doi:10.1006/jmbi.2001.4937 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) **312**, 335–346

JMB

Error Rate and Specificity of Human and Murine DNA Polymerase η

Toshiro Matsuda[1], Katarzyna Bebenek[1], Chikahide Masutani[3] Igor B. Rogozin[5,6], Fumio Hanaoka[3,4] and Thomas A. Kunkel[1,2*]

[1] *Laboratory of Molecular Genetics and*

[2] *Laboratory of Structural Biology, National Institute of Environmental Health Sciences Research Triangle Park NC 27709, USA*

[3] *Institute for Molecular and Cellular Biology, Osaka University and CREST Japan Science and Technology Corporation, 1-3 Yamada-oka Suita, Osaka 565-0871, Japan*

[4] *Institute of Physical and Chemical Research (RIKEN) Wako-shi, Saitama 351-0198, Japan*

[5] *Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences Novosibirsk, Russia*

[6] *National Center for Biotechnology Information NLM, National Institutes of Health, Bethesda MD 20894, USA*

**Corresponding author*

We describe here the error specificity of mammalian DNA polymerase η (pol η), an enzyme that performs translesion DNA synthesis and may participate in somatic hypermutation of immunoglobulin genes. Both mouse and human pol η lack intrinsic proofreading exonuclease activity and both copy undamaged DNA inaccurately. Analysis of more than 1500 single-base substitutions by human pol η indicates that error rates for all 12 mismatches are high and variable depending on the composition and symmetry of the mismatch and its location. pol η also generates tandem base substitutions at an unprecedented rate, and kinetic analysis indicates that it extends a tandem double mismatch about as efficiently as other replicative enzymes extend single-base mismatches. This ability to use an aberrant primer terminus and the high rate of single and double-base substitutions support the idea that pol η may forego strict shape complementarity in order to facilitate highly efficient lesion bypass. Relaxed discrimination is further indicated by pol η infidelity for a wide variety of nucleotide deletion and addition errors. The nature and location of these errors suggest that some may be initiated by strand slippage, while others result from additional mechanisms.

Keywords: DNA polymerase η; replication fidelity; frameshifts; tandem mutations; mismatch extension

Introduction

Human DNA polymerase η (pol η) is encoded by the *POLH* or *XPV* gene.[1,2] Mutations in this gene that inactivate pol η enhance UV-induced mutagenesis and greatly increase susceptibility to sunlight-induced skin cancer. These phenotypes, resulting from the absence of pol η, likely relate to its ability to copy DNA templates containing *cis-syn* thymine-thymine dimers more efficiently than most other DNA polymerases.[3,4] Structure-function studies of several of the other polymerases suggest that efficient and accurate DNA synthesis depends on correct Watson-Crick base-pair geometry.[5] Consistent with the importance of correct base-pair geometry to accurate replication, most DNA polymerases rarely make mistakes when copying undamaged DNA templates; their base-substitution error rates typically range from 10^{-6} to 10^{-4}. In contrast, human pol η is less accurate,[6] with an average base-substitution error rate of 3.5×10^{-2}. Consistent with this high error rate are steady-state

Abbreviations used: pol η, DNA polymerase η; pol ι, DNA polymerase i; pol β, DNA polymerase β; pol α, DNA polymerase α; HIV-1 RT, HIV-1 reverse transcriptase.

E-mail address of the corresponding author: kunkel@neiehs.nih.gov

0022-2836/01/020335–12 $35.00/0

kinetic data indicating that human[6,7] and yeast[8] pol η have low selectivity during the initial nucleotide insertion step of a polymerization reaction. These results imply that pol η has a relaxed requirement for correct base-pairing geometry and suggest that the function of pol η may need to be tightly controlled *in vivo* to prevent potentially mutagenic DNA synthesis.

To fully appreciate this mutagenic potential and to better understand the nucleotide selectivity of a DNA polymerase capable of efficiently bypassing a bulky lesion in DNA that strongly impedes synthesis by other polymerases, we provide a comprehensive view of the error specificity of pol η from two different sources. We show that mouse pol η is as inaccurate as its human homolog, thus demonstrating that low-fidelity synthesis during copying of undamaged DNA is inherent to mammalian pol η. We then describe the error specificity of human pol η in detail. The results reveal that pol η has low selectivity at the initial nucleotide insertion step, and it is promiscuous for subsequent mismatch extension. Both properties may relate to its ability to copy cyclobutane pyrimidine dimers efficiently. pol η is also highly inaccurate for a variety of nucleotide deletion and addition errors, with the specificity suggesting that these errors are made by more than one mechanism.

Results

Fidelity of mouse pol η

We began this study by examining the fidelity of mouse pol η using a forward mutation assay that scores a variety of substitution, addition and deletion errors during synthesis to copy a 407-nucleotide template present as a single-stranded gap in M13mp2 DNA.[6] Correct polymerization produces DNA that yields blue M13 plaques upon introduction of the product DNA into an *Escherichia coli* α-complementation strain and plating on indicator plates. Errors are scored as light blue or colorless plaques. Like its human homolog,[6] recombinant mouse pol η lacks 3′ → 5′ exonucleolytic activity and it completely filled the 407 nucleotide gap (data not shown, but see Figure 1(a) and (b) of Matsuda *et al*.[6]). When the products of gap-filling synthesis were introduced into *E. coli*, the *lacZ* mutant frequency among the resulting M13 plaques was 37 % (average of two independent determinations). This is much greater than the mutant frequency of DNA that had not been copied *in vitro* (0.06 %), indicating that the *lacZ* mutants were generated during gap-filling synthesis. The results with mouse pol η were similar to the 35 % *lacZ* mutant frequency obtained for synthesis catalyzed by human pol η.[6] As with the human enzyme,[6] the majority of *lacZ* mutant plaques generated by mouse pol η were colorless plaque.

To determine the nature and number of errors generated by mouse pol η, we isolated DNA from 20 independent *lacZ* mutants and sequenced nucleotides +191 (the start of synthesis) through −84 (the 5′ end of the *lacZ* gene in M13mp2). For comparison, and to obtain a detailed view of the mutagenic potential of human pol η, we expanded our earlier analysis of the error specificity of human pol η[6] by sequencing all 407 nucleotides in the gap (+191 through −216) in 124 independent mutants. The *lacZ* mutants generated by both enzymes contained multiple sequence changes, consistent with the colorless plaque phenotype of most of the mutants. This and the high *lacZ* mutant frequencies indicate that both mammalian polymerases conduct low-fidelity synthesis when copying an undamaged DNA template. A variety of sequence changes were observed (summarized in Table 1).

Base-substitution error rates and specificity

The majority of pol η errors are base substitutions (Figure 1). Those not located immediately adjacent to a flanking sequence change of any kind are categorized as single-base substitutions (Table 1, Figure 1(a)). The human and mouse enzymes generated 1560 and 122 single-base substitutions, respectively (Table 1). Given the total number of template nucleotides analyzed, the average single-base substitution error rates of human and mouse pol η are 3.2×10^{-2} and 2.2×10^{-2}, respectively†.

Substitutions resulting from formation of all 12 mispairs were generated by both polymerases. The two enzymes have comparable error rates for these mispairs (Figure 2(a)), and statistical analysis (see the legend to Figure 2) indicates that their average error specificities are not sig-

† Control experiments indicate that a single nucleotide misincorporated into the minus strand of M13mp2 DNA during gap-filling synthesis is expressed in *E. coli* with an efficiency of about 60 %. However, minus strand expression efficiencies are somewhat lower for DNA heteroduplexes containing larger numbers of mismatched bases (data not shown, but as an example, see Umar *et al*.[32]) Since *lacZ* mutants generated by pol η contain many sequence changes, the observed mutant frequencies of 35 % (human) and 37 % (mouse) suggest that essentially every gap-filled product of the pol η reactions originally contained mutations, and that the non-mutant (i.e. blue) plaques resulted from loss of the minus strand in *E. coli*. Consistent with this interpretation, no mutation was found when we sequenced DNA samples isolated from six independent blue M13mp2 plaques recovered from reactions by human pol η. Because most of the *lacZ* mutants contain many changes, silent changes are linked to changes that yield a detectable phenotype. Thus, unlike earlier studies with higher-fidelity DNA polymerases, in this study essentially all copied nucleotides within the gap become targets for the calculation of substitution, deletion and addition error rates. The error rate is simply the number of observed mutations of any particular type (from Figures 1 and 4) divided by the total number of nucleotides that were sequenced in all mutants (from Table 1).

Table 1. Summary of sequence analysis of *lacZ* mutants generated by human and mouse DNA polymerase η

	Human	Mouse
Total *lacZ* mutants sequenced	124	20
Bases sequenced per mutant	407	275
Total bases sequenced[a]	50,468	5500
Single-base substitutions	1560	122
Tandem base substitutions	89	6
Substitution + deletion or addition	34	0
Deletions: one base	119	12
two or more bases	87	7
Addition of one to three bases	74	9
Complex deletion/addition	9	0

[a] The number of nucleotides actually sequenced is slightly less due to deletions partly counterbalanced by additions. For simplicity, and because the net number of nucleotides lost is small, we calculated error rates using the total number of nucleotides listed.

nificantly different. Error rates for the 12 mispairs vary over a considerable range for both human and mouse pol η. For example, the human pol η error rate for the T·dGMP mispair is 70-fold higher than for the C·dCMP mispair (Figure 2(a), 6.3×10^{-2} *versus* 9.3×10^{-4}). The substitution spectrum for human pol η (Figure 1(a)) also reveals substantial sequence-context effects on error rates. For example, the error rate for the T·dGMP mispair varies by 18-fold, from 14×10^{-2} at the template T at position −8 to 0.8×10^{-2} at the immediately adjacent template T at position −7 or at positions +6 and +87 (Figure 1(a)). Sequence-dependent variations are observed for other substitutions (Figure 1(a)), yielding a range of error rates for each mispair (Figure 2(b)). The recovery of three base substitutions just upstream of 5′ end of the gap (Figure 1(a)) suggests that human pol η can perform limited strand-displacement synthesis.

To further examine sequence-dependent biases in error rate, we analyzed the largest collection of single-base substitution events, the T → C substitutions shown in Figure 1(a). Considering only the bases immediately flanking these substitutions, we detected two- to threefold biases for misincorporation of dGMP opposite T when its 3′ neighbor is a T·A or A·T base-pair as compared to when it is a G·C or C·G base-pair (Figure 3). Given the large number of template T residues (99) and substitution events (665) considered, these biases are statistically significant (see the legend to Figure 3).

Deletion and addition errors

Both human and mouse pol η also generated a variety of deletion and addition errors (Table 1, Figure 4). Among these, the most frequent changes were single-base deletions (Table 1, 119 occurrences with human pol η) and single-base additions (64 occurrences with human pol η). The average error rates of the human enzyme for single-base deletions and additions are therefore 240×10^{-5} and 130×10^{-5}, respectively. These rates are much higher than those of DNA polymerases in other families when measured using the same forward mutation assay (Table 2). The single-base additions and deletions generated by pol η are distributed throughout the template sequence (Figure 4(a)). From this distribution and the nucleotide composition of the template sequence, error rates for deletions and additions of specific nucleotides can be calculated. Thus, human pol η has an average error rate of 150×10^{-5} for deletion of one of the 204 non-iterated template nucleotides in the 407-nucleotide gap (Table 3). Interestingly, the average rates for deletions in homopolymeric runs of two to five nucleotides are only two- to threefold higher (albeit higher still at specific homopolymeric sites, Figure 4(a)). Also, the average single-base addition rates of human pol η do not increase with increasing run length (Table 3). Moreover, 18 additions involved adding a nucleotide that was different from either of its neighbors. These error specificity data suggest that many pol η deletion and addition errors may not be initiated by strand slippage (see below). In contrast to these

Table 2. Single-base deletion and addition error rates of human DNA pol η compared to other DNA polymerases

		Error rate ($\times10^{-5}$)	
DNA Polymerase	Family	Deletions	Additions
Human pol η (this study)	RAD30	240	130
Human pol β[29]	Pol X	13	0.2
Human pol α[36]	Pol B	3.9	0.3
HIV-1 RT[37,38]	RT	3.8	3.2
Klenow fragment[39]	Pol A	0.8	0.05

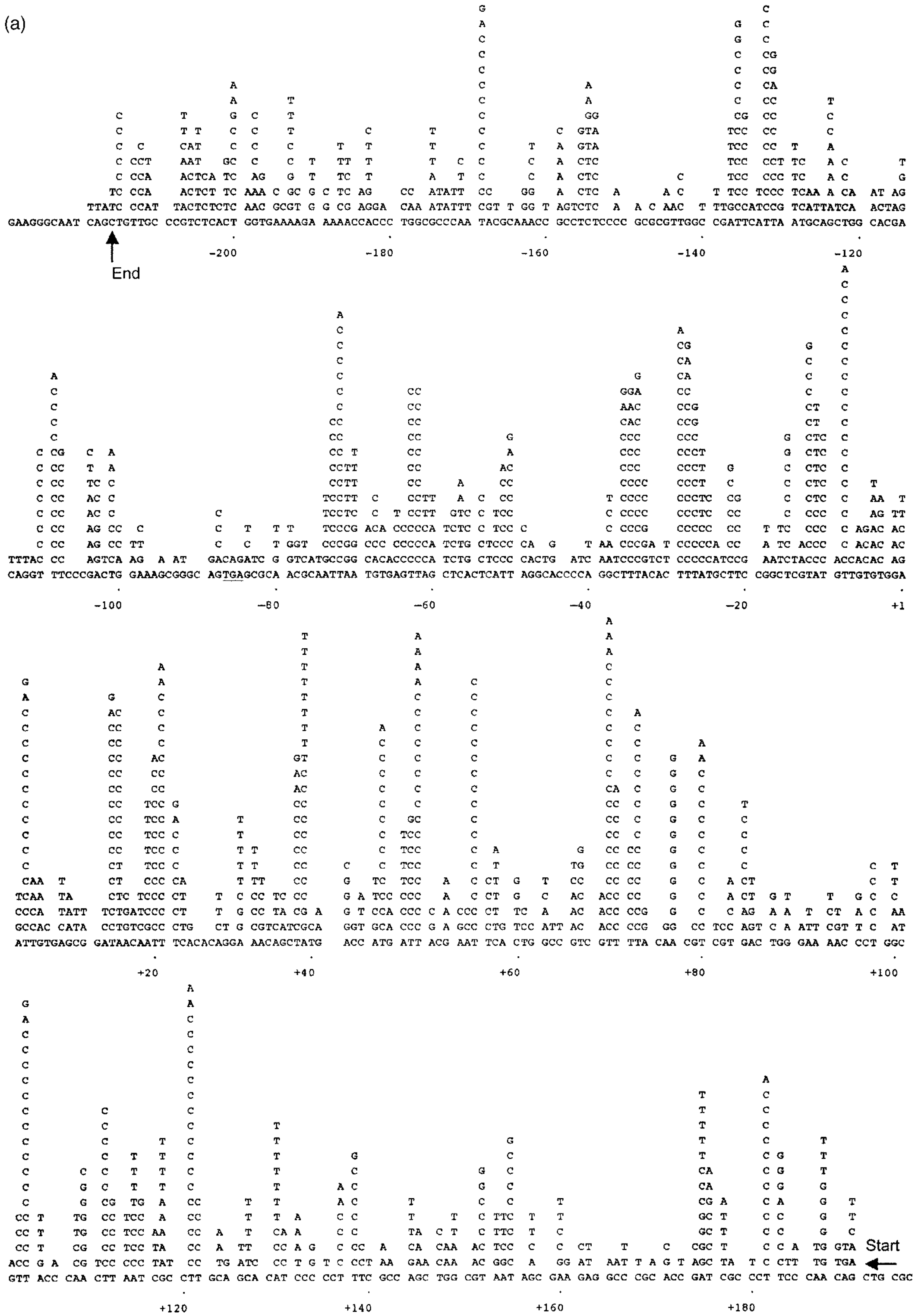


Figure 1 (*legend opposite*)

results with pol η, the single-base frameshift error rates of other DNA polymerases increase substantially with increasing run length,[9] consistent with the strand slippage hypothesis[10]. As one example, the single-base deletion error rate of human pol β in homopolymeric runs of four or five bases is 35-fold higher than for deletion of non-iterated nucleotides (Table 3).

In addition to single-nucleotide frameshift errors, pol η generated additions of two or three nucleo-

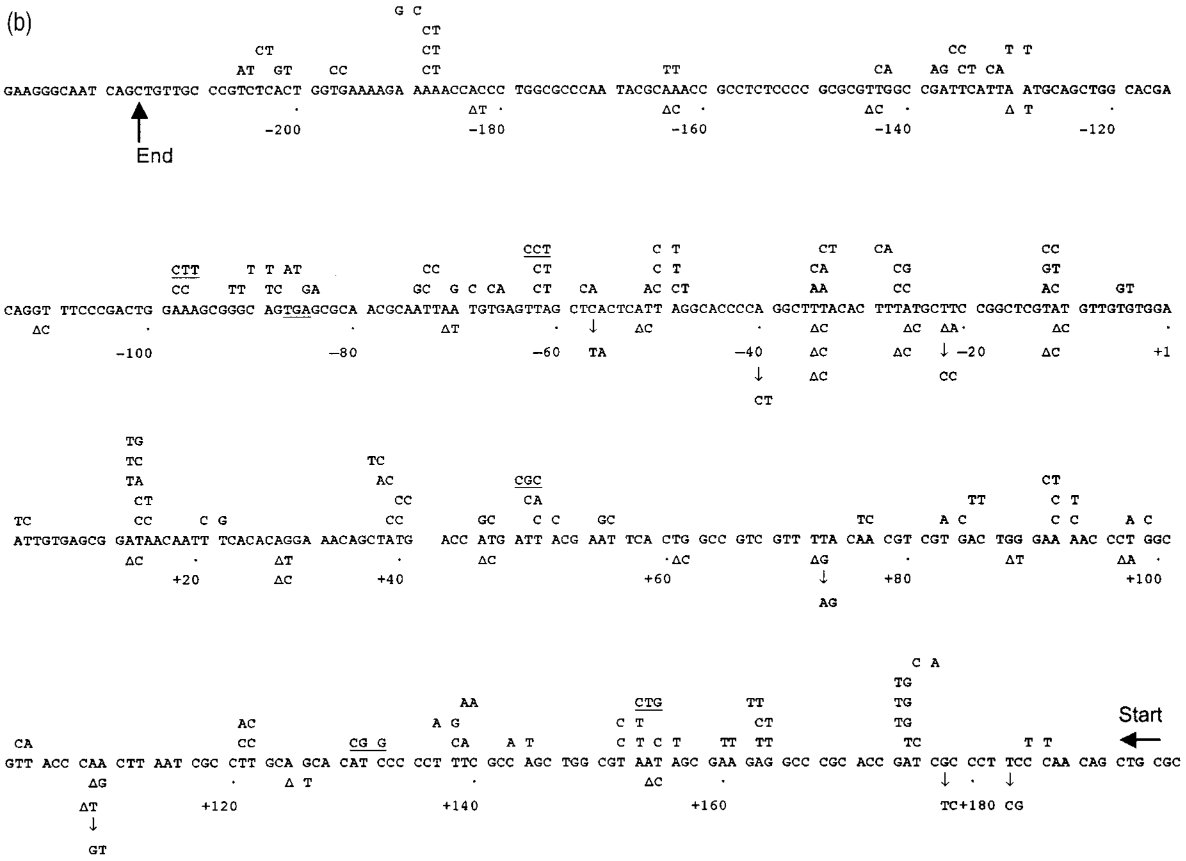


Figure 1. Spectra of base substitutions by human pol η. The 407 nucleotide template is shown as four lines of sequence. Nucleotide +1 is the first transcribed nucleotide of the *lacZ* α-complementation gene in M13mp2 DNA. DNA synthesis begins with incorporation opposite template nucleotide +191, marked with an arrow in the bottom line of sequence. The last single-stranded template nucleotide in the gap is at position −216 (arrow in top line of sequence). The termination codon for the carboxy-terminal end of the *lacI* gene in M13mp2 is underlined (nucleotides −87, −86 and −85) in the second line of template sequence. (a) Spectrum of single-base changes. (b) Spectrum of tandem changes. These include 84 tandem double-substitutions and five tandem triple-substitutions (underlined), shown as adjacent letters above the lines of sequence. Shown below the lines of sequence are 27 events where two nucleotides are replaced by one (ΔX) and seven events in which one nucleotide is replaced by two.

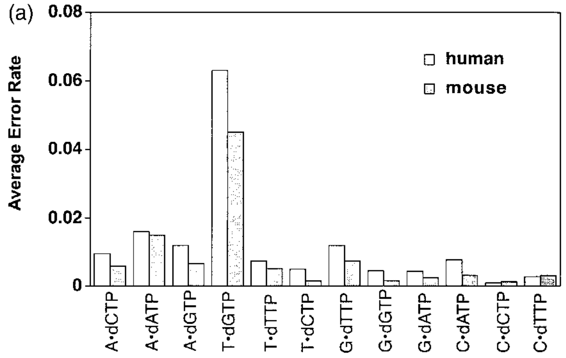
tides (Figure 4(a) and (b)), deletions of two to more than 100 nucleotides (Figure 4(a)-(c)), and a few complex addition/deletion mutations (Table 1). The locations of these errors suggest that some may involve strand slippage and homologous base-pairing (discussed below), but that others may arise by a different mechanism.

Analysis of adjacent point mutations

The errors made by pol η included 84 tandem double-base substitutions. From these data, the calculated error rate for tandem double substitutions is 1.7×10^{-3}. These errors were widely distributed across the target sequence and involved various combinations of mispairs (Figure 1(b)). The extraordinary inaccuracy of human pol η is further illustrated by the recovery of five tandem triple-substitutions (underlined mutants in Figure 1(b)), and tandem events in which two template bases are replaced by one (substitution-deletion, 27 occurrences) and where one base is replaced by two (substitution-addition, seven occurrences).

Extension kinetics for doubly mismatched primer termini

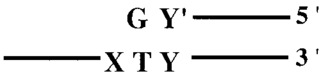
The presence of 84 tandem double and five tandem triple-base substitutions in the error spectrum (Figure 1(b)) indicates that pol η can extend primers containing multiple consecutive terminal mismatches. To investigate the efficiency of this reaction, we determined steady-state kinetic parameters for correct incorporation by human pol η using two different, doubly mismatched termini. The results indicate that pol η extends these substrates (Table 4, lines 7 and 8) with catalytic efficiencies that are only about tenfold lower than for the singly mismatched termini (lines 2, 3, 5 and 6) and 1000-fold lower than for extension of correct termini (lines 1 and 4). The efficiency of extension of these tandem double mismatches by pol η is in



(b)

	A•dCTP	A•dATP	A•dGTP	T•dGTP	T•dTTP	T•dCTP	G•dTTP	G•dGTP	G•dATP	C•dATP	C•dCTP	C•dTTP
Highest	0.097	0.073	0.081	0.15	0.032	0.024	0.04	0.048	0.032	0.048	0.016	0.024

Figure 2. Error rates for 12 mispairs. (a) Average error rates for each of the 12 possible base·base mispairs. (b) Listed is the highest site-specific error rate observed for each mispair, taken from the spectrum in Figure 1(a). In all cases, the lowest error rate was the same at $\leqslant 8 \times 10^{-3}$, because zero or one substitution was seen at (at least) one template site among 124 sequenced *lacZ* mutants. When a Monte Carlo modification of the Pearson χ^2 test of spectra homogeneity was used to compare the human and mouse base substitution spectra,[33,34] no significant difference was found ($P_{\text{hg}} = 0.97$). A strong correlation (tau correlation coefficient $CC = 0.43$, $P_{\text{cc}} < 0.01$[35]) was also observed when the two spectra were compared, as predicted for homogeneous spectra. Finally, various pairwise comparisons of rates indicate that the human and mouse data are similar ($CC = 0.77$, $P_{\text{cc}} < 0.01$, $P_{\text{hg}} = 0.67$). These analyses suggest that the human and mouse enzymes do not differ significantly in error rates or error distribution.



X	Mutations per site
A	9.8
T	8.8
C	6.5
G	6.4

Y•Y'	Mutations per site
A•T	12
T•A	11
C•G	4.9
G•C	4.0

Figure 3. Sequence-context effects on formation of the T·dGMP mispair by human pol η. The analysis is based on the location and number of T → C substitutions shown in Figure 1. We totaled the observed number of T → C substitutions $M(\underline{T}Y)$ in all $\underline{T}Y$ sites, where the substituted site is underlined and Y = A, T, G or C. We calculated the number of TY dinucleotides, $N(TY)$, in the target sequence. The expected number of T → C substitutions in $\underline{T}Y$ sites, $E(\underline{T}Y)$, was estimated as:

$$E(\underline{T}Y) = [N(TY) \times \Sigma M(\underline{T}Y)]/[\Sigma N(TY)]$$

The expected number of T → C substitutions $E(X\underline{T})$ (X = A, T, G, C) in $X\underline{T}$ sites was estimated using the same equation. A chi-square test (χ^2) with three degrees of freedom was then used to evaluate the differences between the observed and expected number of T → C substitutions in $\underline{T}Y$ and $X\underline{T}$ sites. For both $X\underline{T}$ and $\underline{T}Y$ sites, the difference between the expected and the observed number of mutations was highly significant ($P < 0.001$). However, for the $X\underline{T}$ sites, the significance may partly reflect a larger number of WTW trinucleotides (28 sites) in comparison with STS (21 sites), WTS (21 sites) and STW trinucleotides (17 sites) in the target sequence (where W = A or T and S = G or C).

fact similar to the efficiencies with which two other exonuclease-deficient replicative DNA polymerases extend single-base mismatches (Table 4, lines 9-12, from Mendelman *et al.*[11] and Lewis *et al.*[43]).

Discussion

This study of the fidelity and error specificity of DNA polymerase η has both biological and mechanistic implications. Like its human homolog,[6] mouse pol η lacks an intrinsic proofreading exonucleolytic activity and has very low fidelity during copying of undamaged DNA. These facts and steady-state kinetic analysis showing that both yeast[8] and human pol η misinsert nucleotides at rates that are high relative to many other DNA polymerases[6,7] indicate that low-fidelity DNA synthesis is a common property of DNA polymerase η. Following misinsertion, pol η extends mismatched primer termini at rates that, while slower than for extension of matched termini, can nonetheless be substantial.[12,13] In this study, promiscuity during mismatch extension is clearly illustrated by the frequent appearance of tandem double and triple-base substitutions in the human pol η error spectrum (Figure 1(b)). It is indicated also by the ability of human pol η to extend tandem double-mismatched termini at rates similar to those with which other polymerases extend single mismatches (Table 4). This property is interesting in light of our previous suggestion[6] and evidence[12] that slow mismatch extension, and possibly rapid dissociation as suggested by low processivity, may permit proofreading by an extrinsic exonuclease. This could reduce mutagenesis resulting from frequent misinsertions by pol η. However, promiscuous extension of certain mismatches by pol η (e.g. T·dGMP, Table 4) may limit the opportunity for proofreading. In these instances, the cell may reduce the mutagenic potential of low-fidelity pol η in at least two other ways. Strategic protein partnerships may target pol η for use only when needed, e.g. for lesion bypass, or stable misincorporations by pol η may be removed by DNA mismatch repair.

Table 3. Relationship between homopolymeric run length and single-nucleotide deletion and addition error rates of human pol β and pol η

	Error rate (×10^{-5})			
	Single-nucleotide deletions		Single-nucleotide additions	
Homopolymeric run length (nucleotides)	pol β	pol η	pol β	pol η
One	2.0	150	⩽0.4	180
Two	9.3	270	⩽0.7	62
Three	23	410	⩽1.4	57
Four and five	70	320	2.3	160

Both mouse and human pol η generate all 12 single-base·base mispairs (Figure 2). Statistical analysis (see the legend to Figure 2) suggests that the human and mouse enzymes do not differ significantly in either error rate or error distribution. The error rate values in Figure 2 are substantially higher than those of most other DNA polymerases examined to date. The exceptions are African swine fever virus DNA polymerase and human pol ι, which are less accurate than pol η for specific mismatches.[14–17] Structure-function studies[5] suggest that the nucleotide selectivity of more accurate DNA polymerases in the Pol A, Pol X and RT families depends largely on base-pair geometry.[5] The binding pockets of those polymerases are shaped to accommodate a correct Watson-Crick base-pair and to exclude an incorrect pair. The polymerase active site is suggested to exclude water and thus enhance nucleotide selectivity.[18] In light of these ideas, the high base-substitution error rates for pol η (Figure 2) are remarkable, because they are similar to the rates of 1/10 to 1/300 predicted by free-energy differences between correct and incorrect base-pairs in duplex DNA in aqueous solution.[19,20] This suggests that pol η may have a more relaxed requirement for shape complementarity in order to achieve efficient catalysis during bypass of certain helix-distorting lesions, and that its active site may exclude water less effectively than more accurate polymerases.

Under these circumstances, incoming nucleotides may be stabilized by base-stacking interactions. With this possibility in mind, we performed an initial, simple analysis to determine if the nucleo-

Table 4. Kinetic constants for extension of various primer termini by human pol η

Template·primer (substrate)	Correct dNTP	K_m (μM)	$k_{cat(app)}$ (min^{-1})	$(k_{cat}/K_m)_{app}$ (×10^3)(μM·min)$^{-1}$	f_{ext}
G·C (a)	dATP	31 ± 8.8	13 ± 4.4	420 ± 190	1
G·T (a)	dATP	730 ± 160	3.3 ± 1.1	4.5 ± 1.8	0.011
G·A (a)	dATP	440 ± 320	1.6 ± 0.9	3.5 ± 3.3	0.008
T·A (b)	dGTP	37 ± 8.2	11 ± 0.0	310 ± 69	1
T·G (b)	dGTP	85 ± 48	8.4 ± 0.9	98 ± 56	0.32
T·C (b)	dGTP	960 ± 110	3.2 ± 1.8	3.3 ± 1.9	0.011
G·T & T·C (c)	dGTP	1300 ± 140	0.44 ± 0.18	0.32 ± 0.14	0.001
G·A & T·C (c)	dGTP	1100 ± 160	0.30 ± 0.08	0.28 ± 0.09	0.0009
Pol α-G·A[11]	dGTP				0.0006
Pol α-T·C[11]	dTTP				0.0002
HIV-1 RT-T·G[40]	dTTP				0.0009
HIV-1 RT-T·C[40]	dTTP				0.0002

The results shown in the first six lines are from.[15] Reactions were as described in Experimental Procedures, using the following substrates:

```
           dATP
(a)          ↓XTCTTTTGGGACCGCAATGG-5´* - where X = C, T or A
              0•••••••••••••••••••
5´-ACGTCGTGACTGAGAAAACCCTGGCGTTACCCA-3´
          dGTP
(b)         ↓XCTCTTTTGGGACCGCAATGG-5′* - where X = A, G or C
             0••••••••••••••••••••
5´-ACGTCGTGACTGAGAAAACCCTGGCGTTACCCA-3′
          dGTP
(c)         ↓YXTCTTTTGGGACCGCAATGG-5′* - where X = C, T or A
             00•••••••••••••••••••
5´-ACGTCGTGACTGAGAAAACCCTGGCGTTACCCA-3′- and where Y = A or C
```

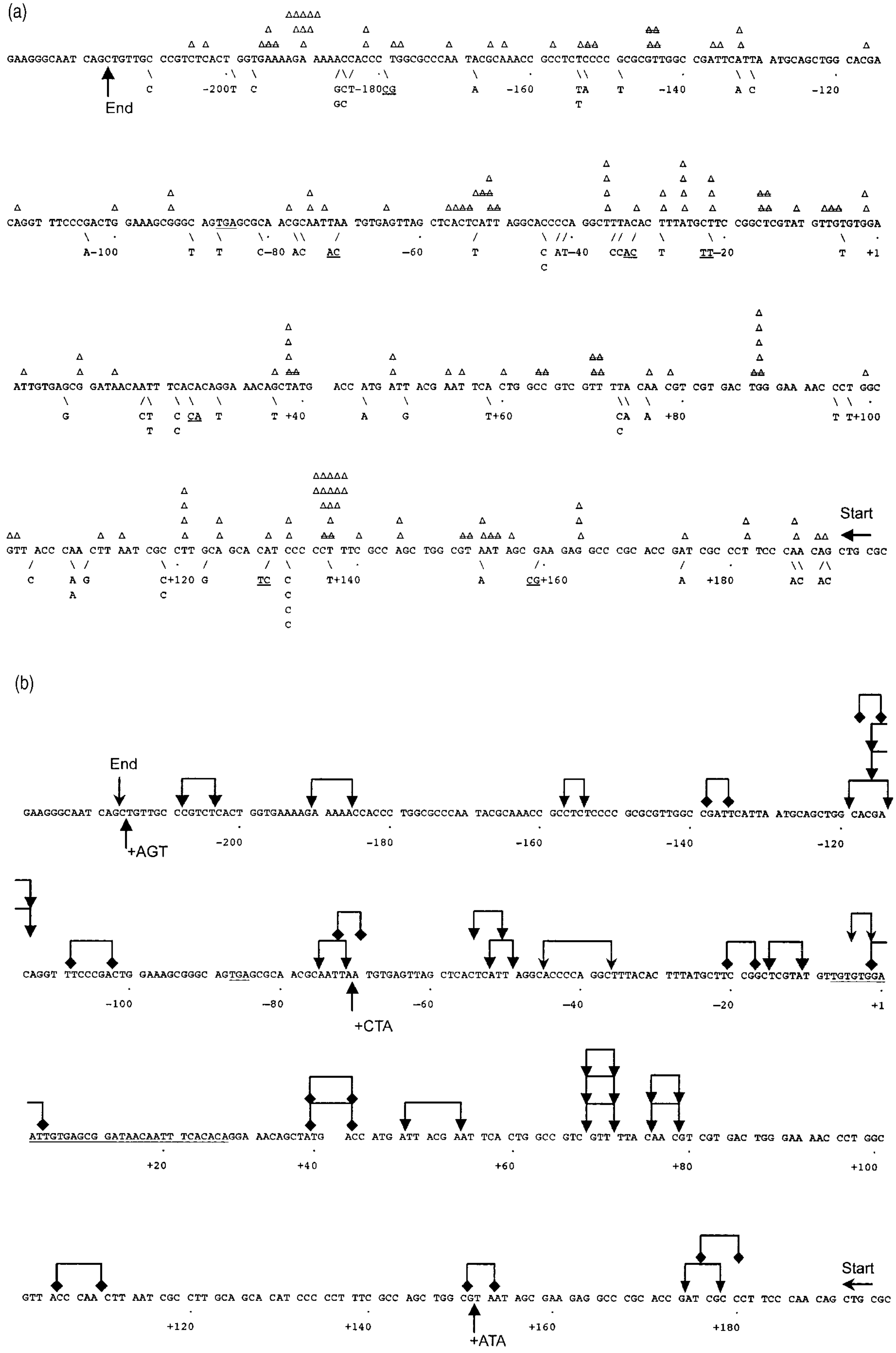

(a)
GAAGGGCAAT CAGCTGTTGC CCGTCTCACT GGTGAAAAGA AAAACCACCC TGGCGCCCAA TACGCAAACC GCCTCTCCCC GCGCGTTGGC CGATTCATTA ATGCAGCTGG CACGA
End
-200
-180
-160
-140
-120
CAGGT TTCCCGACTG GAAAGCGGGC AGTGAGCGCA ACGCAATTAA TGTGAGTTAG CTCACTCATT AGGCACCCCA GGCTTTACAC TTTATGCTTC CGGCTCGTAT GTTGTGTGGA
-100
-80
-60
-40
-20
+1
ATTGTGAGCG GATAACAATT TCACACAGGA AACAGCTATG ACC ATG ATT ACG AAT TCA CTG GCC GTC GTT TTA CAA CGT CGT GAC TGG GAA AAC CCT GGC
+40
+60
+80
+100
GTT ACC CAA CTT AAT CGC CTT GCA GCA CAT CCC CCT TTC GCC AGC TGG CGT AAT AGC GAA GAG GCC CGC ACC GAT CGC CCT TCC CAA CAG CTG CGC
+120
+140
+160
+180
Start
(b)
End
+AGT
+CTA
+20
+ATA
Start

Figure 4 (*legend opposite*)

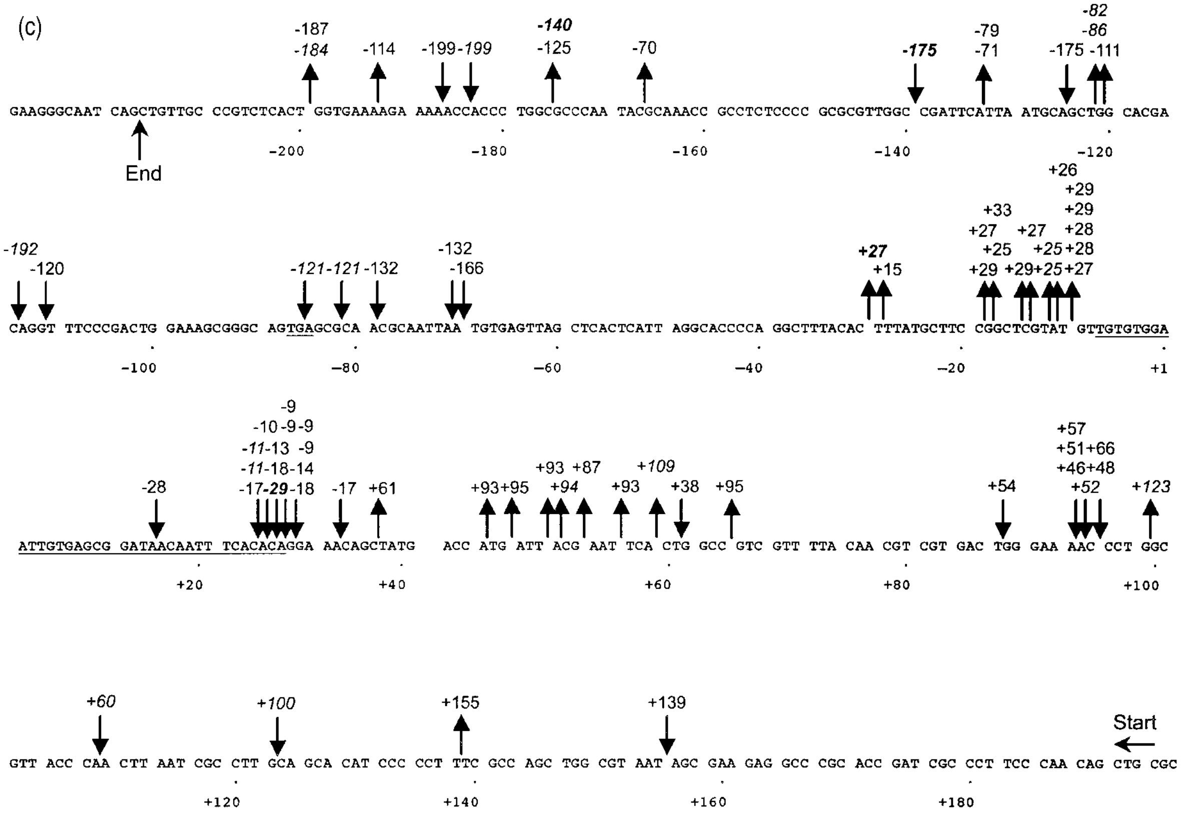

(c)
-187
-184
-114
-199-199
-140
-125
-70
-175
-79
-71
-82
-86
-175 -111
GAAGGGCAAT CAGCTGTTGC CCGTCTCACT GGTGAAAAGA AAAACCACCC TGGCGCCCAA TACGCAAACC GCCTCTCCCC GCGCGTTGGC CGATTCATTA ATGCAGCTGG CACGA
End
-200
-180
-160
-140
-120
-192
-120
-121-121-132
-132
-166
+27
+15
+33
+27 +27
+25 +25+28
+29 +29+25+27
+26
+29
+29
+28
CAGGT TTCCCGACTG GAAAGCGGGC AGTGAGCGCA ACGCAATTAA TGTGAGTTAG CTCACTCATT AGGCACCCCA GGCTTTACAC TTTATGCTTC CGGCTCGTAT GTTGTGTGGA
-100
-80
-60
-40
-20
+1
-28
-9
-10-9-9
-11-13 -9
-11-18-14
-17-29-18
-17
+61
+93+95
+93 +87
+94
+93
+109
+38
+95
+54
+57
+51+66
+46+48
+52
+123
ATTGTGAGCG GATAACAATT TCACACAGGA AACAGCTATG ACC ATG ATT ACG AAT TCA CTG GCC GTC GTT TTA CAA CGT CGT GAC TGG GAA AAC CCT GGC
+20
+40
+60
+80
+100
+60
+100
+155
+139
Start
GTT ACC CAA CTT AAT CGC CTT GCA GCA CAT CCC CCT TTC GCC AGC TGG CGT AAT AGC GAA GAG GCC CGC ACC GAT CGC CCT TCC CAA CAG CTG CGC
+120
+140
+160
+180

Figure 4. Spectra of deletion and addition errors by human pol η. The template sequence is the same as that shown in Figure 1. (a) Spectra of deletions (triangles) and additions (letters) of one or two nucleotides. The deletions are shown above each line of sequence, with single-base deletions as simple triangles and two-base deletions shown as adjacent triangles with lines. When deletions occur within repeat sequences, the exact base deleted is not known. In these cases, the triangle(s) are placed at the left border of the repetitive sequence. Additions are shown below each line of sequence, with single-base events indicated as simple letters and two-base additions shown as consecutive underlined letters. (b) Spectra of additions of three bases and deletions of three to nine bases. Three different three-base additions are shown as letters below the template sequence. Deletions are indicated by connected lines above the primary sequence. The two events designated with small arrowheads are deletions flanked by direct repeats. Events with the larger arrowheads are deletions whose endpoints are flanked by imperfect direct repeats, where a single misinsertion during synthesis of the first repeat would result in a perfect match upon pairing with the downstream repeat. The remaining events, designated with diamonds, are deletions flanked by sequences with little or no apparent homology. (c) The spectrum of 38 deletions of ten or more bases. Each deletion is defined by two arrows; one points down and the other points up. Each arrow is associated with a number indicating the nucleotide position at which the other endpoint for that deletion is located. As one example, a 17 nucleotide deletion was recovered whose endpoints are nucleotides +139 and +155. This deletion is delineated by two arrows just above the lowest line of primary sequence. One arrow is associated with +139, and it points down to indicate the deletion endpoint at nucleotide +155. Similarly, the arrow pointing up at nucleotide +139 indicates the other endpoint and is associated with the number +155. By using a similar pairing strategy, all 38 deletions can be identified with respect to the number of deleted nucleotides, the location of the endpoints and the nucleotides at the junction. Bold italicized numbers indicate deletions flanked by direct repeats. Non-bold italicized numbers indicate deletions whose endpoints are imperfect direct repeats, where a single misinsertion during copying of the first repeat results in a perfect match upon pairing with the downstream repeat. Plain numbers designate deletions flanked by sequences with no apparent homology.

tides immediately flanking the misincorporation site influenced the error rate. As typically seen with other DNA polymerases, pol η error rates vary over a substantial range that depends on the mismatch composition, the mismatch symmetry (template *versus* incoming dNTP), and on the nucleotide sequence surrounding the mismatch. Analysis of the distribution of the 665 observed T·dGMP errors (Figure 3) indicates a higher error rate when the base-pair preceding the error is T·A or A·T than when it is G·C or C·G. This precludes the simple notion that the incoming incorrect dGTP is stabilized for incorporation more effectively when stacked with an adjacent purine rather than a pyrimidine. It may be that misincorporation of dGTP is favored when the primer terminus is slightly less stable (A·T rather than G·C) and therefore more easily positioned for catalysis. The error spectrum shown in Figure 1 is the largest collection yet obtained for any DNA polymerase, providing a rich opportunity for a more detailed analysis of sequence-context effects on fidelity. Whatever the explanation for the non-random distribution of base-substitution errors, the data for human and mouse pol η obtained here has already implied a novel and unanticipated biological role for this inaccurate polymerase. Statistical analysis of the pattern of substitutions in comparison to the specificity of somatic hypermutagenesis of immunoglobulin genes suggests that pol η may participate in strand-specific hypermutation at A·T basepairs.[21] If this hypothesis is correct, then the low fidelity with which pol η copies undamaged DNA may provide an additional selective advantage to mammals beyond its role in translesion synthesis.

While most investigations of the fidelity of pol η have focused on formation and extension of base-base mismatches, the present study clearly illustrates that both human and mouse pol η also generate a variety of nucleotide deletion and addition errors (Tables 1-3, Figure 4). The error-specificity of human pol η is remarkable in both the variety of errors observed (Figure 4) and in the rate at which they are generated. The average human pol η error rates for single-nucleotide additions and deletions greatly exceed the frameshift error rates of most other DNA polymerases (Table 2). The only exception is human pol κ, whose single-base frameshift error rates[22] approach those shown here for pol η. One logical explanation for formation of many of the frameshifts shown in Figure 4 is strand slippage involving repetitive sequences.[10] A slippage mechanism[9,23] may explain many of the frameshifts observed within homopolymeric sequences (Figure 4(a)) as well as those between distant directly repeated sequences (Figure 4(b) and (c)). The ability of pol η to extend termini with one, two or even three mismatches may explain the high frequency and diverse nature of the deletions shown in Figure 4. Those deletions with limited homology at the endpoints, which includes most of those shown in Figure 4(b) and (c), may involve slippage between imperfect repeats. As proposed in earlier studies of single[24–26] and multiple-base deletions,[27] some of the deletions seen here may be initiated by a misinsertion that subsequently finds more favorable homology by relocating the primer to a downstream location. An additional factor contributing to primer relocation could be DNA secondary structure in the template. This is suggested by the clustering of deletion endpoints in the vicinity of a palindrome in the template (underlined sequence in Figure 4(c)) that can form a stable hairpin. Also note that the rate at which pol η misinserts nucleotides is much higher than the deletion rate. Thus, misinsertion followed by primer relocation to create a misalinged template-primer can readily explain the extraordinarily high rate of single-base frameshifts observed at non-iterated template positions (Table 3). Still another mechanism that could explain this class of errors is one in which a deletion intermediate is stabilized by stacking interactions with the incoming dNTP.[28–31]

The fact that the single-base addition error rate of pol η is much higher than those of other polymerases (Table 2), and that additions of two or three bases are observed (Figure 4) implies that pol η has a remarkable capacity to perform synthesis using primers containing unpaired nucleotides. This is further illustrated by the fact that the difference in error rates for additions and deletions by pol η is only twofold, whereas the difference is over tenfold with most other polymerases (Table 2). Moreover, many of the additions are not at repetitive sequence locations, suggesting that the additions may be initiated by a mechanism other than strand slippage. Hypothetically, this could involve misinsertion followed by primer relocation to create an unpaired nucleotide in the primer strand. However, more unusual models may be needed to account for the large number of additions wherein the nucleotide that is added differs from either of its neighbors (Figure 4(a)). Understanding the extraordinary infidelity and error-specificity of this remarkable DNA polymerase will undoubtedly be facilitated by structural information indicating how pol η differs from other polymerases in its interactions with undamaged and damaged DNA at and upstream of the polymerase active site.

Experimental Procedures

Materials

All materials for the fidelity assay were from previously described sources[41]. C-terminal hexahistidine-tagged human and mouse DNA polymerases η were expressed in insect cells and purified as described.[1,42]

DNA synthesis reactions and analysis of fidelity

Reactions (25 μl) were performed as described,[6,11] and contained 1.4 nM M13mp2 DNA with a 407-nucleotide gap (from nucleotide −216 through +191 of the *lacZ* gene), 40 mM Tris-HCl (pH 8.0), 10 mM $MgCl_2$, 10 mM dithiothreitol, 6.25 μg of bovine serum albumin (BSA), 60 mM KCl, 2.5% (v/v) glycerol and 10 μM, 100 μM or 1 mM dNTPs. Synthesis was initiated by adding 36 or 72 nM pol η. Reactions were incubated at 37 °C for one hour and terminated by adding EDTA to 15 mM. Agarose gel electrophoresis, performed as described,[6] revealed that human and mouse pol η filled the gap to completion. DNA products of the reactions with 72 nM pol η were examined for the frequency of *lacZ* mutants as described.[6] DNA from independent *lacZ* mutant phage was sequenced to identify the errors made by pol η during gap-filling synthesis. Error rates were calculated as described above.

Mismatch extension kinetics

Reactions (25 μl) were performed as for fidelity measurements, except that they contained 200 nM template primed at a 1.2:1 molar ratio with a (5′-^{32}P)-labeled primer, 2 nM pol η and correct dATP or dGTP. The template-primers used are shown in Table 4. Duplicate determinations were performed with each template-primer using seven different concentration of the correct nucleotide. Aliquots were removed at two, four, six and eight minutes, and products were separated by electrophoresis in denaturing 16% (w/v) polyacrylamide gels and quantified by phosphorimagery. Kinetic constants were derived as described.[43]

Acknowledgments

We thank Matthew Longley and Anna Bebenek for critical evaluation of the manuscript.

References

1. Masutani, C., Kusumoto, R., Yamada, A., Dohmae, N., Yokoi, M. & Yuasa, M., *et al*. (1999). The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase η. *Nature,* **399**, 700-704.
2. Johnson, R. E., Kondratick, C. M., Prakash, S. & Prakash, L. (1999). hRAD30 mutations in the variant form of xeroderma pigmentosum. *Science,* **285**, 263-265.
3. Masutani, C., Araki, M., Yamada, A., Kusumoto, R., Nogimori, T., Maekawa, T. *et al*. (1999). Xeroderma pigmentosum variant (XP-V) correcting protein from HeLa cells has a thymine dimer bypass DNA polymerase activity. *EMBO J.* **18**, 3491-3501.
4. Johnson, R. E., Prakash, S. & Prakash, L. (1999). Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Pol η. *Science,* **283**, 1001-1004.
5. Kunkel, T. A. & Bebenek, K. (2000). DNA replication fidelity. *Annu. Rev. Biochem.* **69**, 497-529.
6. Matsuda, T., Bebenek, K., Masutani, C., Hanaoka, F. & Kunkel, T. A. (2000). Low fidelity DNA synthesis by human DNA polymerase-η. *Nature,* **404**, 1011-1013.
7. Johnson, R. E., Washington, M. T., Prakash, S. & Prakash, L. (2000). Fidelity of human DNA polymerase η. *J. Biol. Chem.* **275**, 7447-7450.
8. Washington, M. T., Johnson, R. E., Prakash, S. & Prakash, L. (1999). Fidelity and processivity of *Saccharomyces cerevisiae* DNA polymerase η. *J. Biol. Chem.* **274**, 36835-36838.
9. Bebenek, K. & Kunkel, T. A. (2000). *Streisinger revisited: DNA synthesis errors mediated by substrate misalignments*. Cold Spring Harbor Symp. Quant. Biol. **65**, 81-91.
10. Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. & Inouye, M. (1966). Frameshift mutations and the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77-84.
11. Mendelman, L. V., Petruska, J. & Goodman, M. F. (1990). Base mispair extension kinetics. Comparison of DNA polymerase alpha and reverse transcriptase. *J. Biol. Chem.* **265**, 2338-2346.
12. Bebenek, K., Matsuda, T., Masutani, C., Hanaoka, F. & Kunkel, T. A. (2001). Proofreading of DNA polymerase η-dependent replication errors. *J. Biol. Chem.* **276**, 2317-2320.
13. Washington, M. T., Johnson, R. E., Prakash, S. & Prakash, L. (2001). Mismatch extension ability of yeast and human DNA polymerase η. *J. Biol. Chem.* **276**, 2263-2266.
14. Showalter, A. K. & Tsai, M.-D. (2001). A DNA polymerase with specificity for five base-pairs. *J. Am. Chem. Soc.* **123**, 1776-1777.
15. Tissier, A., McDonald, J. P., Frank, E. G. & Woodgate, R. (2000). pol ι, a remarkably error-prone human DNA polymerase. *Genes Dev.* **14**, 1642-1650.
16. Zhang, Y., Yuan, F., Wu, X. & Wang, Z. (2000). Preferential incorporation of G opposite template T by the low-fidelity human DNA polymerase ι. *Mol. Cell. Biol.* **20**, 7099-7108.
17. Johnson, R. E., Washington, M. T., Haracska, L., Prakash, S. & Prakash, L. (2000). Eukaryotic polymerases ι and ζ act sequentially to bypass DNA lesions. *Nature,* **406**, 1015-1019.
18. Petruska, J., Sowers, L. C. & Goodman, M. F. (1986). Comparison of nucleotide interactions in water, proteins, and vacuum: model for DNA polymerase fidelity. *Proc. Natl Acad. Sci. USA,* **83**, 1559-1562.
19. Raszka, M. & Kaplan, N. O. (1972). Association by hydrogen bonding of mononucleotides in aqueous solution. *Proc. Natl Acad. Sci. USA,* **69**, 2025-2029.
20. Mildvan, A. S. (1974). Mechanism of enzyme action. *Annu. Rev. Biochem.* **43**, 357-399.
21. Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T. & Kunkel, T. A. (2001). Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nature Immun.* **2**, 530-536.
22. Ohashi, E., Bebenek, K., Matsuda, T., Feaver, W. J., Gerlach, V. L., Friedberg, E. C. *et al*. (2000). Fidelity and processivity of DNA synthesis by DNA polymerase kappa, the product of the human DINB1 gene. *J. Biol. Chem.* **275**, 39678-39684.
23. Kroutil, L. C. & Kunkel, T. A. (1998). DNA replication errors involving strand misalignments. In *Genetic Instabilities and Hereditary Neurological Diseases* (Wells, R. D. & Warren, S. T., eds), pp. 699-716, Academic Press, San Diego.
24. Kunkel, T. A. & Soni, A. (1988). Mutagenesis by transient misalignment. *J. Biol. Chem.* **263**, 14784-14789.
25. Bebenek, K. & Kunkel, T. A. (1990). Frameshift errors initiated by nucleotide misincorporation. *Proc. Natl Acad. Sci. USA,* **87**, 4946-4950.
26. Bebenek, K., Roberts, J. D. & Kunkel, T. A. (1992). The effects of dNTP pool imbalances on frameshift fidelity during DNA replication. *J. Biol. Chem.* **267**, 3589-3596.
27. Cai, H., Yu, H., McEntee, K., Kunkel, T. A. & Goodman, M. F. (1995). Purification and properties of wild-type and exonuclease-deficient DNA polymerase II from *Escherichia coli*. *J. Biol. Chem.* **270**, 15327-15335.
28. Kunkel, T. A. (1986). Frameshift mutagenesis by eucaryotic DNA polymerases *in vitro*. *J. Biol. Chem.* **261**, 13581-13587.
29. Osheroff, W. P., Beard, W. A., Yin, S., Wilson, S. H. & Kunkel, T. A. (2000). Minor groove interactions at the DNA polymerase beta active site modulate single-base deletion error rates. *J. Biol. Chem.* **275**, 28033-22808.
30. Efrati, E., Tocco, G., Eritja, R., Wilson, S. H. & Goodman, M. F. (1997). Abasic translesion synthesis by DNA polymerase beta violates the ''A-rule''.

Novel types of nucleotide incorporation by human DNA polymerase beta at an abasic lesion in different sequence contexts. *J. Biol. Chem.* **272**, 2559-2569.

31. Hashim, M. F., Schnetz-Boutaud, N. & Marnett, L. J. (1997). Replication of template-primers containing propanodeoxyguanosine by DNA polymerase beta. Induction of base-pair substitution and frameshift mutations by template slippage and deoxynucleoside triphosphate stabilization. *J. Biol. Chem.* **272**, 20205-20212.
32. Umar, A., Boyer, J. C. & Kunkel, T. A. (1994). DNA loop repair by human cell extracts. *Science*, **266**, 814-816.
33. Adams, W. T. & Skopet, T. R. (1987). Statistical test for the comparison of samples from mutational spectra. *J. Mol. Biol.* **194**, 391-396.
34. Piegorsch, W. W. & Bailer, A. J. (1994). Statistical approaches for analyzing mutational spectra: some recommendations for categorical data. *Genetics*, **136**, 403-416.
35. Babenko, V. N. & Rogozin, I. B. (1999). Use of a rank correlation coefficient for comparing mutational spectra. *Biofizika*, **44**, 632-638.
36. Roberts, J. D. & Kunkel, T. A. (1996). Eukaryotic DNA replication fidelity. In *DNA Replication in Eukaryotic Cells: Concepts, Enzymes and Systems* (Pamphilis, M. D., ed.), pp. 217-247, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
37. Bebenek, K., Abbotts, J., Roberts, J. D., Wilson, S. H. & Kunkel, T. A. (1989). Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. *J. Biol. Chem.* **264**, 16948-16956.
38. Bebenek, K., Abbotts, J., Wilson, S. H. & Kunkel, T. A. (1993). Error-prone polymerization by HIV-1 reverse transcriptase: contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots. *J. Biol. Chem.* **268**, 10324-10334.
39. Bebenek, K., Joyce, C. M., Fitzgerald, M. P. & Kunkel, T. A. (1990). The fidelity of DNA synthesis catalyzed by derivatives of *Escherichia coli* DNA polymerase I. *J. Biol. Chem.* **265**, 13878-13887.
40. Yu, H. & Goodman, M. F. (1992). Comparison of HIV-1 and avian virus reverse transcriptase fidelity on RNA and DNA templates. *J. Biol. Chem.* **267**, 10888-10896.
41. Bebenek, K. & Kunkel, T. A. (1995). Analyzing the fidelity of DNA polymerases. *Methods Enzymol.* **262**, 217-232.
42. Masutani, C., Kusumoto, R., Iwai, S. & Hanaoka, F. (2000). Mechanisms of accurate translesion synthesis by human DNA polymerase η. *EMBO J.* **19**, 3100-3109.
43. Lewis, D. A., Bebenek, K., Beard, W. A., Wilson, S. H. & Kunkel, T. A. (1999). Uniquely altered DNA replication fidelity conferred by an amino acid change in the nucleotide binding pocket of human immunodeficiency virus type 1 reverse transcriptase. *J. Biol. Chem.* **274**, 32924-32930.

Edited by M. Yaniv

(Received 30 April 2001; received in revised form 11 July 2001; accepted 11 July 2001)