Published online 6 August 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 22* ***8095–8104***
doi:10.1093/nar/gkq685

Deep sequencing analysis of mutations resulting from the incorporation of dNTP analogs

Katherine L. Petrie[1,2] and Gerald F. Joyce[1,2,*]

[1]Department of Chemistry and [2]Department of Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA

Received June 16, 2010; Revised July 13, 2010; Accepted July 15, 2010

ABSTRACT

Next-generation DNA sequencing technology was used to score >100 000 mutations resulting from exposure of a nucleic acid template to a mutagenic dNTP analog during a single pass of a DNA polymerase. An RNA template of known secondary structure was reverse transcribed in the presence of 8-oxo-dGTP, dPTP or both, followed by forward transcription in the presence of standard NTPs. Each mutagen, whether used alone or in combination, resulted in a highly characteristic mutation profile. Mutations were generated at a mean frequency of 1–2% per eligible nucleotide position, but there was substantial variation in the frequency of mutation at different positions, with a SD close to the mean. This variation was partly due to the identity of the immediately surrounding nucleotides and was not significantly influenced by the secondary structure of the RNA template. Most of the variation appears to result from idiosyncratic features that derive from local sequence context, demonstrating how different genetic sequences have different chemical phenotypes.

INTRODUCTION

Genetic mutations can have deleterious effects in the short term, but also provide the basis for evolutionary innovation, and thus are critical to understanding heritable disease, cancer progression and the development of novel biological traits. The frequency of mutations can be increased by various mutagens, each mutagen resulting in a characteristic spectrum of genetic changes. One of the most common cellular mutagens is 8-oxo-dG, which results from oxidative damage to DNA and increases the frequency of G-to-T mutations due to its ability to pair with either C or A residues (1–3) (Figure 1A). Another potent mutagen is the synthetic compound dP, containing the bicyclic pyrimidine analog 3,4-dihydro-8H-pyrimido-[4,5-C][1,2]oxazin-7-one (4). This compound can pair with either G or A residues, and thus promotes C-to-T mutations (5,6) (Figure 1B). In the laboratory, mutagens such as these are employed to study the mechanisms of genetic alteration and to construct populations of variants that can be culled for individuals possessing desired phenotypic characteristics.

The usual approach for measuring the frequency of mutations that result from exposure to a particular mutagen is to score a phenotypic change that derives from a specific genetic polymorphism. Several different sites of polymorphism can be interrogated to evaluate how the mutation profile differs in different sequence contexts. Any phenotypic assay is indirect and potentially affected by selection bias, although one can couple the phenotypic assay with DNA sequencing to confirm the presence or absence of the indicated mutations. With the advent of 'next generation' sequencing technologies, it should be possible to avoid phenotypic scoring and simply sequence large numbers of DNA molecules that have been exposed to a mutagen. Even then, one must guard against potential selection bias that can arise whenever different DNA sequences are co-amplified.

Ideally one would expose a sample nucleic acid to a mutagen, then sequence large numbers of individual molecules without performing amplification. For mutagens that are provided as a dNTP analog a minimum of two passes of a polymerase are required: one to incorporate the analog into DNA, and the other to copy the potentially modified DNA to bring about either mutation or reversion. Until single-molecule sequencing technologies become available, further amplification is required, which should be carried out in the absence of the mutagen and with minimal selective bias.

The present study involved reverse transcription of an RNA template in the presence of either 8-oxo-dGTP or

*To whom correspondence should be addressed. Tel: +1 858 784 9844; Fax: +1 858 784 2943; Email: gjoyce@scripps.edu

© The Author(s) 2010. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

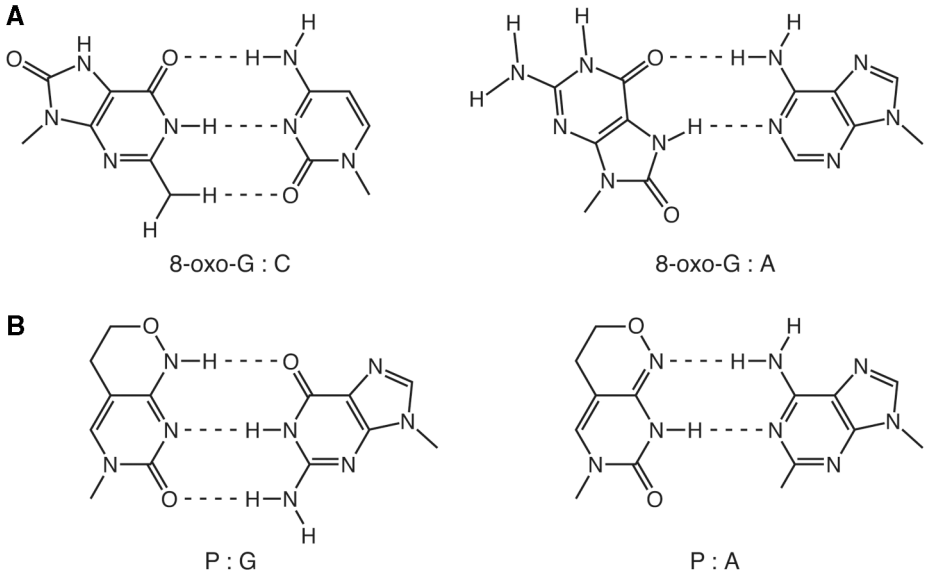

Figure 1. Alternative pairing of mutagenic base analogs. (**A**) 8-oxo-G can pair with either C or A; (**B**) P can pair with either G or A.

dPTP (or both), followed by forward transcription in the absence of a mutagen, then high-fidelity PCR amplification and the sequencing of millions of DNA molecules. Control reactions, carried out in the absence of a mutagen, were processed in a side-by-side manner to assess the basal mutation rate and potential sequencing errors. The template RNA contained 126 nucleotide positions that were potentially mutable, and the sequencing coverage averaged >40 000 reads per position. Thus it was possible to obtain a precise estimate of the frequency of mutation at each position and to survey enough positions to determine the effects of sequence context.

Reaction conditions were chosen that sought to maximize incorporation of the mutagen without compromising the efficiency of polymerization. Under these conditions it was possible with 8-oxo-dGTP to generate A-to-C changes (referring to the starting compared to product RNA) at a frequency of 1.2% per eligible position; and with dPTP to generate A-to-G and G-to-A changes at a frequency of 1.7 and 1.6% per eligible position, respectively. When the two mutagens were combined, their effects were additive. The frequency of mutations at non-eligible positions was <0.1% overall, and did not differ significantly compared to that for control samples that had not been exposed to a mutagen.

Surprisingly, there was a high degree of variation in the frequency of mutagen-induced changes for different eligible positions, with an SD close to the mean. Some of this variation could be attributed to nearest-neighbor effects, but most appears to be idiosyncratic with regard to local sequence context. Compositional biases and the secondary structure of the starting template nucleic acid do not play a significant role. Rather it appears that different genetic sequences have different chemical phenotypes that reflect both the intrinsic properties of the nucleic acid molecule and the way those properties are expressed within the environment of a polymerase active site.

MATERIALS AND METHODS

Mutagenesis procedure

Catalytic RNA having the sequence 5′-AGAAGAAAGA AAUUUCUCUAAUAGUGAUCCUUGUGAUUUGU GUGAUCUCAAUCCUAAGACUGAACGUUAUGG AUCAAUGGGUAGGUUCCAAGUAGAGCAGACG AUAAAGUGUUUCCGUUCCCUAGUAGAUUGCG AGUCGUAUUUUGACGUGGCUGACUCCGUGCC AUCC-3′ was prepared by *in vitro* transcription of the corresponding DNA template (primer binding sites underlined), then purified by denaturing polyacrylamide gel electrophoresis (PAGE). The RNA was allowed to react with a chimeric DNA-RNA substrate having the sequence 5′-CATCGTGCCTTGCTGCTCTAATACGA CTCACUAUU-3′ (T7 promoter sequence underlined; RNA residues in bold), resulting in ligation of the substrate to the 5′-end of the catalytic RNA. The reaction mixture contained 2 μM catalytic RNA, 10 μM substrate, 15 mM $MgCl_2$, 50 mM KCl and 50 mM EPPS (pH 8.5), which was incubated at 37°C for 1 h. The ligated RNA was again purified by PAGE.

Of the ligated RNA, 30 pmols were used to initiate reverse transcription and subsequent forward transcription in a common reaction mixture containing 0.5 μM RNA and 2.5 μM cDNA primer having the sequence 5′-T_{23}XXXGGATGGCACGGAGTCAG-3′, where X corresponds to an abasic DNA spacer (Glen Research). The abasic spacer causes transcription to terminate at a position corresponding to the starting RNA molecule, and the oligo(T) tail provides size discrimination between the cDNA and newly-transcribed RNA. The reaction mixture also contained 4.5 U/μl RNAse H–MMLV reverse transcriptase, 2.5 U/μl T7 RNA polymerase, 0.001 U/μl inorganic pyrophosphatase, various formulations of dNTPs, 2 mM each NTP, $MgCl_2$ at a concentration of 14.2 mM plus the total concentration of dNTPs, 50 mM KCl, 4 mM DTT and 50 mM EPPS (pH 8.5), which was incubated at 37°C for 1 h. The newly transcribed RNAs were purified

by high-resolution PAGE, thus excluding any mutagen-containing cDNAs.

Sequencing and data analysis

Of the mutagenized RNA, 10 pmols were reverse transcribed in a reaction mixture containing 0.5 μM RNA, 2.5 μM cDNA primer having the sequence 5′-GGATGGCACGGAGTCAG-3′, 4.5 U/μl RNAse H–MMLV reverse transcriptase, 0.5 mM each dNTP, 3 mM $MgCl_2$, 75 mM KCl, 10 mM DTT and 50 mM Tris (pH 8.3), which was incubated at 37°C for 1 h. The RNA then was digested using 5 U RNase H and 2 U RiboShredder (Epicentre Biotechnologies), incubating at 37°C for 20 min. The remaining cDNA then was PCR amplified using high-fidelity PfuUltra polymerase (Stratagene), employing the same primer as above and a forward primer having the sequence 5′-AGAAGAAAGA AATTTCTCTAATAGTG-3′. Of the PCR products, 5 pmols were fragmented in a 20-μl reaction mixture containing 10^{-4} U/μl DNase I, which was incubated at 37°C for 20 min. Fragments of 55–105 bp were size-selected in a 2% agarose gel and purified using the Zymo gel extraction kit (Zymo Research). The fragments were end repaired, tailed with deoxyadenosine, and ligated to Illumina-compatible adaptors containing 6-nt barcodes having the sequence: 5′-CAACCT-3′ for the control library; 5′-TAC GTT-3′ for the 8-oxo-dGTP library; 5′-AACCAT-3′ for the dPTP library and 5′-GACTGT-3′ for the combined 8-oxo-dGTP/dPTP library. Ligated products containing 120–200 bp were purified by 2% agarose gel electrophoresis, PCR amplified to enrich fragments containing both adaptors, and again gel purified, selecting molecules containing 150–210 bp.

The four barcoded libraries were pooled and used for cluster generation in a single multiplexed flow cell lane in the Illumina Gene Analyzer IIx system. Single-read sequencing by synthesis (80 cycles) was performed, running phi X 174 genomic DNA as a control in a separate lane of the flow cell. The Illumina Genome Analyzer Pipeline Software (version 1.4.0) was used to carry out image analysis, base calling and quality score calibration. Reads were sorted by barcode and exported in the FASTQ format.

Subsequent data processing was performed using the high-throughput sequencing module of CLC Genomics Workbench (version 3.6; CLC bio). One million sequences were imported for each library. Trimming was done to remove adaptor sequences and ambiguously read nucleotides. Trimmed reads <20 nt were discarded and the remaining reads were aligned relative to the reference sequence. The default local alignment settings for long reads were used to rank all potential matches, with mismatch cost = 2, deletion cost = 3 and insertion cost = 3. The highest scoring matches that shared ≥80% identity with the reference sequence across ≥50% of their length were included in the alignment. This permissive alignment ensured that reads derived from highly mutated molecules would not be discarded.

The SNP detection application in CLC Genomics Workbench was used to evaluate variants that differed compared to the reference sequence. This software uses a modified version of the neighborhood quality standard (7,8) to determine which reads at a given position should be included in the analysis. For inclusion, a position must meet a minimum quality score, must be in the center of an 11-nt window with no more than two gaps and/or mismatches compared to the reference sequence, and must be accompanied by surrounding nucleotides in the window that also meet a minimum quality score. Either increasing the maximum number of gaps/mismatches or decreasing the window length did not significantly alter the calculated mutation frequency. Various minimum quality scores were tested to evaluate sensitivity to this variable.

RESULTS

Mutagenesis with dNTP analogs

A well-characterized RNA molecule, with known secondary structure, was chosen as the starting template for incorporation of mutagenic dNTP analogs. This RNA is the DSL ribozyme, which catalyzes ligation of an oligonucleotide substrate to the 5′-end of the ribozyme (9). If the substrate has the sequence of an RNA polymerase promoter element, then the ribozyme can be made to undergo repeated rounds of ligation and selective amplification at a constant temperature within a single reaction mixture (10). Each round of amplification involves reverse transcription to generate promoter-containing cDNA followed by forward transcription to generate new copies of the RNA. If the substrate for RNA-catalyzed RNA ligation is present in the mixture, then the newly-synthesized RNA also can react and initiate further amplification. If, however, one begins with reacted RNA and does not provide substrate, then the amplification process is limited to a single round of reverse and forward transcription. Such a condition was employed to allow one pass to incorporate the dNTP analog followed by one pass to copy the incorporated analog and bring about mutation.

Preliminary experiments were carried out to determine the maximum concentration of 8-oxo-dGTP or dPTP (or both) that could be included in the reaction mixture without compromising the efficiency of polymerization. The concentrations of the competing dNTPs were reduced as tolerated to maximize the frequency of mutations brought about by the dNTP analogs. For 8-oxo-dGTP, which can pair with either C or A, the concentrations of dGTP and TTP were reduced and for dPTP, which can pair with either G or A, the concentrations of dCTP and TTP were reduced. Compared to the standard condition for reverse transcription, which employed 0.2 mM of each dNTP, the condition for 8-oxo-dGTP mutagenesis employed 8 mM 8-oxo-dGTP, 0.05 mM dGTP and 0.075 mM TTP; the condition for dPTP mutagenesis employed 0.25 mM dPTP, 0.075 mM dCTP and 0.075 mM TTP and the condition for combined 8-oxo-dGTP/dPTP mutagenesis employed 8 mM 8-oxo-dGTP, 0.25 mM dPTP, 0.05 mM dGTP,

0.075 mM dCTP and 0.075 mM TTP (all other dNTP concentrations remained 0.2 mM).

Employing the formulations described, as well as standard dNTP concentrations as a control, the starting RNA was reverse transcribed using RNAse H–MMLV reverse transcriptase, then forward transcribed using T7 RNA polymerase in the presence of the four standard NTPs. This allowed incorporation of 8-oxo-dGTP opposite C or A positions of the starting RNA, which in turn could be read out as either C or A in the product RNA. Similarly the incorporation of dPTP could occur opposite G or A positions of the starting RNA and could be read out as either G or A. The product RNA, which is smaller in size compared to both the starting RNA and the cDNA, was purified by high-resolution PAGE, then reverse transcribed in the presence of standard dNTPs. This provided mutagen-free DNA that was used for sequence analysis. The RNA was degraded using ribonuclease (avoiding alkali which could promote deamination of dC to dU), then the cDNA was PCR amplified using a high-fidelity DNA polymerase. The PCR products were fragmented, size selected, and prepared for sequencing on the Illumina platform. Adapters with distinct barcode sequences were attached to the four different sets of material so that sequencing could be carried out in a multiplex fashion within a single lane of the flow cell of the Illumina instrument.

The starting amount of RNA was 30 pmols, the amount of mutagenized RNA used for RT-PCR was 10 pmols and the amount of PCR DNA used for sequencing was 5 pmols. Following attachment of the adapters, additional PCR amplification was carried out to enrich molecules that contained adapters at both ends. Subsequent amplification occurred clonally within the flow cell of the sequencer. Approximately 10^{-5} pmols of DNA were clonally amplified and sequenced, which is a sparse sampling of the input material, ensuring that individual sequence reads were unlikely to be related by descent.

Each of the steps of DNA amplification, although carried out with a high-fidelity polymerase and using standard dNTPs, provided an opportunity for additional mutagenesis due to polymerase error. In addition, the standard dNTPs inevitably contain a small amount of modified bases that could promote mutations. Even the synthetic DNA primers may contain chemical lesions that could give rise to mutations in portions of the DNA that were not exposed to the dNTP analogs. Thus the standard dNTP formulation and the primer regions of the amplified material serve as controls to assess other sources of mutation and potential sequencing errors.

Frequency and distribution of mutations

For each of the four dNTP formulations, 1 million arbitrarily chosen sequence reads were analyzed using CLC Genomics Workbench. The raw base calls and associated quality scores were exported from the Illumina platform without applying filters. Then the sequences were trimmed to remove adapter sequences and ambiguously read nucleotides. Phi X 174 genomic DNA was sequenced in parallel to provide calibration standards. Because there was no ambiguity in aligning the individual reads relative to the 169-nt reference sequence, the alignment parameters were made highly permissive, requiring only 80% identity across 50% of the read length.

Mutations relative to the reference sequence were evaluated by applying variable quality score standards, thus providing an analysis of the trade-off between inclusiveness of potentially meaningful data and rejection of potential sequencing errors. The quality score is a measure of the confidence of individual base calls, derived from the relative signal intensity for each of the four possible bases at that position (7,8). Quality scores reflect the likelihood that a base is erroneously called, and are calibrated according to how well they predict the known sequence of a reference sample (in this case phi X 174 DNA). Various quality score cut-offs were applied to each nucleotide position and to the 5 nt on either side of that position. The quality scores ranged from 0.06 to 100% tolerable error for the central nucleotide, and from 0.16 to 100% tolerable error for the surrounding nucleotides. The overall mutation rate did not differ substantially for each of the four different dNTP formulations over a range of quality scores of 0.06–10%, but rose substantially when the tolerable error was 100% (Figure 2). This broad insensitivity of the observed mutation rate to quality score cut-off suggests that the observed rate reflects the true frequency of mutation. The most conservative error threshold of 0.06% was adopted for subsequent data analysis to minimize the inclusion of sequencing errors. Even after applying this conservative filter, there were an average of >40 000 reads per nucleotide position. The coverage at each position is shown in Supplementary Figure S1.

The overall mutation frequency with the standard dNTPs was 0.08% per nucleotide position. In the presence of 8-oxo-dGTP this increased to 0.39%, with dPTP it was 0.91% and with combined 8-oxo-dGTP/dPTP it was 1.03% (Table 1). Nearly all of the mutations with 8-oxo-dGTP were A-to-C changes (referring to the starting compared to product RNA), and these occurred at a frequency of 1.20% per eligible position. An 'eligible position' refers to A residues in the RNA for 8-oxo-dGTP mutagenesis and to either A or G residues in the RNA for dPTP mutagenesis. With dPTP, both A-to-G and G-to-A changes were observed, at a frequency of 1.71 and 1.61% per eligible position, respectively. With the combination of 8-oxo-dGTP and dPTP, A-to-C, A-to-G and G-to-A changes all were observed, at a frequency of 1.09, 1.31 and 1.40%, respectively.

It is notable that 8-oxo-dGTP had an asymmetric effect, generating frequent A-to-C changes, but C-to-A changes at a rate than did not differ significantly compared to the control reaction with standard dNTPs. This indicates that 8-oxo-dGTP is readily incorporated opposite A residues of an RNA template, but either 8-oxo-dGTP is not incorporated opposite C residues or ATP is not readily incorporated opposite 8-oxo-dG residues of a DNA template. In contrast, there was no pronounced asymmetry with dPTP, indicating that it can be incorporated opposite A or G residues of an RNA template and that GTP or ATP can be incorporated opposite dP residues of

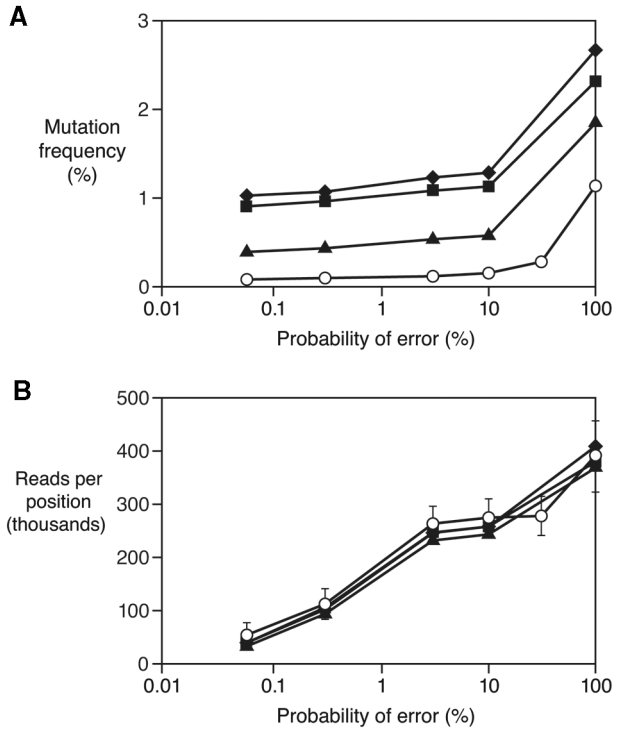

Figure 2. Effect of quality score on DNA sequence analysis. (**A**) Average mutation frequency and (**B**) average coverage for different quality score cut-offs. Error probabilities of 0.06, 0.32, 3.2, 10, 32 and 100% correspond to Phred quality scores for the sequenced nucleotide of 32, 25, 15, 10, 5 and 0, respectively. Quality scores for the five nucleotides on either side of the sequenced nucleotide were 28, 20, 10, 5, 0 and 0, respectively. The Phred quality score is given by: $Q = {}^{-}10\ \log_{10}(P)$, where P is the calculated probability that the base call is incorrect (7). Open circles, standard dNTPs; triangles, 8-oxo-dGTP; squares, dPTP; diamonds, combined 8-oxo-dGTP/dPTP. The trend lines in **A** were similar in form, but with different absolute values, for the individual nucleotide positions. Error bars in **B** correspond to one standard deviation, shown only for the standard dNTPs, but of similar magnitude for the other three formulations.

a DNA template. The combination of 8-oxo-dGTP and dPTP showed roughly additive effects compared to the behavior of the two dNTP analogs alone.

With deep sequencing coverage, it was possible to determine the mutation frequency at each nucleotide position for each of the four dNTP formulations. Nearly all of the eligible positions exhibited substantially higher frequency of mutation compared to the non-eligible positions (Figure 3). There was a surprisingly high degree of variation in mutation frequency at the different eligible positions, with a SD close to the mean (Table 1). For A-to-C changes brought about by 8-oxo-dGTP, the range of mutation frequencies (at 31 eligible positions) was 0.05–3.89%, with an interquartile range of 0.58–1.46%. For A-to-G changes brought about dPTP, the range of mutation frequencies (at the same 31 positions) was 0.52–4.02%, with an interquartile range of 1.16–2.11%. For G-to-A changes brought about dPTP, the range of mutation frequencies (at 33 eligible positions) was 0.38–4.16%, with an interquartile range of 0.96–2.27%. The combination of 8-oxo-dGTP and dPTP

Table 1. Type and frequency of mutations for four different dNTP formulations

dNTP formulation	Reference nucleotide	Mutation frequency (%)			
		A	G	C	U
Standard	A	–	0.03	0.01	0.01
	G	0.05	–	0.01	0.07
	C	0.07	0.01	–	0.04
	U	0.01	0.01	0.03	–
8-oxo-dGTP	A	–	0.04	1.20 ± 1.00	0.01
	G	0.05	–	0.01	0.08
	C	0.08	0.01	–	0.06
	U	0.01	0.01	0.02	–
dPTP	A	–	1.71 ± 0.89	0.01	0.01
	G	1.61 ± 0.86	–	0.01	0.07
	C	0.07	0.01	–	0.05
	U	0.01	0.01	0.03	–
8-oxo-dGTP + dPTP	A	–	1.31 ± 0.70	1.09 ± 0.92	0.01
	G	1.40 ± 0.80	–	0.01	0.08
	C	0.09	0.01	–	0.05
	U	0.01	0.01	0.03	–

resulted in a distribution of mutations that was very similar to the combined effects of 8-oxo-dGTP and dPTP alone (Figure 4). This demonstrates the independent and additive behavior of the two mutagens, as well as the consistency of the observed position-dependent variation among the different sample preparations.

Sources of position-specific variation

An analysis was performed to determine whether the frequency of mutation at a particular position is influenced by the identity of the immediately surrounding nucleotides. For mutations at A positions brought about by 8-oxo-dGTP, there was a significantly enhanced likelihood of either C or U at the upstream (to the 5′-side) nucleotide position of the RNA, and conversely a significantly reduced likelihood of either C or U at the downstream nucleotide position (Figure 5). For mutations at A positions brought about by dPTP, there was no effect of the upstream nucleotide of the RNA, but a significantly enhanced likelihood of C at the downstream position. For mutations at G positions brought about by dPTP, there was both a reduced likelihood of U at the upstream position and a reduced likelihood of A at the downstream position.

Mutations induced by 8-oxo-dGTP are almost exclusively due to its incorporation opposite A residues of the RNA template. This requires 8-oxo-dGTP to adopt the *syn* conformation (11), which appears to occur more readily if the preceding residue of the cDNA is a pyrimidine. Perhaps the reduced stacking energy of a preceding pyrimidine is more conducive to the distorted geometry of *syn*-8-oxo-dG. Once 8-oxo-dG has become incorporated, there is a preference for a purine residue at the subsequent position of the cDNA, perhaps as a means to overcome the distorted geometry through increased stacking energy of the trailing nucleotide.

P pairs ambiguously with either G or A, existing as either the amino or imino tautomer, respectively (12,13). A-to-G changes brought about by dPTP require

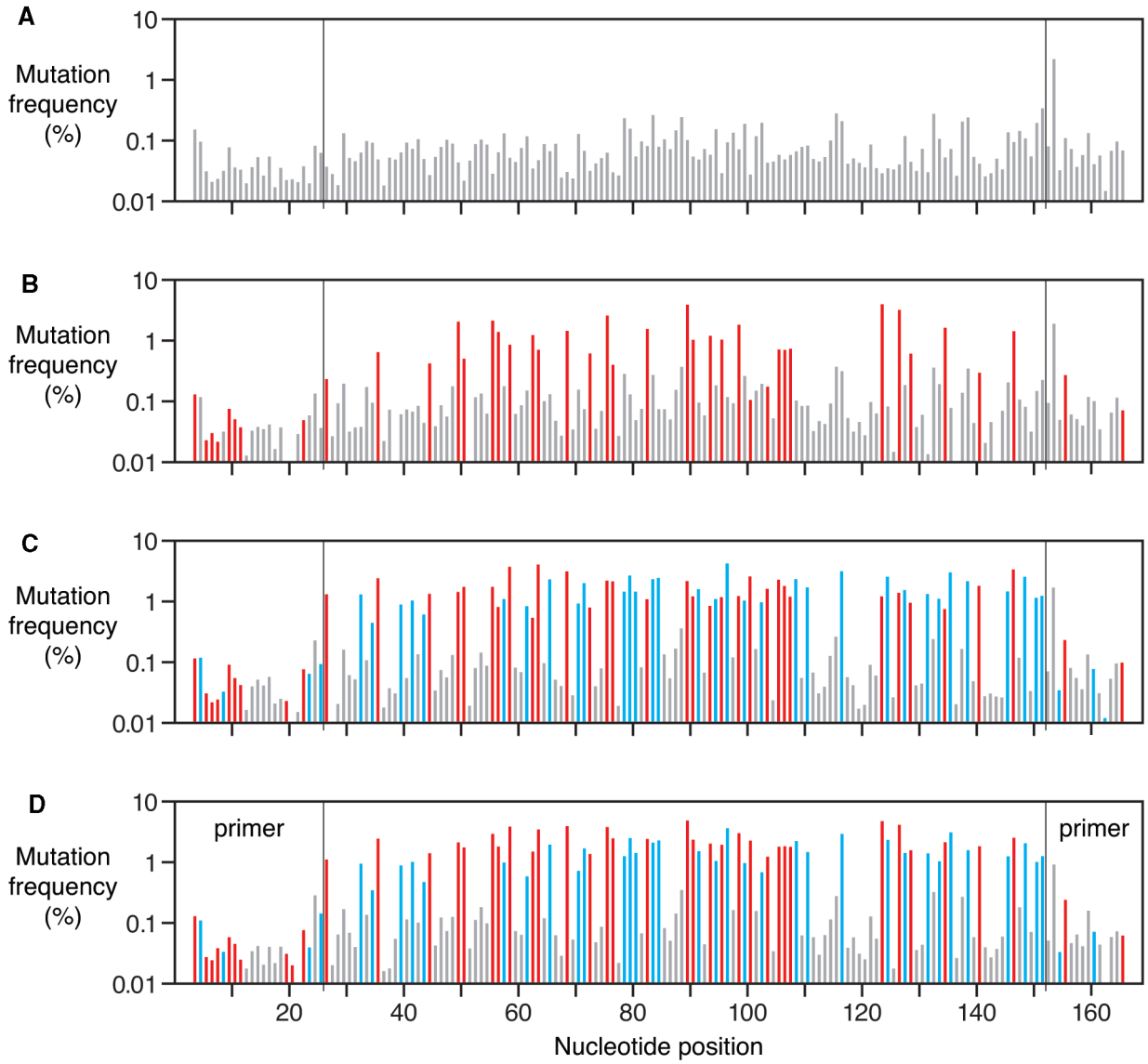

Figure 3. Mutation frequency at each nucleotide position for four different dNTP formulations. (**A**), standard dNTPs; (**B**), 8-oxo-dGTP; (**C**), dPTP; (**D**), combined 8-oxo-dGTP/dPTP. Nucleotide positions refer to the starting RNA, with the two primer regions (nucleotides 1–26 and 153–169) demarcated. Positions that were eligible for mutation are shown in color: A residues (red) were eligible for mutation by either 8-oxo-dGTP or dPTP; G residues (blue) were eligible for mutation by dPTP only. Note that mutation frequency is shown on a log scale.

incorporation of dPTP (as the imino tautomer) opposite A residues of the RNA template. This appears to occur more readily when the preceding residue of the cDNA is dG. P residues within the cDNA must then (as the amino tautomer) direct incorporation of GTP into the product RNA, an event that does not appear to be influenced by the identity of the preceding residue of the RNA. Conversely, G-to-A changes brought about by dPTP require incorporation of dPTP (as the amino tautomer) opposite G residues of the RNA template, which occurs most readily when the preceding residue of the cDNA is not T. P residues within the cDNA must then (as the imino tautomer) direct incorporation of ATP into the product RNA, which occurs most readily when the preceding residue of the RNA is not U.

Taken together, nearest-neighbor effects account for only a small portion of the observed position-specific variation in mutation frequency. Another potential source of variation is the structure of the RNA template, which becomes unfolded during reverse transcription, but nonetheless can affect the propensity of the RNA to track through the polymerase active site, and therefore affect the kinetics of dNTP incorporation (14,15). The secondary structure of the starting molecule has been well characterized (9,16,17). No obvious correlation exists between secondary structural features and the frequency of mutation at nearby positions (Supplementary Figure S2). It is possible that the secondary structure of the cDNA has some influence, although that structure is not known.

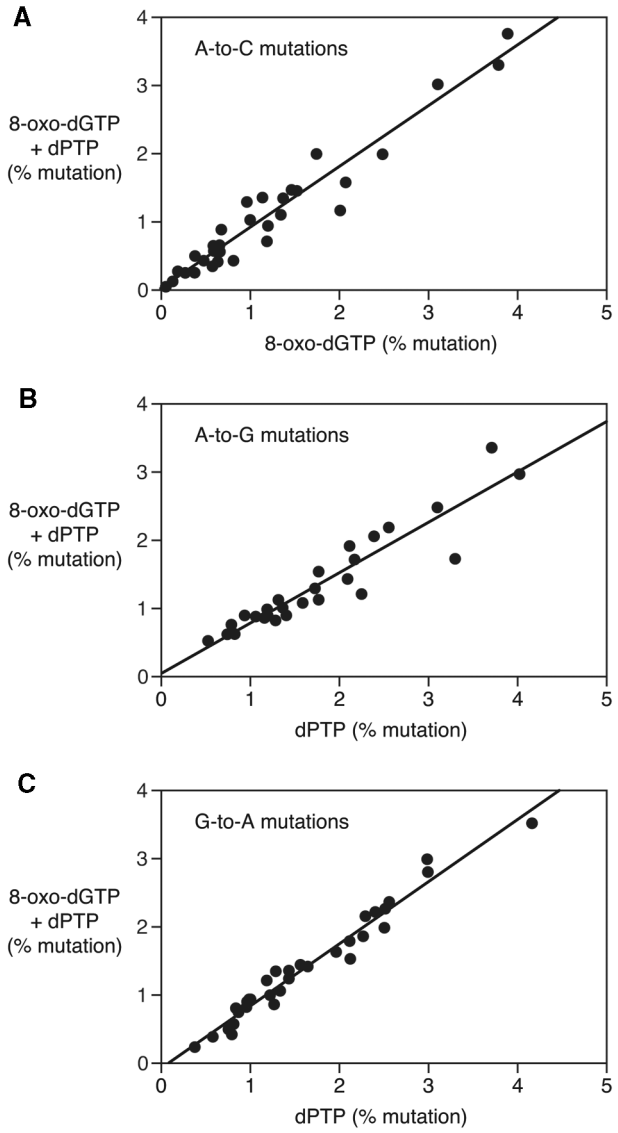

Figure 4. Comparison of mutation frequency at each eligible position when exposed to either one or both mutagenic dNTP analogs. (**A**), A-to-C mutations when exposed to either 8-oxo-dGTP or both mutagens; (**B**), A-to-G mutations when exposed to either dPTP or both mutagens; (**C**), G-to-A mutations when exposed to either dPTP or both mutagens. Linear regression coefficients were 0.97, 0.94 and 0.98, respectively.

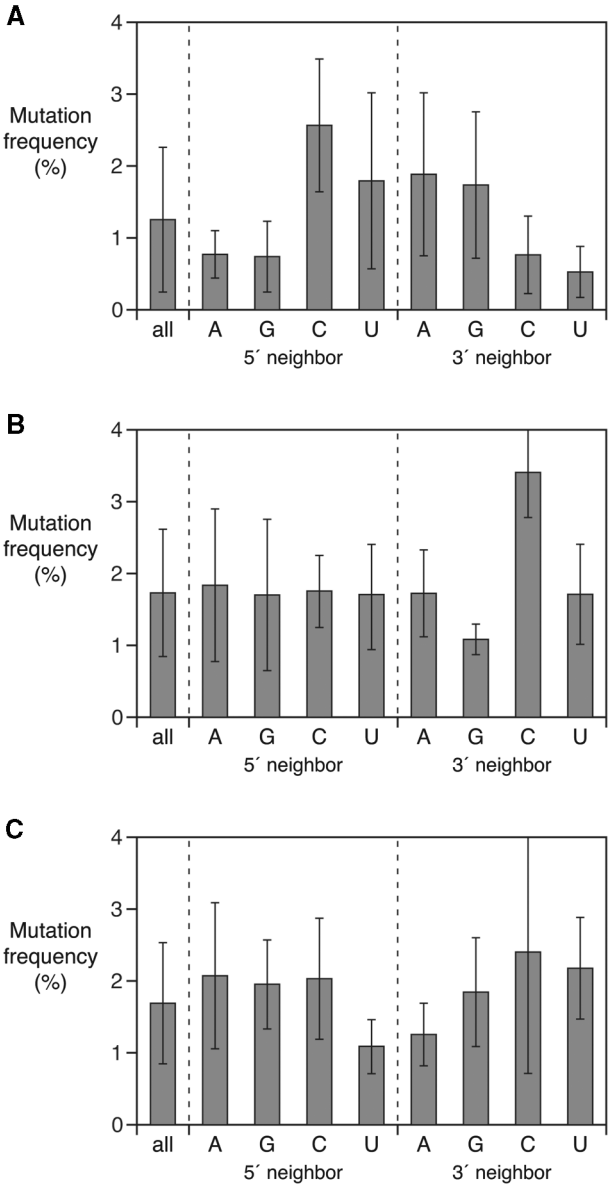

Figure 5. Analysis of the effect of adjacent nucleotides on the mutation frequency at each eligible position. (**A**), Mutations at A positions brought about by 8-oxo-dGTP; (**B**), Mutations at A positions brought about by dPTP; (**C**), Mutations at G positions brought about by dPTP. Adjacent nucleotides are either immediately upstream (5′-neighbor) or immediately downstream (3′-neighbor) within the starting RNA. Error bars correspond to 1 SD.

The preceding analysis of the frequency and distribution of mutations excluded residues within the two primer regions, which might be regarded as immutable. The region of the cDNA primer was never exposed to mutagenic dNTP analogs, and the region of the second primer, although exposed to analogs during cDNA synthesis, would have had any mutations negated during subsequent PCR amplification. However, apparent mutations can occur within the primer regions due to sequence heterogeneity of the primers themselves, mutations that arise during colony amplification within the Illumina sequencer (which employed outside primers) or sequencing errors.

The median frequency of mutation per position within the two primer regions was 0.04%, but there were several positions with a mutation frequency that was at least 2-fold higher (Figure 3). Each of these relative hotspots exhibited a particular favored mutation, and the pattern of favored mutations was the same for all four dNTP formulations. One position in particular, T154 within the region of the cDNA primer, had a frequency of T-to-C changes of 1–2%. This was most likely due to compositional heterogeneity of the synthetic DNA primer itself, although nothing unusual was noted during the corresponding steps of its synthesis. For non-eligible positions between the two primer regions

the median frequency of mutation was 0.06%, similar to that for the two primer regions. This indicates that there is a lower bound for the detection of true mutations of ~0.04–0.06%, which is determined by both false mutations that arise during colony amplification and sequencing errors.

DISCUSSION

Recent advances in DNA sequencing technology make it possible to determine the position-specific frequency of mutation resulting from exposure to a mutagen, and to obtain this information directly, without resorting to phenotypic assays. In the present study, two mutagenic dNTP analogs were examined, one a common cellular mutagen and the other a synthetic compound used in laboratory mutagenesis procedures. A particular set of reaction conditions were chosen to maximize incorporation of the dNTP analogs during a single pass of an RNA-dependent DNA polymerase. Other mutagens, reaction conditions and polymerization protocols might have been chosen and analyzed in a similar manner, but the goal here was to achieve a high level of mutagenesis to enable statistical analysis of the distribution of mutations, while minimizing selective bias due to differential amplification.

As expected, 8-oxo-dGTP resulted in a substantially increased frequency of transversion mutations due to its ambiguous pairing with either C or A residues in the starting RNA template (Figure 1A). These mutations were asymmetrical, with A-to-C changes occurring at a frequency of 1.2% per eligible position, but C-to-A changes occurring at a frequency similar to that of the control reaction with standard dNTPs. Human mitochondrial DNA polymerase has been shown to incorporate 8-oxo-dGTP 13-fold more readily opposite dA compared to dC residues, and to incorporate dGTP 10^4-fold more readily than 8-oxo-dGTP opposite dC residues (18). Thus, even with the 160-fold concentration advantage of 8-oxo-dGTP relative to dGTP employed in the mutagenesis procedure, the incorporation of 8-oxo-dGTP likely was disfavored at C positions. The recently reported crystal structure of 8-oxo-dGTP bound opposite dA in the active site DNA polymerase β confirms the substantially more favorable geometry of this pairing compared to that of 8-oxo-dGTP and dC (19). dPTP, in contrast, gave rise to A-to-G and G-to-A transition mutations in a symmetrical manner, occurring at a frequency of 1.7 and 1.6%, respectively. P is more ambiguous in its pairing compared to 8-oxo-G because P can adopt either the amino or imino tautomer (Figure 1B) in roughly equal proportions (13).

The most striking finding of this study was the high degree of position-specific variation in mutation frequency. With the standard dNTPs the mean mutation frequency was 0.08% per nucleotide position, with a SD of 0.06%. In the presence of one or both mutagenic dNTP analogs the mutation frequency was substantially elevated at all eligible positions, also with a SD close to the mean (Table 1). The mutation frequency at each of the non-eligible positions was highly consistent across the four different dNTP formulations (Figure 3), as was the mutation frequency at each of the eligible positions for formulations involving either one or both mutagens (Figure 4). Thus each nucleotide position appears to have a distinct 'personality', with characteristic features that are elicited by exposure to particular mutagens.

A previous study of 8-oxo-dGTP mutagenesis, employing either an RNA or DNA template and HIV-1 reverse transcriptase, demonstrated 10-fold variation in the frequency of A-to-C mutations at 20 eligible positions (3). That study used a *lacZ* α-complementation assay to score mutants based on phenotype, followed by conventional sequence analysis to confirm the indicated mutations. Only ~200 mutations were scored, compared to more than 100 000 in the present study, but even at that lower level of resolution the distribution of mutations appeared to be highly non-uniform.

With the fine-grained analysis enabled by deep sequencing technology, it becomes clear that 'random mutagenesis' is a misnomer. One can speak of protocols that are intended to introduce mutations in an unpredictable manner (20), and for a given nucleotide position the incidence of mutation is likely to be probabilistic for individual molecules. However, haplotypes that involve combinations of low-frequency mutations will be nearly inaccessible in the population, unless the individual mutations each confer selective advantage and can accumulate successively. Conversely, combinations of high-frequency mutations will be especially well represented, thus allowing high-frequency mutations that are phenotypically neutral to be swept along with adaptive mutations during selective enrichment.

One source of position-specific variation in mutation frequency is the identity of the immediately adjacent nucleotides, most apparent for mutations induced by 8-oxo-dGTP (Figure 5). Mutagenesis by 8-oxo-dGTP depends on G assuming the *syn* conformation, and it is not surprising that this can be influenced by both the upstream nucleotide (for incorporating 8-oxo-G opposite A) and downstream nucleotide (for extending the G-A mismatch). However, the main sources of position-specific variation remain obscure. The secondary structure of the starting RNA template does not appear to play an important role (Supplementary Figure S2). Rather, position-specific variation likely reflects differential rates of dNTP incorporation and subsequent extension that occur in different sequence contexts (14,15). One could investigate this possibility systematically by constructing a family of templates containing, say, all 4096 possible hexameric sequences, exposing them to a mutagen, and using deep sequencing to provide a comprehensive assessment of local sequence effects. Ideally these data should be correlated with stepwise measurements of the rates of reverse and forward transcription, obtained at single-nucleotide resolution (21,22).

The picture that emerges is of a genotype that itself has considerable phenotype. More properly, genotype should be regarded as a purely informational entity that acquires phenotypic characteristics only when embodied as a physical molecule. The exploration of sequence space is

broadest when the opportunity for mutagenesis does not vary for different sequences. Clearly this is not the case for nucleic acid molecules that are copied by a polymerase. Yet sequence space is so vast that even a 10-fold variation in mutation frequency at different nucleotide positions leaves ample opportunity to discover novel sequence combinations that enable evolutionary adaptation. As DNA sequencing technology continues to advance and more genomes are sequenced, it will be interesting to see the extent to which the inherent chemical phenotype of the genetic material can be discerned within biological organisms. In some instances, mutational hotspots themselves may be selectable features that promote the evolvability of adaptive traits.

Laboratory mutagenesis procedures, such as those described here, have practical applications in directed evolution studies. Currently the most popular method for 'random' mutagenesis is error-prone PCR (23–26). The use of 8-oxo-dGTP and/or dPTP provides an attractive alternative (6), especially now that single-pass mutation frequencies of 1–2% can be achieved. However, the resulting mutations are limited to only three types: A-to-C, A-to-G and G-to-A changes (referring to the starting compared to product RNA). If one applied the same technique to an antisense RNA template, generated double-stranded DNA, and transcribed sense-strand RNA products, then those products would contain, in effect, U-to-G, U-to-C and C-to-U changes. If the asymmetry of 8-oxo-dGTP mutagenesis could be overcome, for example, by employing a different polymerase, then C-to-A and G-to-U changes would be accessible as well. Taken together, these procedures might be used to generate 8 of the 12 types of single-nucleotide substitutions.

The continuous *in vitro* evolution system, which has been used to evolve ligase ribozymes (10,16), will benefit especially from having a means to introduce mutations at high frequency and in a continuous manner. Thus far, the selective power of this system has exceeded its ability to generate novel variation, narrowing the breadth of the evolutionary search. With a high level of mutagenesis, approaching the error threshold for the retention of genetic information (27), it should be possible to explore sequence space more broadly, even with the restrictions due to the limited types of mutations and position-specific variation in mutation frequency.

Deep sequencing analysis could be applied more broadly to investigate the mutation spectrum of various polymerases under various reaction conditions, including following exposure of nucleic acid templates or dNTPs to potential mutagens, during polymerization in the presence of dNTP analogs or other mutagens, or following processing of mutated nucleic acids by various DNA repair enzymes. Until recently, it was necessary to carry out phenotypic assays to score a sufficient number of mutations to provide statistically significant information. Next-generation sequencing technology now allows that information to be obtained directly by performing millions of individual sequence reads. A present limitation is that the sample nucleic acid must be amplified prior to sequencing, which can introduce false mutations and sets a lower bound of ~0.05% for the detection of true mutations. With the advent of 'third generation' sequencing technologies, especially those that allow direct and repeated sequencing of individual nucleic acid molecules (28), the threshold for detection of mutations will be substantially lower. Such methods will provide an even better opportunity to uncover the inherent phenotypic properties of individual DNA molecules.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Phillip Ordoukhanian and Lana Schaffer for assistance in processing the DNA sequencing data.

FUNDING

The National Science Foundation (MCB-0614614); the National Institutes of Health Predoctoral Training Program in Molecular Evolution (T32GM080209 to K.L.P.). Funding for open access charge: National Science Foundation (grant no. MCB-0614614).

Conflict of interest statement. None declared.

REFERENCES

1. Pavlov,Y.I., Minnik,D.T., Izuta,S. and Kunkel,T.A. (1994) DNA replication fidelity with 8-oxodeoxyguanosine triphosphate. *Biochemistry*, **33**, 4695–4701.
2. Kamath-Loeb,A.S., Hizi,A., Kasai,H. and Loeb,L.A. (1997) Incorporation of the guanosine triphosphate analogs 8-oxo-dGTP and 8-NH_2-dGTP by reverse transcriptases and mammalian DNA polymerases. *J. Biol. Chem.*, **272**, 5892–5898.
3. Bebenek,K., Boyer,J.C. and Kunkel,T.A. (1999) The base substitution fidelity of HIV-1 reverse transcriptase on DNA and RNA templates probed with 8-oxo-deoxyguanosine triphosphate. *Mutation Res.*, **429**, 149–158.
4. Kong Thoo Lin,P. and Brown,D.M. (1989) Synthesis and duplex stability of oligonucleotides containing cytosine-thymine analogues. *Nucleic Acids Res.*, **17**, 10373–10383.
5. Hill,F., Loakes,D. and Brown,D.M. (1998) Polymerase recognition of synthetic oligodeoxyribonucleotides incorporating degenerate pyrimidine and purine bases. *Proc. Natl Acad. Sci. USA*, **95**, 4258–4263.
6. Zaccolo,M., Williams,D.M., Brown,D.M. and Gherardi,E. (1996) An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Biol.*, **255**, 589–603.
7. Altshuler,D., Pollara,V., Cowles,C., Van Etten,W., Baldwin,J., Linton,L. and Lander,E. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.
8. Brockman,W., Alvarez,P., Young,S., Garber,M., Giannoukos,G., Lee,W.L., Russ,C., Lander,E.S., Nusbaum,C. and Jaffe,D.B. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.*, **18**, 763–770.
9. Ikawa,Y., Tsuda,K., Matsumura,S. and Inoue,T. (2004) *De novo* synthesis and development of an RNA enzyme. *Proc. Natl Acad. Sci. USA*, **101**, 13750–13755.
10. Wright,M.C. and Joyce,G.F. (1997) Continuous *in vitro* evolution of catalytic function. *Science*, **276**, 614–617.

11. Kouchakdjian,M., Bodepudi,V., Shibutani,S., Eisenberg,M., Johnson,F., Grollman,A.P. and Patel,D.J. (1991) NMR structural studies of the ionizing radiation adduct 7-hydro-8-oxodeoxyguanosine (8-oxo-7*H*-dG) opposite deoxyadenosine in a DNA duplex. 8-Oxo-7*H*-dG(*syn*)•dA(*anti*) alignment at lesion site. *Biochemistry*, **30**, 1403–1412.
12. Brown,D.M., Hewlins,M.J.E. and Schell,P. (1968) The tautomeric state of N(4)-hydroxy- and of N(4)-amino-cytosine derivatives. *J. Chem. Soc. C*, 1925–1929.
13. Moore,M.H., Van Meervelt,L., Salisbury,S.A., Kong Thoo Lin,P. and Brown,D.M. (1995) Direct observation of two base-pairing modes of a cytosine-thymine analogue with guanine in a DNA Z-form duplex: significance for base analogue mutagenesis. *J. Mol. Biol.*, **251**, 665–673.
14. Suo,Z. and Johnson,K.A. (1997) Effect of RNA secondary structure on the kinetics of DNA synthesis catalyzed by HIV-1 reverse transcriptase. *Biochemistry*, **36**, 12459–12467.
15. Kim,S., Schroeder,C.M. and Xie,X.S. (2010) Single-molecule study of DNA polymerization activity of HIV-1 reverse transcriptase on DNA templates. *J. Mol. Biol.*, **395**, 995–1006.
16. Voytek,S.B. and Joyce,G.F. (2007) Emergence of a continuously evolving ligase ribozyme. *Proc. Natl Acad. Sci. USA*, **104**, 15288–15293.
17. Fujita,Y., Furuta,H. and Ikawa,Y. (2010) Evolutionary optimization of a modular ligase ribozyme: a small catalytic unit and a hairpin motif masking an element that could form an inactive structure. *Nucleic Acids Res.*, **38**, 3328–3339.
18. Hanes,J.W., Thal,D.M. and Johnson,K.A. (2006) Incorporation and replication of 8-oxo-deoxyguanosine by the human mitochondrial DNA polymerase. *J. Biol. Chem.*, **281**, 36241–36248.
19. Batra,V.K., Beard,W.A., Hou,E.W., Pedersen,L.C., Prasad,R. and Wilson,S.H. (2010) Mutagenic conformation of 8-oxo-7,8-dihydro-2′-dGTP in the confines of a DNA polymerase active site. *Nature Struct. Mol. Biol.*, **17**, 889–890.
20. Rasila,T.S., Pajunen,M.I. and Savilahti,H. (2009) Critical evaluation of random mutagenesis by error-prone polymerase chain reaction protocols, *Escherichia coli* mutator strain, and hydroxylamine treatment. *Anal. Biochem.*, **388**, 71–80.
21. Abbondanzieri,E.A., Greenleaf,W.J., Shaevitz,J.W., Landick,R. and Block,S.M. (2005) Direct observation of base-pair stepping by RNA polymerase. *Nature*, **438**, 460–465.
22. Christian,T.D., Romano,L.J. and Rueda,D. (2009) Single-molecule measurements of synthesis by DNA polymerase with base-pair resolution. *Proc. Natl Acad. Sci. USA*, **106**, 21109–21114.
23. Leung,D.W., Chen,E. and Goeddel,D.V. (1989) A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique*, **1**, 11–15.
24. Cadwell,R.C. and Joyce,G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Applic.*, **2**, 28–33.
25. Vartanian,J.-P., Henry,M. and Wain-Hobson,S. (1996) Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions. *Nucleic Acids Res.*, **24**, 2627–2631.
26. Biles,B.D. and Connolly,B.A. (2004) Low fidelity *Pyrococcus furiosus* DNA polymerase mutants useful in error-prone PCR. *Nucleic Acids Res.*, **32**, e176.
27. Eigen,M. (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwiss.*, **58**, 465–523.
28. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.