in Romain *et al* (2009). Mutants were generated with the QuikChange mutagenesis kit (Agilent). Oligonucleotides were purchased from Eurogentec (Belgium) and dissolved in 10 mM Tris–HCl pH 8.0, 1 mM EDTA (TE) for the elongation assays.

Crystallization and diffraction data collection

The dsDNA (dA₅ and TTTTGTGX, where X = A, C, G or T) were annealed in a buffer containing 50 mM Tris pH 7.8, 5 mM MgCl₂ and 2 mM EDTA. Wild-type Tdt and F401A and F405A mutants were mixed at a final concentration of 10 mg/ml with 1.2 excess of the ssDNA (dA₅) and 1.2 excess of dsDNA in a buffer containing 50 mM MES pH 6.5, 50 mM magnesium acetate, 200 mM potassium chloride, 5 mM DTT and 10% glycerol. The complex was first incubated at 4°C for 1 h then incubated with ddCTP (2 mM) for 1 h. The crystals grew in 3 days in a solution containing 12–17% PEG 4000, 9–12% isopropanol, 100 mM sodium acetate and 100 mM HEPES pH 7.5. Crystals were flash-frozen in liquid nitrogen, with a mix of 50% paraffin and 50% paratone as cryoprotectant. The same oligonucleotides were used both for wild-type Tdt and F401A and F405A mutants, with just one exception: the annealed T₅G₃ and A₅ oligonucleotides were used for F401A, with or without Zn²⁺, in the presence of ddCTP and dA₅ (F401A_CG and F401A_AG).

Refinement and model validation

All the data were processed using XDS (Kabsch, 2010), reduced with POINTLESS (Evans, 2011), scaled and merged with SCALA (Evans, 2005). Data collection statistics are included in Table 1. 5% of the reflections were removed from the refinement and kept aside to calculate the R_{free}. Molecular replacement was performed with PHASER (McCoy *et al*, 2007) using 4I2A (Gouge *et al*, 2013) as a search model for all structures. Manual building was achieved with COOT (Emsley & Cowtan, 2004). BUSTER-TNT (Bricogne *et al*, 2011) was used to refine the model until convergence of the R-factors to a minimum. TLS groups were chosen with TLSMD (Painter & Merritt, 2006) and included in the last stages of refinement; three TLS groups were selected for the protein, one TLS group for each of the DNA strands. All the refined models were validated with MOLPROBITY (Chen *et al*, 2010). Superimpositions of structures and figures were generated with PyMol (DeLano).

Coordinates have been deposited under PDB codes 4QZ8 through 4QZI.

Polymerase activity test

DNA template preparation

Oligonucleotides were purchased from Eurogentec, Belgium, and dissolved in Tris–HCl 50 mM pH 8, 1 mM EDTA. Concentrations were estimated by UV absorbance using an absorption coefficient ϵ at 260 nm provided by Eurogentec. Primer strand was 5′-labeled with γ -³²P-ATP (Perkin Elmer, 3,000 Ci/mM) using T4 polynucleotide kinase (New England Biolabs) for 1 h at 37°C; the labeling

reaction was stopped by heating the kinase at 75°C for 10 min. A label-free duplex was prepared by annealing two complementary oligonucleotides. The primers were mixed, heated for 5 min up to 90°C and slowly cooled to room temperature overnight.

DNA substrates

Primers:

5' TACGATTAGCCTC and 5' TACGATTAGCCTA

Downstream Duplexes:

(GG)

3' GGCCGATTACGCAT 5'
pGGCTAATGCGTA 3'

(CG)

3' CGCCGATTACGCAT 5'
pGGCTAATGCGTA 3'

(AG)

3' AGCCGATTACGCAT 5'
pGGCTAATGCGTA 3'

(CT)

3' CTCCGATTACGCAT 5'
pGGCTAATGCGTA 3'

Activity test

The protein was diluted to the desired concentration using a 1× reaction buffer: for Pol μ , this buffer contained Tris–HCl 50 mM pH 7.1, 1 mM TCEP, 2 mM MgCl₂ and 0.1 mg/ml BSA; for Tdt, it contained 25 mM Tris–HCl pH 6.6, 0.2 M Na cacodylate, 4 mM MgCl₂ and 0.25 mg/ml BSA. 2 μ M polymerase was mixed with 0.05 μ M label-free duplex and reaction buffer and incubated for 10 min on ice. 0.05 μ M labeled primer was added and incubated 5 more minutes on ice and 10 min at room temperature. The reaction was started by addition of 0.25 μ M ddNTP and stopped within 1, 2, 4, 8 or 16 min by adding 10 mM EDTA and 95% formamide.

The products of the reaction were analyzed by gel electrophoresis on a 15% acrylamide gel containing 8 M urea. The 0.4-mm-wide gel was run for 3–4 h at 40 V/cm and scanned by phosphorimager Storm 860 Molecular Dynamics (GE Healthcare).

Affinity test

All the products were diluted in reaction buffer 1× containing 50 mM potassium acetate, 20 mM Tris–acetate pH 7.9 and 10 mM MgCl₂. This buffer was also used for washing the beads. The binding reaction was performed by mixing Oligo(dT)₂₅ cellulose beads to 2 μ M polymerase during 10 min with gentle agitation. The excess of polymerase was removed by sedimentation (20 s micro-centrifuge). Beads were washed three times with the reaction buffer. 0.01 μ M labeled duplex was applied to the polymerase-bound beads for 10 min with gentle agitation. Beads were washed three times, and the extra buffer was removed by sedimentation using a micro-centrifuge (20 s). The complex was re-suspended in reaction buffer and transferred into a tube that contains scintillation liquid. Radioactivity was measured with a Tri-Carb 2800TR Liquid Scintillation Analyser (Perkin Elmer). All measurements were made at least six times, and this was used to estimate standard deviations. We checked that the amount of bound radiolabeled duplex is linearly proportional to the amount of added polymerase.

Sequence analysis

A multialignment of Pol μ and Tdt sequences (starting at position 311 of Tdt) was obtained using MULTALIN (Corpet, 1988) and the following list of sequences, made of two subgroups:

Pol μ (15 sequences): human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), marmoset (*Callithrix jacchus*), naked mole rat (*Heterocephalus glaber*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), Chinese hamster (*Cricetulus griseus*), Brandt's bat (*Myotis brandtii*), camel (*Camelus ferus*), dolphin (*Monodelphis domestica*), newt (*Salamandra*), channel catfish (*Ictalurus punctatus*), Mexican tetra (*Astyanax mexicanus*), zebrafish (*Danio rerio*), cobra (*Ophiophagus hannah*).

Tdt (23 sequences): *Raja eglanteria* (clearnose skate), *Bos mutus* (ox), *Ovis aries* (sheep), *Canis familiaris* (dog), *Sus scrofa* (pig), *Macaca fascicularis* (macaque), *Pongo abelii* (orangutan), *Pan troglodytes* (chimpanzee), *Gorilla gorilla* (gorilla), *Heterocephalus glaber* (rat), *Cavia porcellus* (guinea pig), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Monodelphis* (dolphin), *Sarcophilus harrisii* (Tasmanian devil), *Myotis brandtii* (bat), *Myotis davidii* (bat), *Gallus gallus* (chicken), *Xenopus laevis* (frog), *Oncorhynchus mykiss* (rainbow trout), *Takifugu rubripes* (Japanese pufferfish), *Astyanax mexicanus* (Mexican tetra), *Danio rerio* (zebrafish).

A score representing the relative information in the first subgroup versus the second subgroup was defined, for each position (i), by calculating $S_{\text{cross}}(i) = \sum_{\alpha=1}^{20} p_{\alpha}(i) \log p_{\alpha}(i)/q_{\alpha}(i)$ where $p_{\alpha}(i)$ is the fraction of amino acid α at position (i) of the multi-alignment in the first sub-group (Pol μ) and $q_{\alpha}(i)$ the fraction of amino acid α at the same position (i) in the second sub-group (Tdt). A pseudo-count of 0.1 was added to take care of the case arising when one amino acid type is not represented at position (i). The $S_{\text{cross}}(i)$ score was then normalized by dividing it by the total entropy at this position, $S_{\text{tot}}(i)$, calculated by taking into account all sequences (no sub-groups).

Energy minimization and tautomer modeling

The structure of Tdt complexed with a C–C base pair at the MH-bp locus was inserted in a void cubic box of dimension 120 Å × 120 Å × 120 Å. Ions and water molecules present in the PDB file were kept. The CHARMM36 force field (Best et al, 2012) was used. All simulation runs were performed using NAMD (Phillips et al, 2005). The package PSFGEN was used within VMD (Humphrey et al, 1996) to build missing atoms and create input files for NAMD (Phillips et al, 2005). The tautomeric state of His475 was set to type HSD instead of type HSE (using 'mutate' command) after visual inspection, and the tautomeric state of the cytosine involved in MH-bp on the duplex side was set to its imino form by using the patch CYT1. The tautomeric state of the ddGTP incoming nucleotide was set to its enol form by using the patch GUT1. Patch DEOX was applied to all nucleotides. Patches 3PHO and 5TER were used at the strand termini. Topology information and parameters of types CYT (resp. GUA) and ATP were patched manually to define those of ddCTP (resp. ddGTP). TIP3P water model was used.

Conjugate gradient energy minimization runs were performed with options 'fixedAtoms' and 'rigidBonds' for 10,000 steps. Non-bonded interactions, defined through the '1-4' exclusion policy, were cut off at 12 Å with a switching function starting at 10 Å.

Atoms created by PSFGEN were unrestrained in an initial run, before unrestraining all hydrogen atoms and water molecules in a

second run. When a ddGTP molecule was built instead of the ddCTP found in the structure, atoms belonging to its base were also unrestrained.

Supplementary information for this article is available online: <http://emboj.embopress.org>

Acknowledgements

We thank Denis Ptchelkine for making both the F401A and NSH→ASA mutants of mouse Pol μ and Francois Rougeon (IP) for constant support. We acknowledge the help of the staff of ESRF (Grenoble) for excellent data collection facilities, Synchrotron Soleil (Orsay) for help in data collection and ARC Funding Agency (France) for financial support (Grant #3155).

Author contributions

JG grew crystals of complexes, solved, refined and analyzed all crystal structures. Mutants were constructed, expressed, purified by SR, FR and PB. All activity tests were performed by SR. FP performed energy minimizations and search for the best tautomers in the active site. MD devised research, co-wrote the manuscript and co-analyzed the structures with JG. All authors revised the final manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Aoufouchi S, Flatter E, Dahan A, Faili A, Bertocci B, Storck S, Delbos F, Cocco L, Gupta N, Weill JC, Reynaud CA (2000) Two novel human and mouse DNA polymerases of the PolX family. *Nucleic Acids Res* 28: 3684–3693
- Bebenek K, Pedersen LC, Kunkel TA (2011) Replication infidelity via a mismatch with Watson-Crick geometry. *Proc Natl Acad Sci USA* 108: 1862–1867
- Bebenek K, Pedersen LC, Kunkel TA (2014) Structure-function studies of DNA polymerase λ . *Biochemistry* 53: 2781–2792
- Benedict CL, Gilfillan S, Thai TH, Kearney JF (2000) Terminal deoxynucleotidyl transferase and repertoire development. *Immunol Rev* 175: 150–157
- Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, Mackerell AD (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J Chem Theory Comput* 8: 3257–3273
- Bollum FJ (1978) Terminal deoxynucleotidyl transferase: biological studies. *Adv Enzymol Relat Areas Mol Biol* 47: 347–374
- Bricogne G, Blanc E, Brandl M, Flensburg C, Keller P, Paciorek W, Roversi P, Sharff A, Smart O, Vornhein C, Womack T (2011) BUSTER version 2.11.2. Cambridge, United Kingdom: Global Phasing Ltd.
- Brissett NC, Pitcher RS, Juarez R, Picher AJ, Green AJ, Dafforn TR, Fox GC, Blanco L, Doherty AJ (2007) Structure of a NHEJ polymerase-mediated DNA synaptic complex. *Science* 318: 456–459
- Brissett NC, Martin MJ, Pitcher RS, Bianchi J, Juarez R, Green AJ, Fox GC, Blanco L, Doherty AJ (2011) Structure of a preternary complex involving a prokaryotic NHEJ DNA polymerase. *Mol Cell* 41: 221–231
- Brissett NC, Martin MJ, Bartlett EJ, Bianchi J, Blanco L, Doherty AJ (2013) Molecular basis for DNA double-strand break annealing and primer extension by an NHEJ DNA polymerase. *Cell Rep* 5: 1108–1120
- Chayot R, Danckaert A, Montagne B, Ricchetti M (2010) Lack of DNA polymerase μ affects the kinetics of DNA double-strand break repair and impacts on cellular senescence. *DNA Repair (Amst)* 9: 1187–1199
- Chayot R, Montagne B, Ricchetti M (2012) DNA polymerase μ is a global player in the repair of non-homologous end-joining substrates. *DNA Repair (Amst)* 11: 22–34
- Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66: 12–21
- Corpet F (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* 16: 10881–10890
- Daley JM, Laan RLV, Suresh A, Wilson TE (2005) DNA joint dependence of Pol X family polymerase action in nonhomologous end joining. *J Biol Chem* 280: 29030–29037
- Daley JM, Wilson TE (2008) Evidence that base stacking potential in annealed 3' overhangs determines polymerase utilization in yeast nonhomologous end joining. *DNA Repair (Amst)* 7: 67–76
- DeLano WL *The PyMOL Molecular Graphics System*. Available at www.pymol.org
- Delarue M, Boulé JB, Lescar J, Expert-Bezançon N, Jourdan N, Sukumar N, Rougeon F, Papanicolaou C (2002) Crystal structures of a template-independent DNA polymerase: murine terminal deoxynucleotidyltransferase. *EMBO J* 21: 427–439
- Domínguez O, Ruiz JF, Laín de Lera T, García-Díaz M, González MA, Kirchhoff T, Martínez-A C, Bernad A, Blanco L (2000) DNA polymerase μ (Pol μ), homologous to TdT, could act as a DNA mutator in eukaryotic cells. *EMBO J* 19: 1731–1742
- Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60: 2126–2132
- Esteban V, Martin MJ, Blanco L (2013) The BRCT domain and the specific loop 1 of human Pol μ are targets of Cdk2/cyclin A phosphorylation. *DNA Repair (Amst)* 12: 824–834
- Evans P (2005) Scaling and assessment of data quality. *Acta Crystallogr D Biol Crystallogr* 62: 72–82
- Evans PR (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr D Biol Crystallogr* 67: 282–292
- García-Díaz M, Bebenek K, Krahn JM, Blanco L, Kunkel TA, Pedersen LC (2004) A structural solution for the DNA polymerase lambda-dependent repair of DNA gaps with minimal homology. *Mol Cell* 13: 561–572
- García-Díaz M, Bebenek K, Krahn JM, Pedersen LC, Kunkel TA (2006) Structural analysis of strand misalignment during DNA synthesis by a human DNA polymerase. *Cell* 124: 331–342
- García-Díaz M, Bebenek K, Krahn JM, Pedersen LC, Kunkel TA (2007) Role of the catalytic metal during polymerization by DNA polymerase lambda. *DNA Repair (Amst)* 6: 1333–1340
- García-Díaz M, Bebenek K, Kunkel TA, Blanco L (2001) Identification of an intrinsic 5'-deoxyribose-5-phosphate lyase activity in human DNA polymerase lambda: a possible role in base excision repair. *J Biol Chem* 276: 34659–34663
- Gouge J, Rosario S, Romain F, Beguin P, Delarue M (2013) Structures of intermediates along the catalytic cycle of terminal deoxynucleotidyltransferase: dynamical aspects of the two-metal ion mechanism. *J Mol Biol* 425: 4334–4352
- Hogg M, Osterman P, Bylund GO, Ganai RA, Lundström E-B, Sauer-Eriksson AE, Johansson E (2014) Structural basis for processive DNA synthesis by yeast DNA polymerase ϵ . *Nat Struct Mol Biol* 21: 49–55

- Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* 14: 33–38, 27–28
- Juárez R, Ruiz JF, Nick McElhinny SA, Ramsden D, Blanco L (2006) A specific loop in human DNA polymerase μ allows switching between creative and DNA-instructed synthesis. *Nucleic Acids Res* 34: 4572–4582
- Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* 66: 125–132
- Kapitonov VV, Jurka J (2005) RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol* 3: 4540–4545
- Kato KI, Gonçalves JM, Houts GE, Bollum FJ (1967) Deoxynucleotide-polymerizing enzymes of calf thymus gland. II. Properties of the terminal deoxynucleotidyltransferase. *J Biol Chem* 242: 2780–2789
- Landau NR, Schatz DG, Rosa M, Baltimore D (1987) Increased frequency of N-region insertion in a murine pre-B-cell line infected with a terminal deoxynucleotidyl transferase retroviral expression vector. *Mol Cell Biol* 7: 3237–3243
- Li Y, Schlick T (2010) Modeling DNA polymerase μ motions: subtle transitions before chemistry. *Biophys J* 99: 3463–3472
- Li Y, Schlick T (2013) “Gate-keeper” residues and active-site rearrangements in DNA polymerase μ help discriminate non-cognate nucleotides. *PLoS Comput Biol* 9: e1003074
- Lieber MR (2008) The mechanism of human nonhomologous DNA end joining. *J Biol Chem* 283: 1–5
- Lieber MR, Lu H, Gu J, Schwarz K (2008) Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. *Cell Res* 18: 125–133
- Mahajan KN, Gangi-Peterson L, Sorscher DH, Wang J, Gathy KN, Mahajan NP, Reeves WH, Mitchell BS (1999) Association of terminal deoxynucleotidyl transferase with Ku. *Proc Natl Acad Sci USA* 96: 13926–13931
- Malu S, De Ioannes P, Kozlov M, Greene M, Francis D, Hanna M, Pena J, Escalante CR, Kurosawa A, Erdjument-Bromage H, Tempst P, Adachi N, Vezzoni P, Villa A, Aggarwal AK, Cortes P (2012) Artemis C-terminal region facilitates V(D)J recombination through its interactions with DNA Ligase IV and DNA-PKcs. *J Exp Med* 209: 955–963
- Martin MJ, Juarez R, Blanco L (2012) DNA-binding determinants promoting NHEJ by human Pol μ . *Nucleic Acids Res* 40: 11389–11403
- Martin MJ, Garcia-Ortiz MV, Gomez-Bedoya A, Esteban V, Guerra S, Blanco L (2013) A specific N-terminal extension of the 8 kDa domain is required for DNA end-bridging by human Pol μ and Pol λ . *Nucleic Acids Res* 41: 9105–9116
- Martin MJ, Blanco L (2014) Decision-making during NHEJ: a network of interactions in human Pol μ implicated in substrate recognition and end-bridging. *Nucleic Acids Res* 42: 7923–7934
- McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* 40: 658–674
- Mickelsen S, Snyder C, Trujillo K, Bogue M, Roth DB, Meek K (1999) Modulation of terminal deoxynucleotidyltransferase activity by the DNA-dependent protein kinase. *J Immunol* 163: 834–843
- Moon AF, Garcia-Diaz M, Batra VK, Beard WA, Bebenek K, Kunkel TA, Wilson SH, Pedersen LC (2007) The X family portrait: structural insights into biological functions of X family polymerases. *DNA Repair (Amst)* 6: 1709–1725
- Moon AF, Garcia-Diaz M, Bebenek K, Davis BJ, Zhong X, Ramsden DA, Kunkel TA, Pedersen LC (2007) Structural insight into the substrate specificity of DNA Polymerase μ . *Nat Struct Mol Biol* 14: 45–53
- Moon AF, Pryor JM, Ramsden DA, Kunkel TA, Bebenek K, Pedersen LC (2014) Sustained active site rigidity during synthesis by human DNA polymerase μ . *Nat Struct Mol Biol* 21: 253–260
- Mora T, Walczak AM, Bialek W, Callan CG Jr (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107: 5405–5410
- Motea EA, Berdis AJ (2010) Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochim Biophys Acta* 1804: 1151–1166
- Murugan A, Mora T, Walczak AM, Callan CG Jr (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA* 109: 16161–16166
- Nick McElhinny SA, Havener JM, Garcia-Diaz M, Juárez R, Bebenek K, Kee BL, Blanco L, Kunkel TA, Ramsden DA (2005) A gradient of template dependence defines distinct biological roles for family X polymerases in nonhomologous end joining. *Mol Cell* 19: 357–366
- Painter J, Merritt EA (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr D Biol Crystallogr* 62: 439–450
- Peterson RC, Cheung LC, Mattaliano RJ, Chang LM, Bollum FJ (1984) Molecular cloning of human terminal deoxynucleotidyltransferase. *Proc Natl Acad Sci USA* 81: 4363–4367
- Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L, Schulten K (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26: 1781–1802
- Romain F, Barbosa I, Gouge J, Rougeon F, Delarue M (2009) Conferring a template-dependent polymerase activity to terminal deoxynucleotidyltransferase by mutations in the Loop1 region. *Nucleic Acids Res* 37: 4642–4656
- Sawaya MR, Prasad R, Wilson SH, Kraut J, Pelletier H (1997) Crystal structures of human DNA polymerase β complexed with gapped and nicked DNA: evidence for an induced fit mechanism. *Biochemistry* 36: 11205–11215
- Wang W, Hellinga HW, Beese LS (2011) Structural evidence for the rare tautomer hypothesis of spontaneous mutagenesis. *Proc Natl Acad Sci USA* 108: 17644–17648
- Waters CA, Strande NT, Wyatt DW, Pryor JM, Ramsden DA (2014) Nonhomologous end joining: a good solution for bad ends. *DNA Repair (Amst)* 17: 39–51
- Westhof E (2014) Isostericity and tautomerism of base pairs in nucleic acids. *FEBS Lett* 588: 2464–2469
- Xia S, Konigsberg WH (2014) Mispairs with Watson-Crick base-pair geometry observed in ternary complexes of an RB69 DNA polymerase variant. *Protein Sci* 23: 508–513