The DNA Polymerase β Replication Error Spectrum in the Adenomatous Polypsis Coli Gene Contains Human Colon Tumor Mutational Hotspots

Brindha P. Munniapan and William G. Thilly

ABSTRACT

We have found a significant concordance between the in vitro replication errors of human DNA polymerase β and in vivo point mutations of the adenomatous polyposis coli (APC) gene that leads to colon cancer. We determined the error spectrum of DNA polymerase β in the human APC gene under PCR conditions and compared it with the set of mutations reported in human colon tumors. Polymerase β created seven hotspot mutations within 141 target bp analyzed in APC exon 15. Three of these polymerase β hotspots, 2 frameshifts and a bp substitution mutation, were concordant with 3 of 13 APC hotspots detected in human colon cancers in the same DNA sequences. These 3 concordant hotspots accounted for some 54% of reported in vivo APC hotspot mutations. Using the assumption of a hypergeometric distribution of hotspot mutations among bp of the scanned sequences, the probability of this concordance occurring by chance is \( < 4 \times 10^{-4} \). These data support the hypothesis that DNA polymerase β errors are an important fraction of cancer-causing APC mutations.

INTRODUCTION

Mutations in humans may be caused by exogenous mutagens (1) or endogenous processes, such as reaction with metabolites (2), DNA hydrolysis (3, 4), errors in DNA synthesis (5), or possibly through repair processes acting on undamaged DNA (6). We set out to discover if a significant fraction of human nuclear point mutations occurring in vivo could be attributed to a particular source. A potential cause of point mutations in a defined DNA sequence can be tested by determining if its point mutational spectrum in vitro is a statistically significant subset of in vivo mutations (7–9). An overlap, or concordance, of hotspots between the two sets of mutations can identify a contribution from a hypothesized mutagenic source to mutations that cause human disease. If there were no contributions of the hypothesized mutagenic source to mutations that occur in vivo, then a degree of concordance between the two mutational sets should not rise above what is expected by chance. The null hypothesis may be tested using the hypergeometric distribution in which it is assumed that mutational hotspots are randomly distributed across the target sequence for both the error spectrum of the hypothesized mutagenic source and the in vivo mutational spectrum.

We selected the APC tumor suppressor gene for our study because a large and diverse set of APC mutations are known to initiate most somatic colon cancers (10, 11). Within 8532 bp of the APC gene coding region, we compiled >700 human colon tumor mutations from 38 separate reports to estimate the relative frequencies and distribution of the most frequently observed APC mutations (12). On the basis of these data, we selected two APC gene sequences comprising 141 bp (3896–3970 and 4355–4420) that contain the most frequently observed APC mutations in human colon tumors. Within these 141 bp, 51 distinct mutations had been recorded in 198 separate tumors. The quantitative distribution of these events, which constitute the APC in vivo mutational spectrum, is recorded in Fig. 1A.

In taking a broad view of the possible sources of in vivo point mutations, we noted that one pathway, DNA polymerization, had not been examined using mutational spectrometry methods to study primary or repair DNA replication errors. DNA polymerases, however, have in vitro replication error rates that could account for a significant fraction of in vivo human mutations (13–15). DNA polymerase β in particular, the smallest known human polymerase, is reported to have a fidelity rate (measured in specific DNA sequences using phenotypic selection systems) of \( 2 \times 10^{-3} \) to \( 5 \times 10^{-3} \) errors/base incorporation (13, 16, 17). In vivo, polymerase β is responsible for filling in gaps of one to six nucleotides after the excision of sequences containing DNA damage (18, 19). Because every day, human cells may repair up to one million DNA lesions involving excision repair and filling of the ensuing gap by polymerase β (20), the summation of repair replication errors created by polymerase β could plausibly account for common somatic nuclear mutations in humans. To test this hypothesis, we sought to compare two quantitative point mutational spectra for the same DNA sequence: (a) the human DNA polymerase β in vitro error spectrum; and (b) the set of APC gene mutations observed in human colon tumors.

MATERIALS AND METHODS

The replication error spectra of any DNA polymerase can be determined by analysis of the mutations created when it is used in DNA amplification (21, 22). Mutant sequences generated during replication can then be physically separated from wild-type sequences and from each other using CDCE (23) or denaturing gradient gel electrophoresis (24). Individual mutants may be directly collected, isolated, and identified by sequencing (25–27).

High-Fidelity PCR Conditions. For this study, the target APC sequences (bp 3896–3970 and 4355–4420) were first made suitable for CDCE analysis by the attachment of a thermostable oligonucleotide (“clamp”) through DNA hydrolysis (28) and denaturing gradient gel electrophoresis (24). Individual mutants may be directly collected, isolated, and identified by sequencing (25–27).

Purification of CDCE-suitable Target Sequences. For this study, the target APC sequences were first made suitable for CDCE analysis by the attachment of a thermostable oligonucleotide (“clamp”) through DNA hydrolysis (28) and denaturing gradient gel electrophoresis (24). Individual mutants may be directly collected, isolated, and identified by sequencing (25–27).

Purification of CDCE-suitable Target Sequences. For this study, the target APC sequences were first made suitable for CDCE analysis by the attachment of a thermostable oligonucleotide (“clamp”) through DNA hydrolysis (28) and denaturing gradient gel electrophoresis (24). Individual mutants may be directly collected, isolated, and identified by sequencing (25–27).

Received 11/27/01; accepted 4/2/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

1 Supported by The National Institute of Environmental Health Sciences through Grants T32-ES0702 (to B. P. M.) and P01-ES07168-06 (to W. G. T.), which provided additional support for this project.

2 To whom requests for reprints should be addressed, at Biological Engineering Division, Massachusetts Institute of Technology, 21 Ames Street, 16-743, Cambridge, MA 02139. Phone: (617) 253-6221; Fax: (617) 258-5424; E-mail: thilly@mit.edu.

3 The abbreviations used are: APC, adenomatous polyposis coli; Pfu, Pyrococcus furiosus; CDCE, constant denaturant capillary electrophoresis.
Fig. 1. Comparison of in vivo colon tumor hotspot mutations and polymerase β replication error hotspots within 141 bp of the human APC gene. Concordant hotspots within APC bp 3896–3970 and bp 4355–4420 are shown in gray; discordant hotspots are in black. Mutational spectra are shown as a function of bp position in the APC mRNA (GenBank accession no. MT4088) and number of independent tumors carrying a particular mutation (A) or mutant fractions (average of duplicate independent trials) after five template doublings (B). ---- in the top panel, the minimum number of observations (2) that define an in vivo hotspot. Fifty-one different kinds of in vivo mutations have been identified within the 141 APC gene bp scanned: 39 mutations were reported once, 7 were reported twice, and 6 other mutations were reported 4, 5, 6, 24, 32, and 80 times. In some instances, ≥2 types of mutations observed in vivo were reported at the same bp. For clarity in depiction, the magnitude of the bars in this figure represent the most frequently occurring mutation at each site that contains >1 type of mutation.

μl) were dialyzed on a 0.25 μm filter (Millipore) against MilliQ water for 30–40 min and then speed vacuumed to reduce volumes to roughly 2 μl. Samples of 1–2 μl were loaded into a 1-cm length of Teflon tubing (inner diameter = 254 μm; outer diameter = 1588 μm) and electroinjected into a capillary (inner diameter = 250 μm; outer diameter = 350 μm) at 40 μlA for 2 min. Wild-type homoduplexes were separated from wild-type:mutant heteroduplexes within a 5-cm denaturation zone at 72°C. Wild-type DNA molecules were collected into 4 μl of collection buffer [45 mM Tris-borate, 0.5 mM EDTA (pH 8), and 0.3 mg/ml BSA] and stored at −20°C.

Amplification with Recombinant Human DNA Polymerase β. Human DNA polymerase β was provided in recombinant form by Drs. S. Wilson, R. Prasad, and W. Copeland at the National Institute of Environmental Health Sciences (28, 29). 107 molecules of CDCE-suitable, purified, wild-type APC templates were used for successive cycles of replication by human polymerase β. The purified double-strand template DNA was mixed with 0.2 μM each primer, 450 μM each deoxynucleotide triphosphate, 20 mM Tris-HCl (pH 8), 7 mM MgCl2, 100 mM NaCl, 1.5 mM DTT, 2% glycerol, and 0.2 mg/ml BSA. Manual PCR cycling in 20-μl reaction volumes consisted of denaturation in boiling water for 30 s, then primer-template annealing at 50°C or 53°C for 5 min. Recombinant human polymerase β in 50 mM Tris-HCl (pH 6.5), 100 mM NaCl, 1 mM EDTA (pH 8), and 20% glycerol, was added, mixed, and incubated at 37°C for 5 min for each cycle. Final polymerase concentrations between 50 and 150 nM were found suitable for amplification of the target APC sequences. 109 product copies were removed and further amplified when greater than eight manual PCR cycles were required. The products of DNA amplification with polymerase β were amplified for 14 additional template doublings with Pfu DNA polymerase. This degree of amplification with Pfu polymerase did not create mutations that interfered with observations of the DNA polymerase β spectrum, as observed in comparison of Fig. 2, A or B with Fig. 2C.

High-Fidelity PCR and CDCE. The mutant sequences created by polymerase β were collected in mutant:wild-type heteroduplex form from a first CDCE run to enrich the mutants relative to wild-type homoduplexes. High-fidelity PCR was performed on these enriched heteroduplexes before a second round of CDCE in which mutant homoduplexes were observed, collected through electroelution from the capillary column, checked for purity, and sequenced (22, 26, 27).

Fig. 2. Distinguishing true PCR-induced mutations from mutations arising from preexisting DNA damage. Mutations created from bypass of preexisting template damage will remain constant in mutant fraction as template doublings increase, whereas PCR-induced mutations will increase in mutant fraction with each PCR cycle. PCR with recombinant human DNA polymerase β was performed on purified copies of the target sequence until an appropriate number of template doublings was achieved. Individual mutants that increased linearly in fraction with PCR cycles were recognized as putative DNA polymerase errors. Fig. 2 displays the mutations created in APC target sequence bp 3896–3970. A and B, negative control samples amplified only by Pfu DNA polymerase. C, peaks observed after 5 doublings with DNA polymerase β; D, peaks detected after 10 template doublings; E, peaks detected after 15 doublings. Peak 1, a five base deletion within bp 3939–3949; peak 2, a G:C > CG transversion at bp 3970; peak 3, a G:C > T:A transversion at bp 3943; peak 4, a G:C > A:T transition at bp 3914. All of the seven polymerase β replication errors we identified in the APC gene increased linearly with PCR doublings, indicating none arose from preexisting template damage. The mutant peaks detected between peak 1 and the wild-type peak do not increase with template doublings and, therefore, have not been discussed here, although we have isolated and identified these mutants.
Comparison between two mutational spectra has the greatest statistical power when all of the elements of each spectrum are nonrandom events, denoted as “hotspots.” We selected the target APC sequences (bp 3896–3970 and 4355–4420) because they contained the most frequently observed mutations or hotspots in human colon tumors. Within these two sequences, 198 separate colon tumors were reported to carry 51 separate point mutations (12). Of these mutations, 13 were reported two or more times and were denoted as nonrandom events (hotspots) using \( \alpha = 0.05 \) and a Bonferroni correction to account for the fact that hotspots could arise at any position in the scanned target sequences of 141 bp. These data are summarized as a function of APC position in Fig. 1A. The identity and frequency of each of the 13 hotspots within the target APC sequences are provided in Table 1.

To determine the human DNA polymerase \( \beta \) replication error spectrum, we used the polymerase to copy the two target APC sequences in a PCR-like process (21, 30). The products of the PCR-like reactions were analyzed using CDCE coupled to high-fidelity PCR (22, 23). Under the conditions used, polymerase \( \beta \) created 7 mutational hotspots, each occurring at \( >0.5\% \) of all polymerase \( \beta \)-induced mutants within each target APC sequence. Of the identified polymerase \( \beta \) replication hotspots, four were deletion mutations, and three were bp substitutions. These data are summarized in Fig. 1B. Independent experiments eliminated the important possibilities that these mutations arose from bypass of preexisting damage to the template DNA or bypass of heat-induced damage (Figs. 2 and 3, respectively). We calculated the fidelity of polymerase \( \beta \) during replication of the target APC sequences and found that polymerase \( \beta \) has an average error rate of \( 5.2 \times 10^{-7} / \text{bp/doubling} \) in bp 3896–3970 and an average error rate of \( 9.7 \times 10^{-7} / \text{bp/doubling} \) in bp 4355–4420.4

A comparison of the human polymerase \( \beta \) replication errors in APC bp 3896–3970 with those of commercially available polymerases Pfu and Thermus aquaticus (Taq) revealed that there were no hotspots in common with either the in vivo APC tumor spectrum or the polymerase \( \beta \) spectrum. Pfu, a highly accurate DNA polymerase, created G:C \( \rightarrow \) T:A transversions, whereas Taq, an error-prone polymerase that, like polymerase \( \beta \), lacks \( 3' \rightarrow 5' \) exonuclease activity, created A:T \( \rightarrow \) G:C transition mutations (data not shown). Not one of the Pfu or Taq DNA polymerase replication hotspots was concordant with any of the polymerase \( \beta \) replication errors.

Three polymerase \( \beta \) replication errors were found among the 13 in vivo hotspot mutations of the APC tumor spectrum. It was necessary to test the null hypothesis that this degree of concordance (the number of hotspots shared by the two spectra) could have arisen by chance. For this purpose, we used the hypergeometric distribution to ask what would be the probability of having any three mutations from the in vivo observations overlap with any three of a different set of mutations

Table 1 In vivo hotspot mutations detected within APC gene bp 3896–3970 and bp 4355–4420

<table>
<thead>
<tr>
<th>Basepair</th>
<th>Mutation</th>
<th>Observations</th>
<th>Surrounding APC sequence</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>3925</td>
<td>G:C &gt; A:T</td>
<td>2</td>
<td>TGCAA</td>
<td>Gln &gt; Stop</td>
</tr>
<tr>
<td>3934</td>
<td>G:C &gt; T:A</td>
<td>4</td>
<td>CAGAA</td>
<td>Gln &gt; Stop</td>
</tr>
<tr>
<td>3938-41</td>
<td>TAA del</td>
<td>2</td>
<td>AATAAAAAGA</td>
<td>Frameshift</td>
</tr>
<tr>
<td>3939-42</td>
<td>A ins</td>
<td>2</td>
<td>ATAAAAAGAAGATT</td>
<td>Frameshift</td>
</tr>
<tr>
<td>3939-49</td>
<td>AAAGA/AAAG del</td>
<td>80</td>
<td>AAGAA</td>
<td>Gln &gt; Stop</td>
</tr>
<tr>
<td>3943</td>
<td>G:C &gt; T:A</td>
<td>5</td>
<td>AAGAAAG</td>
<td>Frameshift</td>
</tr>
<tr>
<td>3943-46</td>
<td>GAA del</td>
<td>2</td>
<td>TAGGTC</td>
<td>Frameshift</td>
</tr>
<tr>
<td>3959-60</td>
<td>G del</td>
<td>2</td>
<td>AGCGA</td>
<td>Arg &gt; Stop</td>
</tr>
<tr>
<td>4366</td>
<td>G:C &gt; A:T</td>
<td>32</td>
<td>CTAAAAATA</td>
<td>Frameshift</td>
</tr>
<tr>
<td>4378-82</td>
<td>A del</td>
<td>2</td>
<td>TGAAGAG</td>
<td>Frameshift</td>
</tr>
<tr>
<td>4400-03</td>
<td>AA del</td>
<td>2</td>
<td>AAAGAGAGAGAGTG</td>
<td>Frameshift</td>
</tr>
<tr>
<td>4403-12</td>
<td>AG del</td>
<td>18</td>
<td>AAAGAGAGAGAGTG</td>
<td>Frameshift</td>
</tr>
<tr>
<td>4403-12</td>
<td>AGAG del</td>
<td>6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3. Control for heat-induced mutations within the set of polymerase \( \beta \) replication errors. During PCR, the denaturation step (at 100°C in this case) can create potentially mutagenic lesions (3, 4). Mutational spectra of polymerase \( \beta \) errors determined after PCR with 30-s denaturation steps were thus compared with spectra determined after increasing heat exposure to 300 s to discover if any putative DNA polymerase error was created as a thermal artifact. Fig. 3 displays the mutant peaks detected in APC target bp 3896–3970 after five doublings with either 30-s (shown in duplicate in C and D) or 300-s (shown in duplicate in E and F) denaturation steps. Negative control samples based on amplification with Pfu DNA polymerase alone are shown in A and B. The areas of each polymerase \( \beta \) replication mutation (peaks 1–4, identified in Fig. 2) were determined through comparison with the area of an internal standard. None of the seven identified polymerase \( \beta \) replication errors were affected by the increased period of thermal denaturation, indicating they did not arise from thermally induced DNA lesions during the PCR process.

4 B. P. Muniappan and W. G. Thilly, manuscript in preparation
(the polymerase β replication errors). We estimated that there are ≥564 different phenotypically observable mutations that could occur within the 141 target bp of the APC gene. These include 282 types of frameshift mutations (additions or deletions) and 282 possible bp substitutions that could result in stop codons, splice site errors, and missense mutations that inactivate the APC gene (this latter estimate was calculated by multiplying the 141 target bp by the 3 bp substitutions possible at any base and then subtracting the number of wobble base positions). With ≥564 possible phenotypically observable mutational events, we calculate that the probability of observing concordance of any three hotspot mutations from the set of seven polymerase β replication errors with the set of 13 in vivo APC mutations is <4 × 10⁻⁷. This concordance allows us to reject the null hypothesis that the concordance occurred by chance with reasonable confidence.

**DISCUSSION**

The three tumor hotspots concordant with the human DNA polymerase β replication error spectrum alone contain 87 of the 159 separate APC mutational events that have been reported within the scanned sequences or 54% of the in vivo hotspot mutations within 141 bp. The most frequently observed APC mutation in human colon tumors, representing ~15% of germline and 7% of somatic APC mutations leading to colon cancer (12), was a prominent member of the polymerase β error spectrum. This deletion of AAAAG or AAAGA within bp 3939–3949 occurs in a region containing a twice-repeated motif: AAT(A)AAAAGAAAAG(A)TTG. A second concordant hotspot mutation is a 1-bp deletion that occurs within the repeat sequence CCTAAAAATAA (bp 4378–4382). These two mutations could have resulted from a polymerase slippage error within repeated nucleotide sites as hypothesized by Streisinger (31). The third concordant hotspot is a G:C > T:A transversion mutation at bp 3943 (AAA-GAAA) that changes glutamic acid to a stop codon. This mutation might also have been formed during replication by a “slippage” misalignment (32) of the daughter strand where bp 3940–3942 of the nascent strand is paired with bp 3935–3937 (AGAAATA) of the template strand. Replication past the next bases and then realignment of the daughter strand to the correct position would produce a G:A mismatch at bp 3943 that is stabilized by the correct neighboring A:T pairs.

Four in vivo hotspots were not found in the human polymerase β spectrum of APC replication errors. This was not unexpected because polymerase β would not be likely to mediate all human point mutations. It is of immediate interest to discover if other human DNA polymerases, such as δ and ε, which are involved in primary DNA replication and are implicated in proliferating cell nuclear antigen-mediated base excision repair (19, 33), create any of the observed in vivo APC mutations. Such measurements will require availability of these polymerases in suitable form for use in similar DNA replication experiments.

To our knowledge, the four DNA polymerase β hotspots that were not found in the in vivo spectrum have not appeared as even single occurrences among all APC mutations detected in colon tumors. One of these hotspot mutations is a G:C > A:T transition at bp 3914 that substitutes valine for alanine. This conservative amino acid substitution may not affect APC protein function and therefore not produce a phenotypically observable mutation. Another such mutation exchanges aspartic acid (an acidic amino acid) for histidine (a basic amino acid). Two other discordant mutations, a 35-base deletion within bp 4378–4423 and a 5-base deletion within bp 4416–4420 would, however, be expected to inactivate the APC gene product. The 35-base deletion could have been created through an excision repair and fills gaps of one to several bases (19).

Consequently, it may be that there is no physical opportunity in vivo for polymerase β to create such a large deletion. However, the fourth discordant polymerase β mutation is a 5-base deletion (bp 4416–4420) that would be expected to be both molecularly possible and phenotypically detectable. Plausible reasons for its absence in the set of human tumor mutations could include absence by chance from the limited number of in vivo mutational reports, absence of damage-induced repair at this sequence in vivo, or suppression by undefined editing or repair functions in vivo.

The mutational hotspots created by human DNA polymerase β in APC bp 3986–3990 were compared with mutations created by exonucleolytic Pfu polymerase and error-prone Taq polymerase in the same target sequence. Each DNA polymerase created a distinct mutational spectrum of replication errors, with no discordant mutations. Previous studies have demonstrated that different DNA polymerases create nonidentical replication error spectra within the same DNA sequence (13, 14, 21, 22). Although polymerase β replicates DNA patches of 1–6 bases in a processive manner, replication of longer sequences is performed in a distributive manner, so it might be suggested that our results are biased. Moreover, measurement of polymerase β error rates during synthesis of 1 or >350 bases has shown no significant differences in fidelity (17). Furthermore, simple overexpression of polymerase β in mammalian cells has been found to have strong mutator effects (35). This observation and in vitro experiments demonstrating the ability of polymerase β to compete with primary DNA replication polymerases has led Servant et al. (36) to hypothesize that extra polymerase β molecules could become involved competitively with DNA replication. We would like to offer an alternate hypothesis that excess polymerase β molecules would have a mutator effect by simply increasing the number of unforced excision repair events (37). We further speculate that the number of unforced repair events might well significantly outnumber sites of DNA damage requiring excision repair in an average cell day and therefore leave the mutational fingerprint of excision repair processes using polymerase β throughout the genome of unstable somatic and germinal cells.

In summary, it is clear that the replication error spectrum of human DNA polymerase β contains a significant subset of the in vivo APC hotspots found in human colorectal cancers. To our present knowledge, no other potential cause of human point mutation has been found to account for a statistically significant set of specific point mutations in the same nuclear gene sequence (the findings that UV light induces tandem mutations in several experimental systems and that similar tandem mutations are found in sunlight-irradiated human skin strongly support an inference of causation, although these in vivo and in vitro observations were made in different DNA sequences; Ref. 1). We believe our experiments provide a strong test of the specific hypothesis that errors of human DNA polymerase β are a significant source of APC point mutations leading to colon cancer in humans. The results give credence to a more general hypothesis that a significant fraction of human somatic point mutations arise from DNA polymerase β replication errors and point to the importance of discovering the error spectra of other human DNA polymerases.

**ACKNOWLEDGMENTS**

We thank Dr. Rajendra Prasad for graciously providing recombinant human polymerase β and discussing optimal activity conditions. We also thank Dr.
REFERENCES


Lynn S. Ripley for invaluable discussions of this research and Dr. Stephan Morgenthaler for consultation on the determination of statistical significance.