Misalignment-Mediated DNA Polymerase β Mutations: Comparison of Microsatellite and Frame-Shift Error Rates Using a Forward Mutation Assay[†]

Kristin A. Eckert,* Andrew Mowery, and Suzanne E. Hile

Department of Pathology, Gittlen Cancer Research Institute, and Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey Medical Center, 500 University Drive, Hershey, Pennsylvania 17033

Received April 3, 2002; Revised Manuscript Received June 19, 2002

ABSTRACT: Mutations arising in microsatellite DNA are associated with neurological diseases and cancer. To elucidate the molecular basis of microsatellite mutation, we have determined the in vitro polymerase error frequencies at microsatellite sequences representative of those found in the human genome: $[GT/CA]_{10}$, $[TC/AG]_{11}$, and $[TTCC/AAGG]_9$. DNA templates contained the microsatellites inserted in-frame into the 5′ region of the herpes simplex virus thymidine kinase (HSV-tk) gene. Polymerase β (polβ) error frequencies were quantitated in microsatellite sequences, relative to frame-shift error frequencies in coding sequences, from the same DNA synthesis reaction. The polβ error frequencies within the dinucleotide sequences were $(2-9) \times 10^{-3}$, 14−72-fold higher than the ssDNA template frequencies. The polβ error frequencies within the tetranucleotide sequences were $(4-6) \times 10^{-3}$, a 4−13-fold increase over background. Strand biases were observed for the $[TC/AG]_{11}$ and $[TTCC/AAGG]_9$ alleles, in which more errors were produced when the purine strand served as a template. Mutations within each microsatellite included noncanonical base substitution events and single nucleotide deletions as well as the expected unit length changes. An exponential relationship was observed between the polymerase error frequency per site and both the number of repetitive units and total length of the allele. Our observations are consistent with the strand slippage model of microsatellite mutagenesis and demonstrate that DNA sequence and/or structural differences result in mutational strand biases. To our knowledge, this is the first direct quantitation of DNA polymerase errors in vitro using template microsatellite sequences.

A substantial body of evidence indicates that microsatellite sequences can influence DNA metabolism. The observed effects on replication (*1*−*4*) and recombination (*5*) suggest a genetic role, while influences of microsatellites on gene expression (*6*−*8*) suggest an epigenetic role for this segment of the human genome. Over the past decade, alterations in microsatellite DNA sequences have been shown to be causally related to the development of several human neurological diseases (*9*, *10*) and to be diagnostic of a mutator phenotype arising in certain tumors (*11*, *12*). Microsatellite mutations have been proposed to occur by slipped strand mispairing within the repetitive DNA sequences (*13*). Mechanistically, two distinct pathways of slipped strand mispairing can explain the production of microsatellite mutations. During recombination, unequal crossing over between repetitive arrays located on separate DNA molecules may result in mutant products. Alternatively, a DNA misalignment within the repeat array may occur transiently during DNA synthesis. If used as a substrate by DNA polymerases, such misalignments can result in the addition or deletion of repeat units within the microsatellite. Current in vivo experimental data favor the latter polymerase slippage model (*9*, *14*−*16*). For example, the rate of [GT/CA] instability is similar in *rad*52 recombination-defective and wild-type yeast strains (*17*). In contrast, inactivation of DNA polymerase proofreading (*18*−*20*) and mismatch repair proteins (*21*−*23*) can result in an increased rate of microsatellite mutagenesis. However, these observations do not necessarily indicate that microsatellite mutations arise only during replicative (S-phase) DNA synthesis. Exonuclease-proficient polymerases and mismatch repair proteins operate during most long-tract DNA repair and homologous recombination pathways (*24*, *25*).

DNA polymerase-mediated frame-shift mutations occur in homopolymeric sequences during DNA synthesis in vitro in a manner consistent with the polymerase slippage model (*26*). While a large body of information is available regarding the molecular and structural bases of single base frame-shift mutagenesis (*27*), much less is known about the molecular basis of microsatellite mutation. That mechanistic differences may exist between frame-shift and microsatellite mutagenesis is suggested by physical studies of DNA duplexes containing bulged nucleotides. Increasing the number of unpaired bases within a DNA bulge loop results in an increasing degree of structural perturbations within the duplex (*28*, *29*). Moreover, traditional in vitro polymerase mutation assays utilize protein-coding sequences that do not contain repetitive DNA of the form found interspersed throughout the human genome as microsatellites. Importantly, microsatellites vary extensively by sequence (*30*, *31*), and several common microsatellite sequences found within the genome have the potential for adopting non-B form DNA conformations, including Z-DNA, H-DNA (triplex DNA), and cruciform structures (*32*).

[†] This study was supported by USPHS Grant CA73649.

* Corresponding author. Telephone: (717) 531-4065. Fax: (717) 531-5634. E-mail: kae4@psu.edu.

10.1021/bi025918c CCC: $22.00 © 2002 American Chemical Society
Published on Web 07/25/2002

Previous studies in our laboratory have determined microsatellite mutation rates in nontumorigenic human lymphoblastoid cells using an episomal shuttle vector system (*33*, *34*). On the basis of these and other data, we proposed that the interplay of several biochemical factors, including DNA misalignment potential and DNA secondary structure, regulates both the rate and specificity of somatic cell microsatellite mutation. The goals of the current study were to determine directly the degree to which DNA sequence and structure affect microsatellite mutagenesis and to validate the DNA polymerase strand slippage model. The microsatellites studied, $[GT/CA]_{10}$, $[TC/AG]_{11}$, and $[TTCC/AAGG]_9$, are present in the human genome (*30*), have the potential to form non-B DNA structures (*32*), and are identical to those analyzed previously during ex vivo replication in human cells. The HSV-tk[1] in vitro system developed for these analyses allowed quantitation of absolute polymerase error frequencies in the di- and tetranucleotide microsatellite sequences, relative to error frequencies in protein coding sequences, from the same DNA synthesis reaction.

DNA polymerase β (polβ) was chosen to evaluate the in vitro accuracy of DNA polymerase synthesis through microsatellite sequences for several reasons. First, polβ contains no associated 3′→5′ exonuclease activity which has been reported to affect the frequency of both frame-shift and microsatellite mutations (*18*−*20*). Second, polβ has been shown to produce a significant number of misalignment-mediated errors in coding sequences in vitro (*35*, *36*), allowing a robust number of independent mutants to be generated.

EXPERIMENTAL PROCEDURES

Reagents. Recombinant DNA polymerase β was purified as a hexahistidine fusion protein from *Escherichia coli* as previously described (*37*). Briefly, polβ expression was induced with IPTG, and cell extracts were mixed with Ni-NTA resin. After washing and protein elution, fractions containing protein were loaded on a column containing single-stranded DNA−cellulose, and protein was eluted.

Construction of Vectors. The HSV-tk-containing vector pGTK3 (chloramphenicol sensitive, Cm^s) is a derivative of the pGem3Zf(−) phagemid and has been previously described (*36*). The *Eco*RI site at position 5927 of pGTK3 was altered to a *Bam*HI site by oligonucleotide site-directed mutagenesis, creating a *Bam*HI cassette that contains the HSV-tk coding sequence and promoter region. Repeated attempts to rescue ssDNA from pGem3Zf(+) phagemid clones containing this *Bam*HI cassette were unsuccessful. Therefore, this cassette, along with a 958 bp *Bam*HI fragment that contains either a functional (Cm^R) or nonfunctional (Cm^s) chloramphenicol acetyltransferase gene (cat), was cloned into the pGem3Zf(−) phagemid. This strategy produced clones which varied as to the relative orientations of the HSV-tk and cat genes, as well as the functionality of the cat gene (Figure 1A). The pRS1 and pRAS1 vectors contain a functional cat gene and confer chloramphenicol resistance to plasmid-bearing bacteria. The HSV-tk gene is orientated 5′ to 3′, relative to the f_1 origin, in pRS1, such that the HSV-tk sense strand is produced as ssDNA. In pRAS1, the HSV-tk gene is oriented 3′ to 5′, relative to the f_1 origin, and the antisense strand is produced as ssDNA. The pSS1 and pSAS1 vectors are analogous to the pRS1 and pRAS1 vectors, except that the vectors contain a nonfunctional *cat* gene; hence, bacteria containing these vectors are chloramphenicol sensitive. Construction of plasmids containing $[GT/CA]_{10}$, $[TC/AG]_{11}$, and $[TTCC/AAGG]_9$ microsatellite sequences has been described (*34*, *38*). All artificial microsatellite sequences are inserted in-frame between bases 111 and 112 of the HSV-tk gene, in the sequence context [GT (insert) TCTC]. To construct DNA templates for use in the in vitro synthesis reactions, we subcloned the *Bgl*II to *Bss*HII fragment of the HSV-tk gene from these vectors into the pSS1 and pSAS1 vectors, creating vectors pSS/pSAS 2, 4, and 5.1 (see Table 1 for microsatellite sequence present in each vector). The DNA sequence of each HSV-tk gene was confirmed by DNA sequence analysis, and the HSV-tk phenotype of each construct was confirmed by

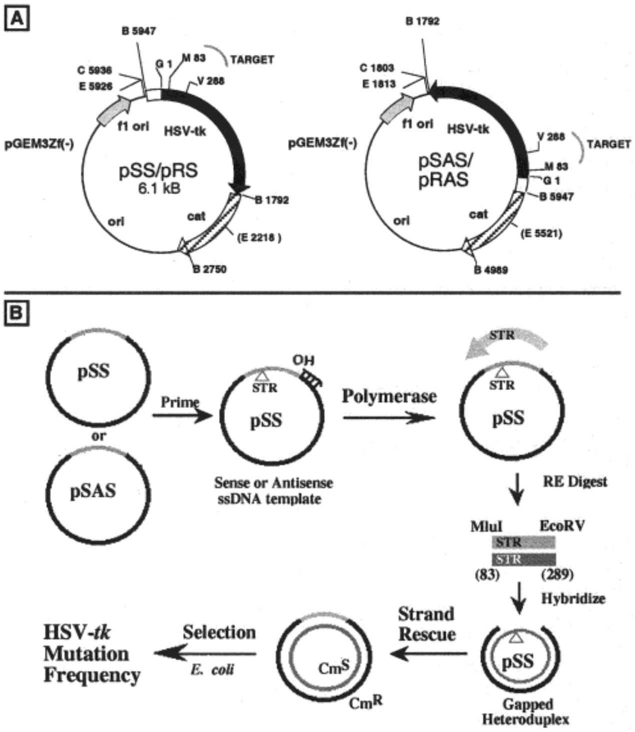

FIGURE 1: (A) Vectors for production of ssDNA. The base vector used for all constructs is the pGEM3Zf(−). Numbering for all vectors begins with the unique *Bgl*II site within the HSV-tk gene. The HSV-tk gene and 3′ noncoding sequences (black arrow) driven by the tetracycline promoter (open bar) were cloned in two orientations, such that the HSV-tk sense strand is produced from the f_1 origin in the SS vectors, while the antisense strand is produced in the SAS vectors. Two versions of each type of construct were produced, one of which encodes a functional cat gene (pRS or pRAS) and one of which encodes a nonfunctional cat gene (pSS or pSAS). The sites of various restriction enzymes are indicated: B, *Bam*HI; C, *Sac*I; E, *Eco*RI; G, *Bgl*II; M, *Mlu*I; V, *Eco*RV. (B) Schematic of the in vitro assay. The STR sequences (shown as an inverted triangle) are located within the *Mlu*I to *Eco*RV fragment (dark gray) of the HSV-tk gene. Templates for DNA synthesis are produced from ssDNA and oligonucleotides. The products of the DNA polymerase reactions (light gray arrow) are digested with *Mlu*I and *Eco*RV restriction enzymes. The polymerase-synthesized strand is rescued by hybridization to a gapped duplex molecule and introduced into *E. coli*. Chloramphenicol selects for the heteroduplex Cm^R strand containing the polymerase synthesis product. The HSV-tk mutation frequency is determined by selection using FUdR.

[1] Abbreviations: Cm, chloramphenicol; FUdR, 5-fluoro-2′-deoxyuridine; GD, gapped duplex; HSV-tk, herpes simplex virus type 1 thymidine kinase; polβ, DNA polymerase β; STR, short tandem repeat.

Table 1: Observed Polymerase β and Background HSV-tk Mutation Frequencies

vector/ microsatellite	HSV-tk mutation frequency × 10^{-4}			
	ssDNA[a]	dsDNA	gapped heteroduplex[b]	polβ synthesis
sense constructs				
pSS1/none	0.62	0.59	2.5 (1.3)	54 ± 9.5 (4)[c]
pSS2/$[GT]_{10}$	1.4	2.1	7.3	55 ± 9.5 (4)
pSS4/$[TC]_{11}$	1.6	2.2	9.4	78 ± 2.5 (4)
pSS5.1/$[TTCC]_9$	10	18	21	150 ± 12 (4)
antisense constructs				
pSAS1/none	1.7	2.2	3.2 (1.8)	98 (2)
pSAS2/$[CA]_{10}$	0.53	4.7	30	120 (2)
pSAS4/$[AG]_{11}$	1.5	5.9	31	200 ± 10 (3)
pSAS5.1/$[AAGG]_9$	5.1	42	55	140 (2)

[a] Mean of two or three independent determinations. [b] GD electroporated into FT334. Number in parentheses is the frequency of mutations occurring within the HSV-tk duplex region of the molecule, as determined by DNA sequence analyses of 41−46 independent mutants. [c] Mean mutation frequency ± standard deviation (number of independent determinations in parentheses).

growth of plasmid-bearing bacteria on selective media (see below). Single-stranded DNA was prepared for each construct by R408 helper phage infection of a plasmid-bearing, F′ *E. coli* strain DH5αIQ.

In Vitro Polymerase Reactions. DNA synthesis templates were created by hybridization of oligonucleotides TK-282 or TK-MluR to pSS or pSAS ssDNAs, respectively, at a 1:1 molar ratio. Priming of these oligonucleotides initiates synthesis at HSV-tk positions 282 (pSS vectors) and 90 (pSAS vectors). The in vitro reactions contained 2 pmol of template DNA at 40 nM concentration. Reaction conditions for polβ were 50 mM Tris-HCl (pH 8.5), 50 mM NaCl, 10 mM $MgCl_2$, 1 mM dithiothreitol, 1 mM dNTPs, 200 μg/mL BSA, and 10 pmol of enzyme/pmol of template DNA. All reactions were incubated at 37 °C for 60 min and terminated with 15 mM EDTA. The extent of DNA synthesis was determined by parallel reactions (0.2 pmol of DNA, same molar ratios of enzyme to substrate as above) supplemented with 5 μCi of [α-^{32}P]dCTP (3000 Ci/mmol). The DNA products of these reactions were analyzed on an 8% denaturing polyacrylamide gel to ensure that DNA synthesis had proceeded past the mutational target.

Mutational Analysis of in Vitro DNA Reaction Products. Linear DNA fragments and ssDNA were prepared as described (*36*) and used to construct gapped duplex (GD) molecules. Hybridization of a linear DNA fragment from pRS1 to ssDNA from pSS1, pSS2, etc. at a 1:1 molar ratio created a single-stranded region complementary to the small fragment purified from the sense strand polymerase reactions (Figure 1B). Analogous GD constructs were prepared to rescue the small fragment from antisense polymerase reactions, using linear DNA prepared from pRAS1 and ssDNA from pSAS1, pSAS2, etc. The GD was separated from other DNA forms by preparative agarose gel electrophoresis and purified using silica, according to the manufacturer's instructions. The background mutation frequencies for each GD preparation were determined by electroporation of *E. coli* with 10 ng of GD, followed by selective plating and DNA sequence analyses, as described below.

To sample the polymerase reaction products for mutations, small fragments were prepared by *Mlu*I and *Eco*RV restriction digestion and hybridized to the corresponding GD as described (*36*), thus forming heteroduplex plasmid molecules (see Figure 1B). An aliquot (~30 ng) of the hybridization was removed for mutational analysis. For all polymerase reactions reported, successful hybridization to GD was achieved, as determined by agarose gel electrophoresis.

To select for HSV-tk mutations, the aliquoted DNA from the final hybridization was used to transform *rec*A13, *upp*, *tdk E. coli* strain FT334 by electroporation and plated on VBA selective media as previously described (*36*). To select for HSV-tk mutant plasmids, the bacteria are plated in the presence of 40 μM 5-fluoro-2′-deoxyuridine (FUdR). FUdR is cytotoxic upon conversion by the plasmid-encoded HSV thymidine kinase enzyme to the nucleotide 5-fluoro-2′-deoxyuridylate, an irreversible inhibitor of thymidylate synthetase. Thus, FT334 bacteria bearing wild-type HSV-tk plasmids catalyze the production of the inhibitor species, 5-fluoro-2′-deoxyuridylate, resulting in thymine-less death of the cells during growth in minimal medium. Because strain FT334 is resistant to the toxic effects of FUdR, bacteria bearing HSV-tk-deficient plasmids survive the FUdR selection. The presence of 50 μg/mL chloramphenicol selects progeny of the polymerase-synthesized DNA strand, as the inner strand of the GD is made from Cm^s vectors. Therefore, the HSV-tk mutant frequency is defined as the number of FUdR-resistant + Cm^R colonies divided by the total number Cm^R colonies. The background mutation frequency for each ssDNA and pSS or pSAS duplex DNA was determined by electroporation of FT334, followed by selective plating on media containing 250 μg/mL carbenicillin in place of chloramphenicol, with or without FUdR.

Independent mutants for DNA sequence analyses were isolated as described (*36*) from one polymerase reaction per template. The DNA sequence of the HSV-tk gene in the *Mlu*I−*Eco*RV region of each mutant was determined by dideoxy DNA sequence analysis. Differences in proportions of specific types of mutations were analyzed statistically using Fisher's exact test (two tailed).

RESULTS

The abundance of microsatellites of varying sequence and their ability to form distinct DNA structures prompted us to analyze DNA polymerase errors within various di- and tetranucleotide microsatellites. To elucidate the protein−DNA interactions relevant for microsatellite mutation, we have developed an in vitro DNA polymerase microsatellite mutation assay for the quantitative analysis of error frequency and specificity.

Design of the DNA Polymerase Microsatellite Mutation Assay. In our experimental system, all mutations are analyzed using DNA vectors that encode an HSV-tk gene containing an artificial, in-frame microsatellite sequence (STR, short tandem repeat). The design of the in vitro STR mutagenesis assay (Figure 1B) follows that previously described to analyze damage-induced DNA polymerase errors (*36*). Forward mutations that inactivate the HSV-tk protein are scored after transfection of *E. coli* and plating in the presence or absence of 5-fluoro-2′-deoxyuridine (FUdR). Mutations which add or delete any number of repeat units within the STR that are not a multiple of three will result in a frameshift mutation. Therefore, mutations occurring in either the repetitive or the HSV-tk coding sequence motif will produce an inactive thymidine kinase protein that is detectable by the same selection scheme.

Two sources of background mutation are present in the design of the in vitro STR assay (Figure 1B). To control for preexisting mutations present within the DNA synthesis template, we determined the HSV-tk mutation frequency for each ssDNA. The mean HSV-tk mutation frequency observed for the control and microsatellite-containing ssDNA templates ranges from 0.53×10^{-4} to 10×10^{-4}, values similar to or lower than those observed for the corresponding dsDNA forms (Table 1). The differences may result from the different biochemical modes of DNA replication for the two forms: rolling circle replication (ssDNA) and origin-dependent, unidirectional replication (dsDNA). We assume that all mutations within the ssDNA occur within the artificial microsatellite allele, similar to what has been observed for the dsDNA forms (*38*). We also controlled for inactivating HSV-tk mutations present within the duplex portion of the GD molecule that is used to rescue the polymerase-synthesized DNA fragment. The observed mutation frequency for unhybridized GD DNA is substantially greater than either the ssDNA or dsDNA forms (Table 1). This increase is likely due to both inaccurate gap-filling DNA synthesis in *E. coli* after transfection and DNA damage produced during the biochemical manipulations necessary to form the GD. DNA sequence analyses revealed that 24% and 34% of the mutations arising after transfection of the pSS1 and pSAS1 GD molecules, respectively, were due to deletion of the entire *Mlu*I−*Eco*RV target region. The number of mutations arising within the duplex region of the GD molecule (outside of the mutational target) was used to determine the GD background mutation frequency, which was calculated to be $(1.3-1.8) \times 10^{-4}$ for each type of construct (Table 1).

We observed complete DNA synthesis by polβ through the mutational target region for all eight templates under our standard reaction conditions (data not shown). The observed mean HSV-tk mutation frequency for all polβ reactions was elevated 15-fold to greater than 100-fold, relative to the corresponding template ssDNA mutation frequency (Table 1). For each template, the polymerase error frequency was calculated by subtracting both the ssDNA and GD background mutation frequencies from the observed polβ HSV-tk mutation frequencies.

Polβ-Mediated Errors in Microsatellite Sequences. To determine the polymerase error frequency within each target region, a mutational spectrum was generated for polβ using each of the eight templates. For all six STR-containing templates, 50% or less of the mutants produced during polβ DNA synthesis arose within the microsatellite sequence. The polymerase error frequency of the six microsatellite sequences examined ranged from 19×10^{-4} to 90×10^{-4} and differed by less than 3-fold from the coding region error frequencies (Table 2).

The polβ error frequencies within the dinucleotide STR sequences (Table 2) were 14−72-fold greater than the corresponding ssDNA template mutation frequencies (Table 1), whereas only a 4−13-fold increase over ssDNA background was observed at the tetranucleotide STR sequences. Mutations within each microsatellite included the expected gain or loss of one or two repeat units (Table 3), as well as noncanonical base substitution events and single nucleotide deletions (Table 4). The noncanonical STR mutations were more prevalent within the sense vectors and primarily involved template thymine residues.

Table 2: Polymerase β Error Frequencies in Coding and Microsatellite Sequences

	polymerase EF × 10^{-4} (no. obsd)[a]	
allele	microsatellite	coding
none (sense)		46 (34)
$[GT]_{10}$	19 (16)	31 (27)
$[TC]_{11}$	29 (28)	42 (41)
$[TTCC]_9$	42 (22)	110 (56)
none (antisense)		84 (32)
$[CA]_{10}$	38 (8)	82 (17)
$[AG]_{11}$	90 (32)	110 (39)
$[AAGG]_9$	65 (31)	65 (31)

[a] Polymerase error frequency = proportion of mutants [MF observed − ssDNA MF − gap duplex MF] (see Table 1).

A pronounced strand bias was observed for the $[TC/AG]_{11}$ allele. The frequency of polβ errors produced using the $[AG]_{11}$-containing template was 4-fold greater than that observed for the $[TC]_{11}$ template (Table 3), a difference that is not observed in the starting ssDNA templates (Table 1). Smaller biases were also observed for the $[GT/CA]_{10}$ and $[TTCC/AAGG]_9$ allelic pairs in which the polymerase error frequency was higher for the adenine-containing template strand (Table 3).

Strand biases in mutational specificity also were observed for the $[TC/AG]_{11}$ and $[TTCC/AAGG]_9$ alleles (Table 3). When the $[TC]_{11}$ sequence was the template for DNA synthesis, 40% of the unit length errors were expansion mutations; in contrast, when the $[AG]_{11}$ sequence was the template for synthesis, only 3% of the unit length errors were expansions, a significant difference ($p = 0.0013$; Fisher's exact test, two sided). For the $[TC/AG]_{11}$ and $[TTCC/AAGG]_9$ alleles, the absolute frequency of microsatellite deletion mutations was 6- and 3-fold higher, respectively, when the purine strand served as the template for DNA synthesis, relative to the pyrimidine strand.

Polβ-Mediated Misalignment Errors in Coding Sequences. An extensive literature exists describing DNA polymerase frame-shift fidelity in homopolymeric sequences (reviewed in ref *27*). In our experimental design, the HSV-tk coding region serves as an internal control for the microsatellite region, as inactivating mutations in both motifs are quantitated using the same selection scheme. Thus, we analyzed the coding region frame-shift mutations for comparative analyses. The mean polymerase error frequency in the HSV-tk coding regions among vectors is $(57 \pm 36) \times 10^{-4}$ for the sense strand and $(85 \pm 19) \times 10^{-4}$ for the antisense strand (Table 2), 92-fold and 50-fold greater than the pSS1 and pSAS1 ssDNA template frequencies, respectively. The coding region mutation frequency for five STR-containing templates varied less than 2-fold from the no STR (HSV-tk gene only) control template. The one exception occurred for the $[TTCC]_9$ template, in which the coding region frequency was 2.4-fold higher than the control (Table 2). However, an in-depth comparison of the types of coding region mutations produced on the $[TTCC]_9$ template versus the other three sense templates failed to uncover any significant differences in mutational specificity. Furthermore, as no significant differences in the proportion of base substitution versus

Table 3: Polymerase β Error Specificity and Error Frequency in Microsatellite Sequences

	no. of independent errors[a]					
	sense template			antisense template		
type of mutation	$[GT]_{10}$	$[TC]_{11}$	$[TTCC]_9$	$[CA]_{10}$	$[AG]_{11}$	$[AAGG]_9$
expansion	2 (0.20)	8 (0.40)	6 (0.43)	4 (0.50)	1 (0.03)	10 (0.32)
1 unit	2	8	4	4	0	8
2 units	0	0	2	0	1	2
deletion	8 (0.80)	12 (0.60)	8 (0.57)	4 (0.50)	30 (0.97)	21 (0.68)
1 unit	7	11	8	2	23	21
2 units	1	1	0	2	7	0
pol EF[b]	12×10^{-4}	21×10^{-4}	27×10^{-4}	38×10^{-4}	87×10^{-4}	65×10^{-4}

[a] Number in parentheses is the proportion of the specific error type among all canonical microsatellite errors. [b] Polymerase error frequency for canonical microsatellite errors.

Table 4: Noncanonical Microsatellite Mutations Produced by Polymerase β

wild-type allele	mutant allele (no. of independent events)	mutational event[a]
gc $[GT]_{10}$ tc	gc $[GT]_5$ T $[GT]_4$ tc (1)	ΔG_{STR6}
	gc $[GT]_9$ G tc (2)	ΔT_{STR10}
	gc $[GT]_4$ TT $[GT]_4$ tc (1)	ΔGT, $G_6 \rightarrow T$
	gc $[GT]_8$ TT tc (2)	ΔGT, $G_{10} \rightarrow T$
gcgt $[TC]_{11}$ ga	gcgt C $[TC]_{10}$ ga (2)	ΔT_{STR1}
	gcgt T $[TC]_{10}$ ga (1)	ΔC_{STR1}
	gcgt TC C $[TC]_9$ ga (1)	ΔT_{STR2}
	gcgt $[TC]_3$ C $[TC]_7$ ga (1)	ΔT_{STR4}
	gc $[TC]_{11}$ ga (1)	$\Delta gt_{110-111}$
	gcg C $[TC]_{10}$ ga (1)	$\Delta T_{111}, \Delta T_{STR1}$
	gcgtgc $[TC]_{12}$ ga (1)	$+gc_{111}$, +TC unit
gt $[TTCC]_9$ tc	gt TCC $[TTCC]_8$ tc (6)	ΔT_{STR1}
	gt $[TTCC]_2$ TCC $[TTCC]_6$ tc (1)	ΔT_{STR3}
	gt $[TTCC]_6$ TCC $[TTCC]_2$ tc (1)	ΔT_{STR7}
cgca $[AG]_{11}$ ct	cgca $[AG]_3$ GG $[AG]_6$ ct (1)	ΔAG, $A_5 \rightarrow G$

[a] One-base deletion, Δ.

Table 5: Polymerase β Misalignment-Mediated Mutation Specificity in the HSV-tk Coding Sequence

type of mutation	sequence motif	no. obsd[a]	pol EF $\times 10^{-4}$
dislocation	T → G at 244	27	19
frame shifts[b]	expansion	13	3.3
(2−4 units)	deletion	109	28
frame shifts[c]	purine (AA, GG)	3 (2)	0.77
2 nt sequence	pyrimidine (TT,CC)	36 (16)	9.2
3 nt sequence	purine (AAA, GGG)	3 (3)	0.77
	pyrimidine (TTT, CCC)	31 (6)	8.0
4 nt sequence	purine (AAAA, GGGG)	17 (3)	4.4
	pyrimidine (TTTT, CCCC)	21 (4)	5.4

[a] Number in parentheses indicates number of sites. [b] Mutations involving one or two bases within 2 nt, 3 nt, or 4 nt homopolymeric sequences. [c] One- or two-base deletion or expansion mutations within the designated sequence.

frame-shift mutations arising within the coding sequences were observed among any of the vectors, the mutations within the coding regions are presented as combined spectra (Figure 2).

The majority of polβ frame-shift mutations occurred within repetitive sequence motifs (Figure 2), consistent with previous observations (*36*, *39*). To determine the polβ error frequencies within the endogenous HSV-tk repetitive sequences, the data for both the sense and antisense HSV-tk mutational spectra were combined so as to minimize the contribution of mutational hotspots. Within the HSV-tk gene, an overall 8:1 bias in favor of frame-shift deletion events over addition errors is observed within mononucleotide repeat sequences two to four units in length (Table 5), consistent with previous observations at the lacZ target (*39*). A strong bias in favor of frame-shift mutations within pyrimidine sequences over purine sequences is also apparent within short (two or three) mononucleotide repeats, but this bias is eliminated as the length of the repeat increased to four nucleotides. A base substitution mutational hotspot (T → G at 244) is present within the antisense template sequence (Figure 2) with an estimated polymerase error frequency of 1.9×10^{-3}. These errors can be explained by a misalignment-mediated dislocation model (*26*), in which slippage of the template strand occurs within the TTTT sequence at positions 241−244. Consistent with this model, several one-base deletion errors also were observed within the TTTT motif.

Testing the Strand Slippage Model for Microsatellite Mutagenesis. Previously, a detailed examination of polymerase error rate as a function of the length of a homopolymeric sequence demonstrated that the error rate increases as the length of the repeat increases (*40*). These results were interpreted as being consistent with the strand slippage, DNA misalignment model for frame-shift mutagenesis. To determine whether polymerase errors produced within microsatellites of the sequence and length found in the human genome are produced by a similar mechanism, we analyzed polβ error frequency as a function of the number of units within repetitive sequences. For this analysis, the polymerase error frequency per site was calculated for each type of repetitive motif present in the coding and microsatellite regions. As noted above, a strong mutational bias for template pyrimidine over template purine sequences is present for mononucleotide repeats of two or three units in length. Interestingly, this relationship is reversed for both the $[TC/AG]_{11}$ and $[TTCC/AAGG]_9$ microsatellite pairs (Figure 3); the polβ error frequency was 2−4-fold greater for the template purine microsatellites, relative to the template pyrimidine microsatellites. An exponential relationship was observed when the error frequencies per site were plotted as a function of the number of units comprising the motif (Figure 3A). The relationship was strongest between the mononucleotide (2, 3, and 4 units) and dinucleotide (10 and 11 units) repetitive motifs ($r = 0.98$). This relationship is consistent with a previous report for homopolymeric sequences in which over a 400-fold increase in the exonuclease-proficient T7 DNA polymerase deletion error rate was measured between a $(T)_3$ and a $(T)_8$ template sequence (*40*). An analysis of the polymerase error frequency per site as a function of the total

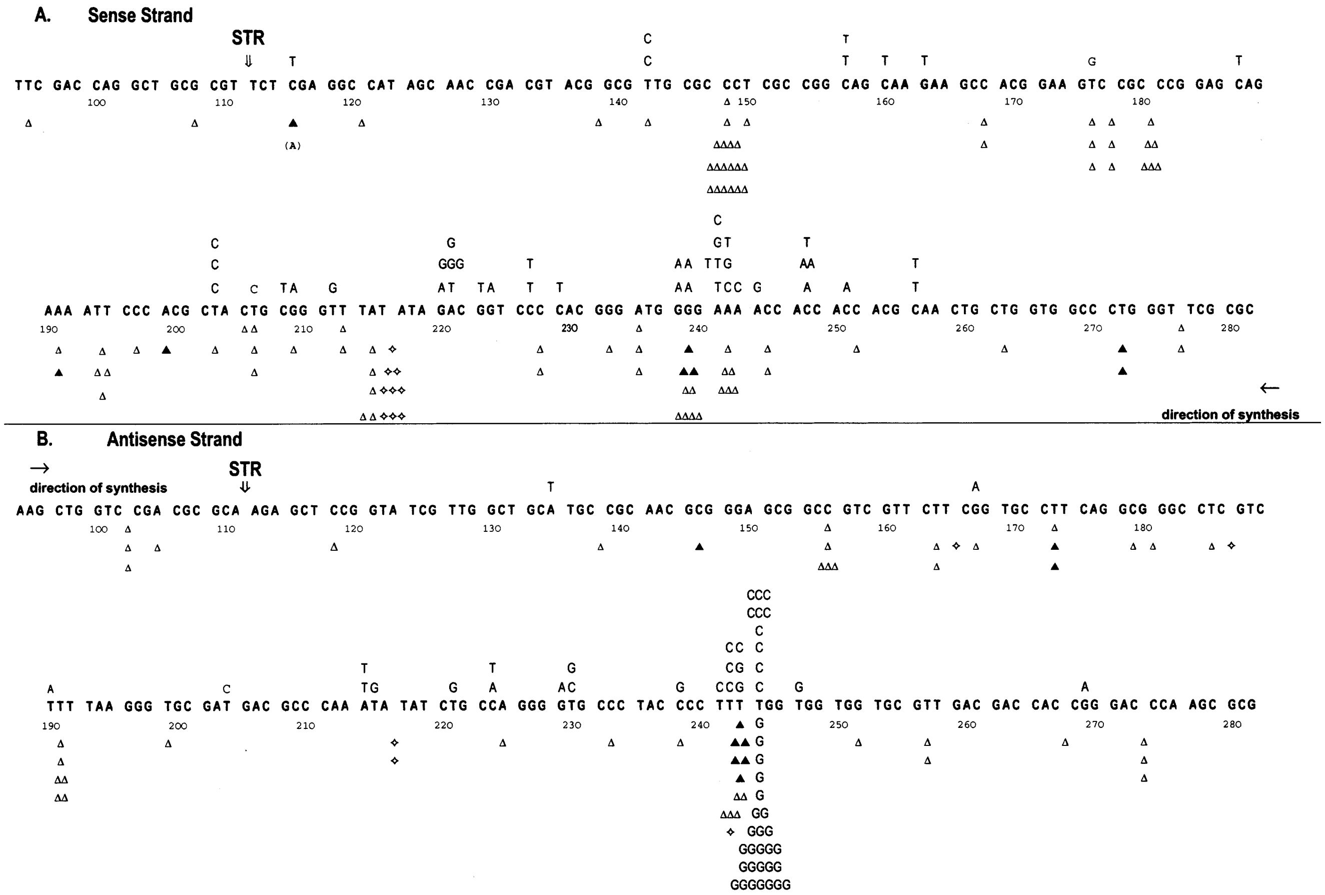

FIGURE 2: Mutational spectra for polβ errors within the HSV-tk coding region. (A) Mutations produced when the sense strand is the DNA template. DNA synthesis is initiated at position 282 and proceeds right to left through the indicated sequence. (B) Mutations produced when the antisense strand is the DNA template. DNA synthesis is initiated at position 90 and proceeds left to right through the indicated sequence. The middle line of each sequence is the wild-type HSV-tk sequence. Base substitution mutations are indicated above the sequence, and misalignment-mediated errors are indicated below the sequence. Base substitutions at position 244 that are postulated to occur by a dislocation mechanism are depicted below the line, as these are initiated by a DNA misalignment (*26*). Each symbol represents one independent mutant: Δ, one-base deletion; ◇, two-base deletion; ▲, one-base addition. Microsatellite sequences were inserted between nucleotides 111 and 112.

length of the repetitive sequence also revealed an exponential relationship between the mononucleotide and dinucleotide repeats, up to 22 nucleotides in length (Figure 3B; $r = 0.97$). Both tetranucleotide motifs (allele length 36) were outliers to this relationship; in this analysis, the observed polymerase error frequency was approximately 20-fold lower than predicted by the exponential. This low polymerase error frequency for tetranucleotide alleles is consistent with the observed low magnitude of increase in HSV-tk mutation frequency, relative to the ssDNA background, for these templates (Tables 1 and 2).

DISCUSSION

To elucidate the molecular basis of microsatellite mutation, we have determined the in vitro error frequencies for DNA polymerase β at sequences representative of those found in the human genome: $[GT/CA]_{10}$, $[TC/AG]_{11}$, and $[TTCC/AAGG]_9$. Templates for DNA synthesis were constructed containing the microsatellites inserted in-frame into the 5′ region of the HSV-tk mutational target. Our assay directly compared the frequency and specificity of errors occurring within the microsatellite sequences to frame-shift errors occurring within the HSV-tk coding sequence. The polβ error frequencies within the microsatellites examined ranged from 2×10^{-3} to 9×10^{-3} (Table 2). Alterations in microsatellites have been proposed to occur by slipped strand mispairing within the repetitive DNA sequences (*13*). Consistent with this model, analysis of polβ-induced frame-shift and microsatellite mutations revealed an exponential relationship between the polymerase error frequency per site and the number of units within the repeated sequences (Figure 3). These in vitro results are comparable to previous in vivo studies of spontaneous frame-shift mutation frequency and number of units in *E. coli* (*41*) and of allele length versus [GT/CA] microsatellite mutagenesis in mismatch repair-deficient *Saccharomyces cerevisiae* (*42*).

The determinants of DNA polymerase frame-shift fidelity have been discussed in a recent review (*27*). As applied to DNA polymerases, the DNA strand slippage model predicts that the error frequency will increase as a function of the number of repeat units due to stabilization of the premutational intermediates by surrounding correct base pairs. A corollary hypothesis (*27*) is that increasing the length of the repetitive allele increases the physical distance between the bulged nucleotides and the 3′-OH of the nascent strand, thereby increasing the efficiency of continued DNA synthesis from the premutational intermediate. Thus, polymerase error frequencies for misalignment mutations are a product of the probability of forming premutational intermediates and the efficiency of utilizing these intermediates as substrates (see

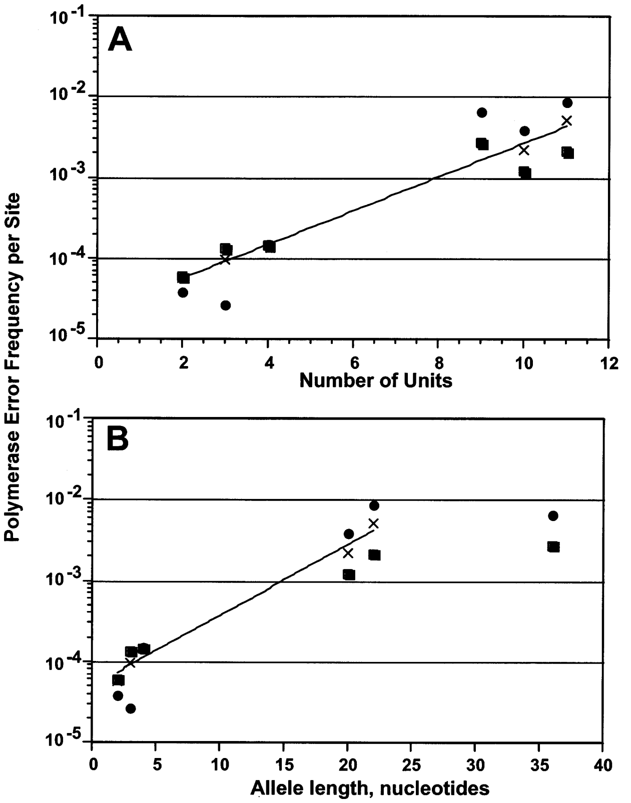

FIGURE 3: Mathematical relationships between polβ error frequency and composition of the repetitive sequence. (A) Number of repetitive units. (B) Total length of repetitive allele. Key: filled circles, error frequencies for template purine sequences (9 units = [AAGG], 11 units = [AG]); filled squares, error frequencies for template pyrimidine sequences (9 units = [TTCC], 11 units = [TC]). Filled square for 10 units is the observed [GT] frequency; filled circle for 10 units is the observed [CA] frequency. Solid line (X) is data for the combined error frequency of both strands at indicated unit length or allele length, fitted to an exponential curve.

Figure 4). As expected for the strand slippage model, we measured an exponential relationship between polymerase error frequency and number of repetitive units which held for di- and tetranucleotide microsatellites as well as homopolymeric coding sequences (Figure 3A). Because several polβ binding events were required to complete DNA synthesis through the microsatellites and because misalignments may be initiated during the polymerase association/dissociation phases of DNA synthesis (*27*), an examination of other DNA polymerases is warranted before the generality of this finding can be established. We have observed a similar exponential relationship between errors produced by DNA polymerase α-primase within the HSV-tk homopolymeric sequences and polypyrimidine/polypurine microsatellites.[2]

Examination of polymerase errors in homopolymeric sequences has demonstrated that factors in addition to length of the repeated sequence affect frame-shift fidelity (*27*). Our data for polβ suggest that additional factors also contribute to polymerase microsatellite errors. For example, the precise sequence of the microsatellite will determine the potential to form premutational intermediates other than strand misalignments during DNA synthesis. A precedent for this in the literature exists for certain trinucleotide microsatellite alleles that can adopt hairpin loop and tetraplex structures (*43*). During our analyses of polβ errors, we measured a high frequency of deletion mutations within purine microsatellites (AG_{11} and $AAGG_9$). These deletions may arise, in part, by a mechanism analogous to that proposed to explain the production of deletion mutations in quasi-palindromic sequences (*44*). Microsatellite purines in ssDNA templates may form internal loop structures similar to internal structural motifs seen in RNA, in which loops are stabilized by cross-strand purine−purine stacking interactions and non-Watson−Crick base pairing (*45*). Such internal loops may be formed within the ssDNA 5′ to the DNA polymerase, such that DNA synthesis proceeds across the base of the loop, resulting in deletion errors (Figure 4). A minimum of four purines may be required for this mechanism, as we did not observe a high frequency of deletions within homopolymeric purine sequences two or three nucleotides in length (Table 5).

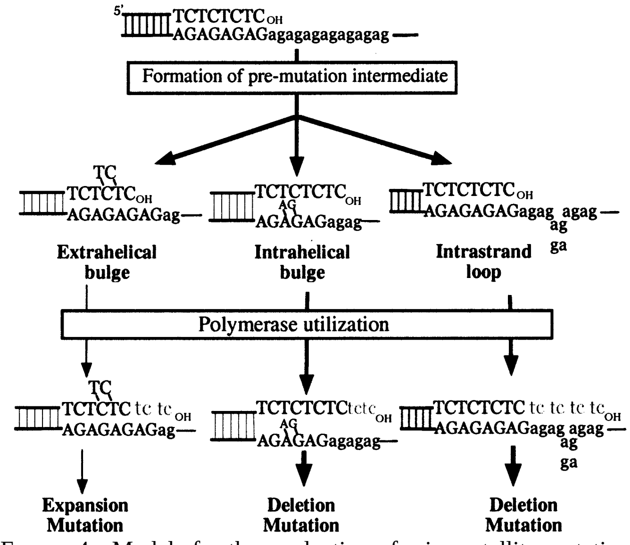

FIGURE 4: Models for the production of microsatellite mutations in vitro. The $[AG]_{11}$ template sequence is representative of the microsatellites examined in this study. Uppercase letters indicate DNA synthesis occurring prior to the mutational event, while lowercase letters indicate ssDNA template sequences. DNA synthesis resulting from the use of premutational intermediates as DNA substrates is indicated in lowercase gray lettering. The extent of polymerase utilization is indicated by the intensity of the arrows.

The physical structure of misaligned, premutational intermediates is a second factor that may lead to mutational biases (Figure 4). Solution studies of DNA duplexes containing bulged nucleotides have revealed that single purine bulges adopt intrahelical positions with minimal disruption of flanking Watson−Crick base pairing, whereas single pyrimidine bulges adopt extrahelical structures when flanked by pyrimidines (reviewed in ref *46*). Bulges of two or three adenines retain this intrahelical structure, and displacement of bases opposite the bulge results in DNA bending (*28*). We have observed that the frequency of polβ deletion mutations within the $[TC/AG]_{11}$ microsatellite is 6-fold greater when the purine strand served as the template for DNA synthesis (8.4×10^{-3}), relative to the pyrimidine template strand (1.3×10^{-3}). This bias may reflect the more efficient utilization by polβ of intrahelical purine bulges than extrahelical pyrimidine bulges (Figure 4). This model predicts that a difference in expansion mutation frequency should be observed when the purine strand is the nascent strand (e.g., the pyrimidine strand is the template). Consistent with this model, the frequency of expansion mutations produced using

[2] S. E. Hile and K. A. Eckert, manuscript in preparation.

the $[TC]_{11}$ template (8.4×10^{-4}, AG as the nascent strand) was 3-fold higher than that using the $[AG]_{11}$ template (2.6×10^{-4}, TC as the nascent strand). Interestingly, this bias appears to be specific for microsatellite sequences, as the polβ error frequency for either deletion or expansion mutations did not vary between template purine ($[A]_4$ or $[G]_4$) and template pyrimidine ($[T]_4$ or $[C]_4$) homopolymeric sequences within the coding region. We note that these two structural factors, formation of premutational intermediates and polymerase utilization, are not mutually exclusive, and both may contribute to microsatellite mutagenesis.

Our analysis of polymerase error frequency versus total length of the repetitive allele revealed an exponential relationship that was valid up to an allele length of ~20 nucleotides (Figure 3B). As explained above, increasing the length of the microsatellite may increase the distance between the bulged nucleotides and the primer terminus, thereby increasing the efficiency of polymerase extension. However, this component of microsatellite error frequency is relevant only over a finite distance dictated by the polymerase−DNA binding interaction. A polβ−ssDNA complex that includes the 31 kDa catalytic domain has been estimated to occlude 16 nucleotides (*47*). Therefore, the 36-nucleotide allele length data may deviate from the exponential relationship (Figure 3B) because the structural perturbation caused by the bulged nucleotides is too distal from the primer terminus and thus does not affect polymerase binding or substrate utilization. Alternatively, the 36-nucleotide allele deviation may be due to differences between the structures of dinucleotide and tetranucleotide bulged templates, such that the four nucleotide bulges are utilized less efficiently by polβ. However, our analysis of polymerase error frequency versus number of units argues that the tetranucleotide templates (9 units) are utilized at least as efficiently as the dinucleotide templates (10 and 11 units) (Figure 3A). Testing dinucleotide and tetranucleotide templates of varying allele lengths in our system can be used to better resolve the contribution of these factors to microsatellite mutagenesis.

We have observed that approximately 30%−40% of the mutations produced by polβ within the $[GT]_{10}$, $[TC]_{11}$, and $[TTCC]_9$ microsatellites were other than unit-length gains or losses (Table 4). These noncanonical changes include single base substitutions as well as single base deletions and primarily involved dipyrimidine residues either adjacent (5′) to the allele or within the microsatellite. The noncanonical errors occur at a frequency of $(0.71-1.5) \times 10^{-3}$, very similar to that observed within the HSV-tk coding region at dipyrimidine sites (Table 5). In vivo, microsatellites are known to exist as interrupted as well as pure arrays (*31*). Direct sequence analyses of the human genome have revealed that the precise sequence composition of microsatellites is very heterogeneous, ranging from pure arrays of a single repetitive sequence to complex arrays containing three or more types of repetitive units per allele (*31*). Our data suggest a very dynamic scenario for microsatellites in which not only the length but also the homogeneity of the microsatellite is in continual flux.

In conclusion, the concomitant increase in DNA polymerase error rates at homopolymeric sequences in vitro and the length of the reiterated sequence have been used previously as data supporting a misalignment-mediated model of frame-shift mutagenesis (*26*). Our observed polβ error frequencies for di- and tetranucleotide microsatellite sequences are consistent with the misalignment model, assuming an exponential relationship between polymerase error frequency and number of repetitive units. In addition, DNA structural differences are postulated to give rise to the observed strand biases for polymerase microsatellite mutations.

ACKNOWLEDGMENT

We thank Dr. Keith Cheng, Dr. Erin Gestl, and Mandy Maneval for helpful discussions and critical reading of the manuscript. We gratefully acknowledge the generous contributions made to the Jake Gittlen Cancer Research Institute.

REFERENCES

1. Brinton, B. T., Caddle, M. S., and Heintz, N. H. (1991) *J. Biol. Chem. 266*, 5153−5161.
2. Gacy, A. M., Goellner, G. M., Spiro, C., Chen, X., Gupta, G., Bradbury, E. M., Dyer, R. B., Mikesell, M. J., Yao, J. Z., Johnson, A. J., Richter, A., Melancon, S. B., and McMurray, C. T. (1998) *Mol. Cell 1*, 583−593.
3. Kang, S., Ohshima, K., Shimizu, M., Amirhaeri, S., and Wells, R. D. (1995) *J. Biol. Chem. 270*, 27014−27021.
4. Samadashwily, G. M., Raca, G., and Mirkin, S. M. (1997) *Nat. Genet. 17*, 298−304.
5. Wahls, W. P., Wallace, L. J., and Moore, P. D. (1990) *Mol. Cell. Biol. 10*, 785−793.
6. Aoki, T., Koch, K. S., and Leffert, H. L. (1997) *J. Mol. Biol. 267*, 229−236.
7. Meloni, R., Albanese, V., Ravassard, P., Treilhou, F., and Mallet, J. (1998) *Hum. Mol. Genet. 7*, 423−428.
8. Rothenburg, S., Koch-Nolte, F., Rich, A., and Haag, F. (2001) *Proc. Natl. Acad. Sci. U.S.A. 98*, 8985−8990.
9. Richards, R. I., and Sutherland, G. R. (1992) *Cell 70*, 709−712.
10. Pearson, C. E., and Sinden, R. R. (1998) *Curr. Opin. Struct. Biol. 8*, 321−330.
11. Loeb, L. A. (1998) *Adv. Cancer Res.*, 26−56.
12. Sidransky, D. (1997) *Science 278*, 1054−1058.
13. Levinson, G., and Gutman, G. A. (1987) *Mol. Biol. Evol. 4*, 203−221.
14. Sia, E. A., Jinks-Robertson, S., and Petes, T. D. (1997) *Mutat. Res. 383*, 61−70.
15. Richards, R. I., and Sutherland, G. R. (1994) *Nat. Genet. 6*, 114−116.
16. Tautz, D., and Schlotterer, C. (1994) *Curr. Opin. Genet. Dev. 4*, 832−837.
17. Henderson, S. T., and Petes, T. D. (1992) *Mol. Cell. Biol. 12*, 2749−2757.
18. Iyer, R. R., Pluciennik, A., Rosche, W. A., Sinden, R. R., and Wells, R. D. (2000) *J. Biol. Chem. 275*, 2174−2184.
19. Strauss, B. S., Sagher, D., and Acharya, S. (1997) *Nucleic Acids Res. 25*, 806−813.
20. Kroutil, L. C., and Kunkel, T. A. (1999) *Nucleic Acids Res. 27*, 3481−3486.
21. Levinson, G., and Gutman, G. (1987) *Nucleic Acids Res. 15*, 5323−5338.
22. Sia, E. A., Kokoska, R. J., Dominska, M., Greenwell, P., and Petes, T. D. (1997) *Mol. Cell. Biol. 17*, 2851−2858.
23. Strand, M., Prolla, T. A., Liskay, R. M., and Petes, T. D. (1993) *Nature (London) 365*, 274−276.
24. Harfe, B. D., and Jinks-Robertson, S. (2000) *Annu. Rev. Genet. 34*, 359−399.
25. Haber, J. E. (1999) *Trends Biol. Sci. 24*, 271−275.
26. Kunkel, T. A. (1990) *Biochemistry 29*, 8003−8011.
27. Kunkel, T. A., and Bebenek, K. (2000) *Annu. Rev. Biochem. 69*, 497−529.
28. Aboul-ela, F., Murchie, A. I. H., Homans, S. W., and Lilley, D. M. J. (1993) *J. Mol. Biol. 229*, 173−188.
29. Rosen, M. A., Live, D., and Patel, D. J. (1992) *Biochemistry 31*, 4004−4014.
30. Beckmann, J. S., and Weber, J. L. (1992) *Genomics 12*, 627−631.
31. Chambers, G. K., and MacAvoy, E. S. (2000) *Comp. Biochem. Physiol. B 126*, 455−476.

32. Sinden, R. R. (1994) *DNA Structure and Function*, Academic Press, New York.
33. Eckert, K. A., Yan, G., and Hile, S. E. (2002) *Mol. Carcinogen*. (in press).
34. Hile, S. E., Yan, G., and Eckert, K. A. (2000) *Cancer Res. 60*, 1698−1703.
35. Kunkel, T. A. (1986) *J. Biol. Chem. 261*, 13581−13587.
36. Eckert, K. A., Hile, S. E., and Vargo, P. L. (1997) *Nucleic Acids Res. 25*, 1450−1457.
37. Opresko, P. L., Shiman, R., and Eckert, K. A. (2000) *Biochemistry 39*, 11399−11407.
38. Eckert, K. A., and Yan, G. (2000) *Nucleic Acids Res. 28*, 2831−2838.
39. Kunkel, T. A. (1985) *J. Biol. Chem. 260*, 5787−5796.
40. Kroutil, L. C., Register, K., Bebenek, K., and Kunkel, T. A. (1996) *Biochemistry 35*, 1046−1053.
41. Streisinger, G., and Owen, J. (1985) *Genetics 109*, 633−659.
42. Weirdl, M., Dominska, M., and Petes, T. D. (1997) *Genetics 146*, 769−779.
43. Wells, R. D. (1996) *J. Biol. Chem. 271*, 2875−2878.
44. Ripley, L. S. (1990) *Annu. Rev. Genet. 24*, 189−213.
45. Moore, P. B. (1999) *Annu. Rev. Biochem. 68*, 287−300.
46. Joshua-Tor, L., Frolow, F., Appella, E., Hope, H., Rabinovich, D., and Sussman, J. L. (1992) *J. Mol. Biol. 225*, 397−431.
47. Rajendran, S., Jezewska, M. J., and Buljalowski, W. (1998) *J. Biol. Chem. 273*, 31021−31031.

BI025918C