Deletion errors generated during replication of CAG repeats

Lisa C. Kroutil and Thomas A. Kunkel*

Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA

Received May 19, 1999; Revised and Accepted July 6, 1999

ABSTRACT

Triplet repeat sequence instability is associated with hereditary neurological diseases and with certain types of cancer. Here we study one form of this instability, deletion of triplet repeats during replication of template (CAG)ₙ sequences by DNA polymerases. To monitor loss of triplet codons, we inserted (CAG)ₙ and (CAG)₁₇ repeats into the lacZ sequence in M13mp2 and changed one repeat to a TAG codon to yield DNA substrates with colorless plaque phenotypes. Templates containing these inserts within gaps were copied and errors were scored as blue plaque Lac revertants whose DNA was sequenced to determine if loss of the TAG codon resulted from substitutions or deletions. DNA synthesis by either DNA polymerase β or exonuclease-deficient T7 DNA polymerase produced deletions involving loss of from 1 to 8 of 9 or 15 of 17 repeats. Thus, these polymerases utilize misaligned template–primers containing from 3 to 45 extra template strand nucleotides. Deletion frequencies were much higher than substitution frequencies at the TAG codon in certain repeats, indicating that triplet repeats are at high risk for mutation in the absence of error correction. Proofreading-proficient T7 DNA polymerase generated deletions at 2- to 10-fold lower frequencies than did its exonuclease-deficient derivative. This suggests that misaligned triplet repeat sequences are subject to proofreading, but at reduced efficiency compared to editing of single-base mismatches.

INTRODUCTION

Eukaryotic genomes contain large numbers of repetitive sequence elements, including many tandem repeats of one to a few nucleotides (microsatellites). Interest in mutation rates in repetitive sequences has been stimulated by associations between microsatellite instability and human diseases. Microsatellite instability is characteristic of cells lacking DNA mismatch repair, including tumor cells from patients with hereditary non-polyposis colorectal cancer (HNPPC) (1,2) and from a variety of sporadic cancers (3). This form of instability is genome-wide and usually consists of the loss or gain of one or a few repeat units from mono-, di-, tri- and tetranucleotide microsatellites. Microsatellite instability is also associated with numerous hereditary neurological diseases (reviewed in 4). This form of instability is locus-specific. Extensively studied examples involve (CAG), (CGG), (GCC) and (GAA) triplet repeats that are 6–52 repeat units in normal individuals and are expanded to 36 to >1000 repeat units in patients afflicted with these diseases.

Several mechanisms have been considered for the origins of microsatellite instability. The large expansions observed for diseases such as fragile X syndrome, Friedreich's ataxia and myotonic dystrophy may result from recombination or from aberrant processing of replication intermediates involving unusual DNA structures arising in certain repeat sequences (4,5 and references therein). Template–primer slippage during DNA synthesis (6,7) may underlie the loss or gain of a smaller number of repeat units in tumors exhibiting microsatellite instability and possibly in certain hereditary neurological diseases. Strand misalignments could arise during replication or during DNA resynthesis associated with any of several different excision repair processes. Microsatellite instability results if insertion/deletion mismatches remain uncorrected by exonucleolytic proofreading or DNA mismatch repair.

The objective of this study is to investigate the strand slippage hypothesis during copying of a triplet repeat sequence in the absence of DNA mismatch repair and using DNA polymerases that differ in structure, primary physiological function and biochemical properties, including the ability to proofread errors. We focus on the (CAG) triplet repeat that is unstable in Huntington’s disease, Haw River syndrome, spinobulbar muscular atrophy and spinocerebellar ataxia types 1–3, 6 and 7 (4). We developed an in vitro assay to measure deletion errors in (CAG)ₙ and (CAG)₁₇ sequences. These numbers of repeat units are in the range of those at the relevant loci in normal individuals, i.e. 6–41 repeats (4). We examined the fidelity of DNA synthesis by human pol β, whose primary function is in base excision repair, and by T7 DNA polymerase, a replicative enzyme. The results suggest that strand slippage occurs during copying of template (CAG)ₙ sequences by these polymerases, that a substantial number of repeats can be deleted in a single cycle of synthesis and that deletion mismatches can be proofread by T7 pol but with relatively low...
efficiency. The data are compared to base substitution error rates and considered in light of the crystal structures of pol β (8,9) and T7 pol (10).

MATERIALS AND METHODS

Strains, reagents and DNA substrates

*Escherichia coli* strains and reagents have been described (11). To generate DNA templates with CAG repeat inserts in the *LacZ* gene, a M13mp2SVOriL vector (12) was modified by oligonucleotide-directed mutagenesis as described (13). The resulting M13 derivatives (Table 1) were used to prepare gapped DNA substrates as described (11). T7 DNA polymerase and its exonuclease-deficient D5A, E7A derivative, both purified in the absence of *E.coli* thioredoxin, were kind gifts of S. Patel and K. Johnson (14). Recombinant human DNA polymerase β, purified as described (15), was kindly provided by S. Wilson (NIEHS).

DNA synthesis reactions and reversion frequency analysis

Pol β reactions (20 µl) contained 20 mM Tris–HCl (pH 8.0), 2 mM dithiothreitol, 10 mM MgCl₂, 45 fmol gapped DNA, 500 µM dATP, dGTP, dCTP and TTP and 1.2 pmol enzyme. T7 DNA polymerase reactions (25 µl) contained 40 mM Tris–HCl (pH 7.5), 2 mM DTT, 10 mM MgCl₂, 45 fmol gapped DNA, 1000 µM dATP, dGTP, dCTP and TTP and 3.0 (exo⁻) or 9.0 pmol (exo⁺) enzyme. Reactions were incubated at 37°C for 60 (pol β) or 45 (T7 pol) min, then quenched by adding EDTA to a final concentration of 15 mM. All reactions completely filled the gap, as determined by agarose gel electrophoresis performed as described (11). The reversion frequencies of control and copied product DNA samples and DNA sequence analysis of *LacZ* revertants were determined as described (11).

RESULTS

Fidelity assay involving (CAG)$_n$ repeats

As substrates for DNA polymerase reactions, we prepared circular, double-stranded DNA molecules containing single-stranded gaps. Within these gaps were template sequences containing either nine or 17 in-frame CAG (Gln) codons in the N-terminus of the *LacZ* α-complementation gene in bacteriophage M13mp2. In each substrate, one CAG codon was converted to a TAG nonsense codon by introducing a single C→T base substitution. This inactivates α-complementation of β-galactosidase activity and results in a colorless M13 plaque phenotype. Two substrates contained nine repeats, one with the imperfection at the fourth repeat codon to be encountered during synthesis (Table 1, first substrate) and the other with the imperfection at the eighth repeat codon. Three substrates contained 17 repeats, with the imperfection present at the fourth, ninth or sixteenth repeat codon (Table 1). In all cases, correct gap filling synthesis maintains the colorless phenotype, while in-frame deletions or base substitutions that eliminate the TAG codon yield blue plaque revertants. The background reversion frequencies for the DNA templates when introduced into *E.coli* without DNA synthesis in vitro are relatively low (Table 1). Sequence analysis of blue revertants from these control reversion frequency determinations revealed that each *LacZ* revertant contained either a single base substitution at the fourth, ninth or sixteenth repeat codon (Table 1). In all cases, correct gap filling synthesis maintains the colorless phenotype, while in-frame deletions or base substitutions that eliminate the TAG codon yield blue plaque revertants. The background reversion frequencies for the DNA templates when introduced into *E.coli* without DNA synthesis in vitro are relatively low (Table 1). Sequence analysis of blue revertants from these control reversion frequency determinations revealed that each *LacZ* revertant contained either a single base substitution at the TAG codon or an in-frame deletion of one or more triplet repeats, always including loss of the TAG codon.

Fidelity of human pol β

We used pol β for the initial attempt to study triplet repeat deletion fidelity. It is an essential enzyme that has a major role in base excision repair in *vivo* (16). It lacks intrinsic 3'→5' exonuclease activity, such that fidelity reflects polymerase selectivity in the absence of proofreading. Pol β fills gaps of the lengths used here in a distributive manner, dissociating from the

| Table 1. Triplet repeat template sequences used in this study |
|---------------------------------|-----------------|----------------|
| Sequence                      | Designation     | Reversion frequency of DNA substrate |
| (CAG)$_5$ TAG (CAG)$_3$ AAGCTT  | (CAG)$_{1+4}$   | $1.1 \times 10^{-5}$ |
| (CAG)$_3$ TAG (CAG)$_3$ AAGCTT  | (CAG)$_{3+8}$   | $1.2 \times 10^{-5}$ |
| (CAG)$_7$ TAG (CAG)$_3$ AAGCTT  | (CAG)$_{7+4}$   | $1.3 \times 10^{-5}$ |
| (CAG)$_{1,0}$ TAG (CAG)$_3$ AAGCTT | (CAG)$_{17+9}$ | $1.4 \times 10^{-5}$ |
| (CAG)$_{1,0}$ TAG (CAG)$_{14}$ AAGCTT | (CAG)$_{1,16}$ | $1.5 \times 10^{-5}$ |

The triplet repeat template sequences within gapped substrates are shown, listed 5'→3', inserted between nucleotides 59 and 60 (where +1 is the first transcribed base of the gene) in the N-terminus of the *LacZ* α-complementation gene. The 3'- and 5'-termini in the complementary strand of all gaps are nucleotides 146 and ~217, respectively (defined by *Pvu*I digestion; see 11). Thus, the gaps contain ~400 template nucleotides, the repeat sequence is encountered by the polymerase after synthesis of ~100 nt and synthesis across the repeat sequence is completed more than 200 nt before the gap is filled. The reversion frequencies are for single-stranded template DNA introduced into *E.coli* without being subjected to DNA synthesis in vitro and are based on the following blue/colorless plaque counts for each substrate, in the order shown: 4/370,000; 5/400,000; 14/170,000; 7/130,000; 26/130,000. Sequence analysis of DNA extracted from revertants revealed the following numbers of substitutions and in-frame deletions for each substrate, in the order shown: 2 and 2; 5 and 11; 2 and 2; 5 and 11; 2 and 9; 2 and 10; 17 and 7. Sequenced revertants of the (CAG)$_{1+4}$ and (CAG)$_{3+8}$ substrates include a few that were obtained from platings performed in addition to those yielding the plaque counts listed above.
template–primer after each nucleotide incorporation event. Also, pol β has low fidelity for single base deletions within homopolymeric runs (17–19).

DNA synthesis by recombinant human pol β was performed to fill gaps containing the triplet repeat sequences. For four of the five substrates examined, this synthesis generated products having reversion frequencies that were 10- to 30-fold above the control DNA (Table 2). Sequence analysis revealed that 62 of 169 LacZ revertants analyzed had a single base substitution at the TAG codon (see last column of Table 3), yielding an average substitution frequency of $15 \times 10^{-5}$. [When this reversion frequency is used to calculate a single-base substitution error rate per detectable nucleotide polymerized (for calculation see 11), the error rate is $8.3 \times 10^{-4}$. This rate is remarkably similar to previous values obtained for substitutions at other locations in the LacZα gene (19), and confirms that pol β has an average base substitution error of $\sim10^{-4}$ when distributively filling gaps of a few hundred nucleotides.] Ninety-three revertants contained in-frame deletions of one to 15 repeat units (Table 3), always including loss of the TAG codon. When these data are used to calculate deletion frequencies (Fig. 1), the values are 40-, 3-, 10-, 35- and 2-fold greater than the deletion frequencies of the control DNA for the (CAG)9-4, (CAG)9-8, (CAG)17-4, (CAG)17-9 and (CAG)17-16 substrates, respectively. In addition to simple substitutions and deletions, 14 revertants contained sequences (not shown) indicating complex deletion errors (designated 'other' in Table 3).

**Fidelity of T7 DNA polymerase**

We next examined the triplet repeat deletion fidelity of T7 DNA polymerase. It is a replicative polymerase with an intrinsic

### Table 2. Reversion frequencies after copying by DNA polymerases

<table>
<thead>
<tr>
<th>DNA substrate</th>
<th>Plaques</th>
<th>Reversion frequency ($\times 10^{-5}$)</th>
<th>Fold increase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Blue</td>
<td></td>
</tr>
<tr>
<td>Copied by DNA polymerase β</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CAG)9,4</td>
<td>150 000</td>
<td>55</td>
<td>37</td>
</tr>
<tr>
<td>(CAG)9,8</td>
<td>110 000</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>(CAG)17,4</td>
<td>64 000</td>
<td>55</td>
<td>86</td>
</tr>
<tr>
<td>(CAG)17,9</td>
<td>92 000</td>
<td>151</td>
<td>160</td>
</tr>
<tr>
<td>(CAG)17,16</td>
<td>110 000</td>
<td>34</td>
<td>31</td>
</tr>
<tr>
<td>Copied by wild-type T7 DNA polymerase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CAG)17,9</td>
<td>140 000</td>
<td>225</td>
<td>160</td>
</tr>
<tr>
<td>Copied by exonuclease-deficient T7 DNA polymerase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CAG)17,9</td>
<td>110 000</td>
<td>638</td>
<td>580</td>
</tr>
</tbody>
</table>

Reactions and product analyses were performed as described in Materials and Methods. The values in the last column were calculated by comparing the reversion frequencies observed following DNA synthesis to those observed with control DNA (from Table 1).

### Table 3. DNA sequence analysis of revertants

<table>
<thead>
<tr>
<th>Number of revertants that had lost the following number of repeat units</th>
<th>Total deletions</th>
<th>Total substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pol β</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CAG)9,4</td>
<td>5 2 2 2 1 9 3 0 0 – – – – – – – – 4 28</td>
<td>20</td>
</tr>
<tr>
<td>(CAG)9,8</td>
<td>1 0 1 0 0 0 0 0 0 – – – – – – – – 1 3</td>
<td>14</td>
</tr>
<tr>
<td>(CAG)17,4</td>
<td>2 0 1 0 1 1 1 1 1 1 1 1 2 2 2 1 2 0 2</td>
<td>20</td>
</tr>
<tr>
<td>(CAG)17,9</td>
<td>0 0 0 1 1 1 0 3 2 9 5 13 1 2 0 0 5</td>
<td>43</td>
</tr>
<tr>
<td>(CAG)17,16</td>
<td>2 4 0 1 1 0 1 0 0 1 0 1 0 0 0 0 2</td>
<td>13</td>
</tr>
<tr>
<td>Wild-type T7 pol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CAG)17,9</td>
<td>1 0 0 0 0 1 8 7 29 14 22 7 9 2 1 0 0 0</td>
<td>101</td>
</tr>
<tr>
<td>Exo-deficient T7 pol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(CAG)17,9</td>
<td>0 0 0 0 0 0 0 0 0 15 11 30 19 20 4 1 0 0 1</td>
<td>101</td>
</tr>
</tbody>
</table>
3′→5′ exonuclease activity that proofreads base-base mismatches (20) and single base insertion/deletion mismatches (21). The fidelity of the wild-type enzyme can be directly compared to that of a mutant polymerase that lacks exonuclease activity due to amino acid substitutions in the exonuclease active site. The exonuclease active site is physically distant from the polymerase active site (10) and the DNA polymerase activity of the mutant protein is relatively normal. Thus any differences observed in the triplet repeat deletion fidelity of the wild-type and mutant protein may reflect the contribution of exonucleolytic proofreading. Also of interest to deletion fidelity is the processivity of polymerization. Low processivity is associated with low frameshift fidelity (7,22). When complexed with the accessory protein thioredoxin, T7 pol is highly processive and can synthesize >1000 nt without dissociating from the template–primer (23,24). However, even in the absence of thioredoxin, T7 pol synthesizes DNA in a more processive manner than does pol β, perhaps reflecting the more extensive contacts between the polymerase and the duplex template–primer observed in the crystal structure (see for example 25). Having previously shown that the exonuclease-deficient DNA polymerase lacking thioredoxin has low frameshift fidelity when copying homopolymeric runs (21), here we chose to study triplet repeat deletion fidelity with the wild-type and exonuclease-deficient polymerase catalytic subunit in the absence of thioredoxin.

Having seen the highest repeat deletion frequency for pol β with the (CAG)17–9 substrate (Fig. 1), we copied this substrate with T7 pol. The products of gap filling synthesis by the wild-type and exonuclease-deficient enzymes had LacZα reversion frequencies that were 30- and 110-fold higher (Table 2) than the value for the control DNA. The 3.6-fold difference in reversion frequency observed between the wild-type and exonuclease-deficient T7 DNA polymerase (Table 3) suggests that the exonuclease can proofread triplet repeat deletion intermediates.

Sequence analysis of 204 revertants (Table 3) revealed that all but two contained in-frame deletions of 1 to 14 repeat units, always including loss of the TAG codon. The wild-type enzyme most frequently yielded deletions of 6 to 12 triplet repeat units while the exonuclease-deficient polymerase most frequently generated deletions of 8 to 13 triplet repeat units (Table 3 and Fig. 2A). The frequency differences between the two forms of T7 pol for deletions range from only 1.9-fold for loss of eight repeats to 10-fold for loss of 11 repeats (Fig. 2B).

**DISCUSSION**

We show here that DNA polymerases involved in DNA replication and excision repair generate deletions of variable numbers of triplet repeats during a single cycle of DNA synthesis in vitro. This has several implications. Such errors could account for some of the microsatellite instability associated with mismatch repair-defective human tumor cells and perhaps some of the deletion instability associated with hereditary degenerative diseases. Deletion frequencies for substrates copied in vitro are readily observed above control values (Figs 1 and 2A). This is despite the fact that the CAG repeats are imperfect, i.e. they are interrupted by a C→T transition to permit errors to be scored by loss of a TAG codon (Fig. 3A), and the fact that the number of repeat units in these substrates

![Figure 1](image1.png)  
**Figure 1.** Frequencies of in-frame deletions by human pol β. Black bars, control template DNA; gray bars, substrates copied by human pol β. The values were calculated from the reversion frequencies in Tables 1 and 2 and the sequencing data in Table 3 and in the legend to Table 1.

![Figure 2](image2.png)  
**Figure 2.** Reversion frequencies for deletions by T7 pol. (A) Black bars, wild-type T7 pol; gray bars, exonuclease-deficient T7 pol. The values were calculated from the reversion frequencies in Table 2 and the sequencing data in Table 3. (B) The ratio of reversion frequencies, calculated from the data in (A).
The deletion frequencies observed following DNA synthesis by pol β with the five different substrates are highly variable. The fact that deletion frequencies with the (CAG)₉ substrates are generally higher than those with the (CAG)₉ substrates (Fig. 1) is consistent with the strand slippage model first proposed by Streisinger et al. (6). As reviewed in detail elsewhere (7), template–primer slippage within microsatellites yields misaligned intermediates whose number and stability increases with increasing repeat unit number. The deletion frequencies detected here likely provide a minimum estimate of instability, since some deletions will remain undetected because they do not involve loss of the nonsense codon (see for example Fig. 3B). For example, only about half of the hypothetical deletion intermediates in the (CAG)₁₇₋₉ substrate will yield blue plaque revertants (Fig. 3C). An even lower proportion of misaligned intermediates will result in loss of the TAG codon if this codon is closer to the 5’-end of the template, e.g. as with the (CAG)₁₇₋₁₈ substrate. This may explain why synthesis by pol β with the (CAG)₁₇₋₁₈ substrate increased the deletion frequency to a lesser extent than for the (CAG)₁₇₋₄ or (CAG)₁₇₋₉ substrates (Fig. 1).

The deletions that result from synthesis by pol β and T7 pol involve loss of from one to 15 repeat units (Table 3). This clearly shows that large changes in the number of repeat units in a microsatellite can result from only a single round of gap filling DNA synthesis. This is consistent with the possibility that a large number of repeat units could be deleted from (or added to?) a microsatellite during a single excision repair or replication event in vivo. The data in Table 3 further indicate that pol β and T7 pol can bind to and extend misaligned template–primers containing as many as 45 unpaired nucleotides in the template strand. The ability to continue synthesis from misaligned template–primers in triplet repeat sequences is consistent with biochemical and structural data, indicating that polymerases contact the template–primer for only a few nucleotides upstream of the active site (8–10, 25). The exact length of the duplex template–primer stem and the exact location of the nucleotides located only 3–9 bp from the active site. In some cases, the stability of large deletion intermediates may be increased by an inverted repeat in the template (see for example fig. 4–4 in 7). More relevant to this study may be the well-known ability of single-stranded CAG repeats to form hairpins (see for example 26–30). This may contribute to the high deletion frequencies seen with the (CAG)₁₇₋₉ substrate (Fig. 1), since, among all substrates examined, (CAG)₁₇₋₉ has the highest probability of forming hairpins that would contain the TAG codon that must be lost in order to score a deletion error. Hairpin formation may also contribute to the pattern of deletions seen with the (CAG)₁₇₋₉ substrate, where losses of eight to 12 CAG repeats (small hairpins?) predominate (Table 3).

Previous studies have correlated low fidelity for deletions and additions within homopolymeric sequences with low polymerization processivity (reviewed in 7). This led to the hypothesis that misalignments may arise during polymerase dissociation–reassociation with template–primer. Consistent with this, pol β, which has very low processivity when filling
gaps of the sizes used here, is also inaccurate for repeat unit deletions when copying triplet repeat sequences (Fig. 1 and Table 3). Exonuclease-deficient, thioredoxin-deficient T7 pol is even less accurate than is pol β (Table 2), despite the fact that it can incorporate from 1 to 50 nt/reaction cycle (24). A high rate of both deletion and addition errors within homopolymeric sequences has been observed before with this enzyme (21). Such error-prone synthesis likely results from aberrant interactions of the polymerase with the duplex template–primer stem upstream of the active site (10) when T7 pol is missing its processivity subunit.

The results obtained with wild-type and exonuclease-deficient T7 pol (Fig. 2) indicate that misalignments in the (CAG)₁₉₉ sequence associated with loss of up to 13 repeat units are corrected by exonucleolytic proofreading. This suggests a possible role for proofreading in stabilizing triplet repeat sequences in vivo. However, the apparent contribution to stability is <10-fold (Fig. 2B), in contrast to the ≥100-fold contribution of proofreading by T7 pol to base substitution fidelity (20). This difference is predicted on the hypothesis (31) that misalignments will be stabilized by correct base pairs and will therefore partition in favor of continued polymerization rather than exonucleolytic proofreading. The results in Figure 2 are in agreement with the small contribution of proofreading to deletion rates in long repetitive sequences in vitro (31) and in vivo (32–34).

This study reveals a high potential for deletion mutagenesis during copying of CAG repeat sequences by pol β or thioredoxin-free T7 DNA pol. It will be of interest to extend this approach to other repeat sequences and to more processive polymerization reactions, especially those involving additional proteins known to participate in DNA replication and excision repair.

ACKNOWLEDGEMENTS

We thank K. Bebenek and W. Osheroff for critical comments on the manuscript.

REFERENCES