

FROM CONTEXT-DEPENDENCE OF MUTATIONS TO MOLECULAR MECHANISMS OF MUTAGENESIS

IGOR B. ROGOZIN

*National Center for Biotechnology Information NLM, National Institutes of Health,
Bethesda, MD 20894, USA; Institute of Cytology and Genetics RAS, Novosibirsk 630090,
Russia; rogozin@ncbi.nlm.nih.gov*

BORIS A. MALYARCHUK

*Institute of Biological Problems of the North, Far-East Branch of the Russian Academy
of Sciences, Magadan 685000, Russia; malyar@ibpn.kolyma.ru*

YOURI I. PAVLOV

University of Nebraska Medical Center, Omaha, NE 68198, USA; ypavlov@unmc.edu

LUCIANO MILANESI

*Institute of Biomedical Technologies CNR, Milano 20090, Italy;
luciano.milanesi@itb.cnr.it*

Mutation frequencies vary significantly along nucleotide sequences such that mutations often concentrate at certain positions called hotspots. Mutation hotspots in DNA reflect intrinsic properties of the mutation process, such as sequence specificity, that manifests itself at the level of interaction between mutagens, DNA, and the action of the repair and replication machineries. The nucleotide sequence context of mutational hotspots is a fingerprint of interactions between DNA and repair/replication/modification enzymes, and the analysis of hotspot context provides evidence of such interactions. The hotspots might also reflect structural and functional features of the respective DNA sequences and provide information about natural selection. We discuss analysis of 8-oxoguanine-induced mutations in pro- and eukaryotic genes, polymorphic positions in the human mitochondrial DNA and mutations in the HIV-1 retrovirus. Comparative analysis of 8-oxoguanine-induced mutations and spontaneous mutation spectra suggested that a substantial fraction of spontaneous A•T→C•T mutations is caused by 8-oxoGTP in nucleotide pools. In the case of human mitochondrial DNA, significant differences between molecular mechanisms of mutations in hypervariable segments and coding part of DNA were detected. Analysis of mutations in the HIV-1 retrovirus suggested a complex interplay between molecular mechanisms of mutagenesis and natural selection.

1. Mutation spectra and mutation hotspots

Genetic variation is a necessary prerequisite of evolution. Genomes are replicated at a level of fidelity that "determined" by deep evolutionary forces, by

high fitness, for example hotspots in the p53 gene might reflect both intrinsic mutability and selection during tumor progression [3]. Thus, study of mutation hotspots can help reveal mutagenic mechanisms, or can reveal information about the functional domains of a target protein [2,3,9].

2. Nucleotide context of mutation hotspots

2.1. Local context

Intrinsic mutation hotspots are frequently caused by mutable motifs (hotspot motifs) (reviewed in [2,3,10]). One well-known example is CpG dinucleotides which are correlated with mutation hotspots in mammalian genomes [11]. The mutational mechanism for this effect is likely to involve deamination of 5-methyl cytosine, which is frequently found at CpG dinucleotides. Thus, C•G→T•A mutations occur at CG mutable motifs (hotspot bases are underlined) due to deamination of 5-methyl cytosine followed by replication of the resulting T/G mismatch. Another well-known example of obvious mutational hotspots is hotspots of somatic mutations in immunoglobulin V genes [12]. In this case mutation hotspots are associated with RGYW/WRCY and WA/TW mutable motifs (potential hotspot sites are underlined, W = A/T) [6,7]. Usually mutation hotspots emerged at a specific position of a mutable motif, for example, only G•C bases are mutation-prone within RGYW/WRCY motifs (potential hotspot sites are underlined). Many other nucleotide sequence context effects on mutation rate have been studied and characterized, some examples are shown in the Table 1.

Table 1. Examples of mutable motifs.

Spectrum/test system/mutagen	Mutable motif	Comments
Spontaneous G•C→A•T mutations in mammalian genomes	<u>CG</u>	May result from the spontaneous deamination of 5-methylcytosine [11]
Somatic mutations in immunoglobulin V genes	<u>RGYW</u> <u>WA</u>	<u>AGYW</u> is more mutable compared to <u>GGYW</u> <u>TA</u> is more mutable compared to <u>AA</u> [7]
Hotspots of error produced by cytidine deaminase APOBEC3G	<u>GG</u>	<i>in vivo</i> experiment
8-Oxoguanine induced hotspots	<u>AA</u>	This motif was found to be mutable in pro- and eukaryotic genes
Target signal of retroposable elements	TTAAAA	LINEs and SINEs [17]

Hotspot positions are underlined. R = A or G; Y = T or C; S = G or C; W = A or T; K = G or T; M = A or C; B = T, C or G; H = A, T or C; V = A, C or G; D = A, T or G.

Alternatively, repetitive sequences such as homonucleotide runs, direct and inverted repeats and microsatellite repeats are involved in specific types of high frequency mutational events (reviewed in [13]). For these mutations, the exact DNA sequence is not critical but only the fact that a sequence motif is repeated. The theoretical basis of these observations was suggested by Streisinger and co-workers [14]: it was proposed that short deletions and insertions within homonucleotide or homopolymeric tracts arise by misalignment of DNA strands during replication. This misalignment can lead to heterogeneity in the length of homopolymeric tracts; similar arguments apply to the more complex tandemly repeated structures of microsatellites (reviewed in [2,13]). Dislocation mutagenesis is similar to misalignment mutagenesis, but involves transient strand slippage in a monotonous run of nucleotides in the primer or template strand which is followed by incorporation of the next correct nucleotide (Figure 2) [15]. This mechanism was proposed based on studies of the *in vitro* mutation spectra of DNA polymerase β [15]. Dislocation mutagenesis may also play an important role *in vivo* generating base substitution hotspots in the control region of human mitochondrial DNA [16].

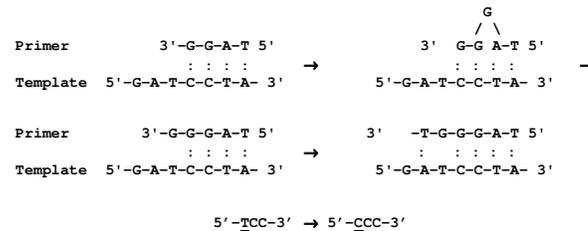


Figure 2. Dislocation mutagenesis. The primer strand dislocation, a three-nucleotide subsequence of the template strand is shown below a schematic representation of dislocation model.

There is strong evidence that short direct repeats mediate deletions and duplications in DNA [13]. Two possible mechanisms for these events are: 1) recombination between short homologous repeats or 2) DNA polymerase slippage between short repeated sequences [13]. In addition, if heteroduplexes form between imperfect direct repeats, repair of the mismatches could cause base substitutions and frameshift mutations in a concerted manner [18]. This mechanism has been suggested for some classes of spontaneous mutations in bacterial and eukaryotic genes [19] and somatic mutations in immunoglobulin genes [20].

Long inverted repeats (40-150 bases) are also particularly unstable in bacterial cells [21]. This instability is likely due to formation of hairpin

structures in single-stranded DNA and/or DNA polymerase “jumps” [13]. Correction of a quasipalindrome to a perfect inverted repeat may occur by either inter- or intramolecular strand switch [18]. Many mutations of this type have been observed in bacteria, yeast and human cells [18].

2.2. Global factors

Mutable motifs alone are not enough for emergence of hotspots, this is illustrated by the distribution of somatic mutations across AGCT sites in an artificially synthesized EPS sequence inserted into immunoglobulin gene (Figure 1). The EPS sequence contains AGCT sequences matching well-known RGYW/WRCY mutable motifs [6] repeated six times, respectively (PA-PF monomeric units in Figure 1) [4,5]. The number of mutations at G:C bases within AGCT motifs varied from 4 (the PF monomer) to 21 mutations (the PD monomer). Two significantly different classes of AGCT motifs was revealed by the CLUSTERM program (www.itb.cnr.it/webmutation/) [22], the hotspot class includes PA, PB, PC, and PD sequences (Figure 1), while another class consists from PE and PF sequences which have significantly lower frequency of mutations. This result shows that a significant heterogeneity of the mutation rate exists even in monotonously repeated AGCT motifs. Notably, the frequency of mutations in AGCTs significantly dropped at the end of the EPS sequence. This illustrates that mutation hotspots are not equivalent to mutable motifs. Emergence of hotspots is a complex process depending on high-order structures which are hard to detect.

Many factors may influence mutation frequency in a particular nucleotide sequence. However, in most cases, only local nucleotide sequence context was studied. It is likely that other higher-level features of gene or chromatin structure also have significant influence on mutation frequency of a mutable motif at a specific site. An important factor could be the rate of DNA repair. DNA repair rates vary for transcribed and non-transcribed strands of the same gene and for more and less highly expressed genes [23]. Inherent asymmetry between the two DNA strands at the replication fork could also influence mutation frequency and specificity [24]. Other potential factors include asymmetric base composition or higher order chromatin structure (reviewed by Boulikas [25]). In general, the impact of mutation rate heterogeneity is not clear, and there are some contradictions about neutral mutation rate variation across genomes. It was suggested that mutation rates differ substantially among regions of mammalian genomes [26]. However, analysis of genomic alignments of human, chimpanzee, and baboon suggested that since the time of the human-chimpanzee ancestor, there has been little or no regional variation in mutation [27]. The controversy

about mutation rate variation can be crucial for estimates of a fraction of human non-coding DNA which is under purifying selection. In general, the problem of mutation rate variation is important for understanding of fundamental problems of molecular biology and evolution. In this paper we will discuss three examples of mutation spectra analysis.

3. 8-Oxoguanine-induced mutations

Chemical agents, ionizing radiation and oxidative stress cause DNA oxidation [28]. 8-Oxoguanine (8-oxoG) is one of the most prominent base oxidation products and has been implicated in mutagenesis, carcinogenesis and aging [29]. It has been shown to cause G•C→T•A and A•T→C•G mutations *in vivo* and *in vitro*, depending whether guanine is oxidized in DNA or in the DNA precursor pools, respectively [30]. A spontaneous mutation spectrum in the *mutT* deficient *E.coli* strain (186 mutations; *lacI^d* test system) is composed almost exclusively, of A•T→C•G transversions which is in general consistent with mutagenic properties of 8-oxoGTP [31]. Hotspot context analysis of these transversions [31] using the CLUSTERM program [22] and regression trees [32] revealed AA mutable sequence (the hotspot position is underlined) (Figure 3). Comparison of the *mutT* spectrum and A•T→C•T transversions in a spectrum of spontaneous mutations in the *lacI* gene (*lacI^d* test system) [33] did not reveal significant differences between them (probability that these two spectra are different [34] $P(\chi^2) = 0.69$). Furthermore, a highly significant positive correlation was found (Kendell's τ correlation coefficient [35] = 0.65, $P < 0.01$). This result suggested that a substantial fraction of spontaneous A•T→C•T mutations in *E.coli* is caused by 8-oxoGTP in nucleotide pools.

Table 2. Comparison of A•T→C•G transversion in *lacI* gene from *mutT* and wild-type strains of *E.coli* [31, 33].

Position																			
41	81	72	64	87	168	79	189	192	195	167	83	117	96	128	177	77	141	105	54
A•T→C•G mutations in <i>mutT</i> strain																			
4	10	5	2	4	9	7	<u>23</u>	<u>18</u>	10	<u>37</u>	5	1	5	4	7	<u>20</u>	5	2	8
A•T→C•G spontaneous mutations																			
2	3	2	2	1	2	2	<u>8</u>	<u>1</u>	6	<u>10</u>	0	1	0	0	3	<u>4</u>	0	1	3

Positions of AA mutable motifs are underlined.

Reconstructed spontaneous mutations in human pseudogenes [36] (<ftp.bionet.nsc.ru/pub/biology/dbms/PSEUDO.ZIP>) were also analyzed, and the frequencies of nucleotides surrounding A•T→C•T transversions are shown in Table 3. Notably, AA and TT are the most frequent dinucleotide combinations. Such excess is statistically significant ($P(\chi^2) < 0.01$) as compared to dinucleotide

frequencies in reconstructed ancestral sequences (Table 3). A substantially higher frequency of A in the position +1 was observed for A/C SNPs ($A \gg T \approx C \gg G$) [37] which is consistent with the AA mutable motif.

Table 3. Frequencies of bases in position +1 and -1 in a set of spontaneous A•T → C•G transversions found in human pseudogenes [36].

Substitution	Position -1				Position +1			
	A	T	G	C	A	T	G	C
A→C T→G	0.25	<u>0.32</u>	0.21	0.22	<u>0.35</u>	0.24	0.17	0.24
Expected	0.25	0.22	0.22	0.31	0.26	0.23	0.29	0.22

Expected values (frequencies of AN and NT dinucleotides, N=A/T/G/C) were calculated in ancestral sequences used for reconstruction of spontaneous mutations [36].

These results suggested that mutagenesis due to 8-oxoG is significantly influenced by nearest neighboring bases and the context is quite evolutionarily stable. The revealed context properties could be fingerprints of interactions between DNA and repair/replication/modification enzymes. There is, at present, no data on evolutionary conservation of context specificity of such interactions between pro- and eukaryotes [3]. Thus, the alternative view that the revealed context properties reflect intrinsic properties of interactions between 8-oxoG and DNA might be the current model of choice.

4. Human mitochondrial DNA

Most of mitochondrial DNA (mtDNA) variability studies have been based on sequence variation of the fast-evolving major non-coding (or control) region, which spans 1122 bases between the tRNA genes for proline (tRNA^{Pro}) and phenylalanine (tRNA^{Phe}) [38]. The majority of mutations are concentrated in two hypervariable segments, HVS I (positions 16024-16365) and HVS II (positions 73-340) [39]. Our analysis of phylogenetically reconstructed mutation spectra of the mtDNA HVS I and II regions has suggested that the dislocation mutagenesis (Figure 2) plays an important role for generating base substitutions in these regions [16,40]. However, an impact of the dislocation mutagenesis on the remaining part of mtDNA remains unclear.

To study spontaneous base substitutions in regions of human mtDNA other than HVS I and II, we reconstructed mutation spectra of the mtDNA region containing ND3, tRNA^{Arg}, and ND4L genes (positions 10171-10659) using published data on polymorphisms in various human populations. We have analyzed different phylogenetic haplogroups of mtDNA revealed by means of median network analysis [39] (<http://fluxus-engineering.com>, the Network 3.1 program). We used only the published population data comprising mtDNA

sequences with known phylogenetic status as described by Malyarchuk and co-workers [16,40]. The reconstructed mutation spectrum contained 93 mutations in 489 bases.

The dislocation mutagenesis model was analyzed for the reconstructed mutation spectrum using a Monte-Carlo procedure [40]. No statistically significant support for this model was found ($P(W \leq W_{\text{random}}) = 0.68$). This result suggested that the dislocation mutagenesis does not play an important role for generating substitutions in the coding regions of mtDNA. A higher rate of molecular evolution in HVS regions than in the remaining part of mtDNA can be explained by differences in either mutation or selection pressure [39]. The observed differences in dislocation mutagenesis suggested that a higher rate of mutations in HVS regions is caused by intrinsic properties of mutations. HVS regions are associated with initiation/termination of mtDNA replication and RNA/DNA transition, and one of these processes may be error-prone for the DNA strand dislocation mutagenesis.

5. Hypermutation in HIV-1

Genomic heterogeneity is a hallmark of retroviruses, especially of HIV, which helps virus to escape the host immune system. Hypermutability is linked to pathogenicity and it was generally attributed to relatively low fidelity of reverse transcriptase [41]. A high rate of mutagenesis in retroviruses, in addition to being a way to elude the immune system, may lead to their low viability. It was hypothesized that even a relatively small increase of the mutation rate in a retrovirus will lead to the accumulation of many deleterious mutations and virus extinction. Indeed, the treatment of HIV-infected human cells by mutagenic nucleoside analogs resulted in the loss of viral replicative potential [42]. Recent discoveries suggest that nature already exploited this mechanism for protection from retroviruses. The unique cellular gene CEM15 was found that conferred resistance to HIV. Its antiviral action could be overcome by the presence of virion infectivity factor (Vif), encoded by the viral genome [43]. CEM15 appeared to be identical to the cytidine deaminase APOBEC3G. It is already known that APOBEC3G is a strong mutator when expressed in *E. coli*, suggesting that it could deaminate cytosines in DNA [44]. The viruses without Vif experienced hypermutation and all these mutations were transitions that could be explained by deamination of a (-) DNA strand of the virus [45]. The current model for APOBEC3G antiviral action proposes that the deaminase is packaged into *Vif*⁻ HIV virions and induces massive deamination in the viral (-) strand [45,46]. This deamination can lead to hypermutagenesis or the destruction of the viral genome during repair of uracil [46].

comparisons of mutational spectra produced in the same sequence using statistical methods to compare two or more spectra based on an exact or pseudo-probability test (a Monte Carlo modification of the exact test) [34] or correlation analyses [35,48]. This strategy was employed in studies of somatic hypermutation in immunoglobulin genes. Mutational spectra analysis of 15 immunoglobulin genes suggested that the WA mutable motif is a universal descriptor of somatic hypermutation at A•T pairs. Highly mutable sites, "hotspots", that matched WA were preferentially found in one DNA strand. Analysis of base-substitution hotspots in DNA polymerase η spectra showed that 33 of 36 hotspots in the human polymerase η spectrum conformed to the WA consensus. A correlation between the WA motif and the error specificity of human DNA polymerase η suggested that this polymerase contributes to the WA hotspots [7]. Additional analysis of this correlation using the same mouse immunoglobulin target sequence for *in vivo* and *in vitro* spectrum generation combined with studies of mutable motifs and frequencies of substitutions greatly improved the power of comparisons, allowing use of the pairwise and multiple regression analysis [48]. Obtained results supported the hypothesis that DNA polymerase η contributes to somatic hypermutation of immunoglobulin genes at A•T pairs via short patches of low fidelity DNA synthesis of both strands, but with a preference for the non-transcribed strand [48].

It should be emphasized that the context of hotspots may be very helpful in deep understanding of underlying molecular mechanisms of mutagenesis. However, the determinants of mutation frequency and specificity are complex and there are many analytical methods for their study. The most reliable results can be obtained if several methods are combined or used sequentially and if many different sources of information are considered. It is a challenging task to analyze mutation spectra, and in some cases, the effort will primarily be descriptive in nature. However, in several well-documented studies, the analysis of mutation spectra has contributed substantially to understanding molecular mechanisms of mutagenesis. As analytical methods continue to be developed, more theoretical and experimental studies will contribute insights into the complex process of mutagenesis.

Acknowledgements

We thank Eugene Koonin and Galina Glazko for helpful comments on the manuscript. This work was partially supported by RFBR, by Far-East Branch of the Russian Academy of Sciences, MIUR "Functional Genomics" 449/97, "High

through put Grid computational platforms for virtual scalable organizations” and “Bioinformatics” Italian FIRB projects.

References

1. J.W. Drake, B. Charlesworth, D. Charlesworth and J.F. Crow, *Genetics* **148**, 1667 (1998).
2. J.H. Miller, *Annu. Rev. Genet.* **17**, 215 (1983).
3. I.B. Rogozin, V.N. Babenko, L. Milanese and Y.I. Pavlov, *Brief. Bioinform.* **4**, 210 (2003).
4. U. Storb, E.L. Klotz, J. Hackett, K. Kage, G. Bozek and T.E. Martin, *J. Exp. Med.* **188**, 689 (1998).
5. N. Kim, G. Bozek, J.C. Lo and U. Storb, *J. Exp. Med.* **190**, 21 (1999).
6. I.B. Rogozin and N.A. Kolchanov, *Biochim. Biophys. Acta* **1171**, 11 (1992).
7. I.B. Rogozin, Y.I. Pavlov, K. Bebenek, T. Matsuda and T.A. Kunkel, *Nat. Immunol.* **2**, 530 (2001).
8. S. Benzer, *Proc. Natl. Acad. Sci. USA* **47**, 403 (1961).
9. A. Siepel and D. Haussler, *Mol. Biol. Evol.* **21**, 468 (2004).
10. M. Zavolan and T.B. Kepler, *Curr. Opin. Genet. Dev.* **11**, 612 (2001).
11. D.N. Cooper and H. Youssoufian, *Hum. Genet.* **78**, 151 (1988).
12. A.G. Betz, C. Rada, R. Pannell, C. Milstein and M.S. Neuberger, *Proc. Natl. Acad. Sci. USA* **90**, 2385 (1993).
13. D.A. Gordenin and M.A. Resnick, *Mutat Res* **400**, 45 (1998).
14. G. Streisinger, Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi and M. Inouye, *Cold Spring Harb. Symp. Quant. Biol.* **31**, 77 (1966).
15. T.A. Kunkel, *J. Biol. Chem.* **260**, 5787 (1985).
16. B.A. Malyarchuk, I.B. Rogozin, V.B. Berikov and M.V. Derenko, *Hum. Genet.* **111**, 46 (2002).
17. J. Jurka and P. Klonowski, *J. Mol. Evol.* **43**, 685 (1996).
18. L.S. Ripley, *Proc. Natl. Acad. Sci. USA* **79**, 4128 (1982).
19. G.B. Golding and B.W. Glickman, *Proc. Natl. Acad. Sci. USA* **82**, 8577 (1985).
20. N.A. Kolchanov, V.V. Solovyev and I.B. Rogozin, *Dokl. Akad. Nauk* **281**, 994 (1985).
21. D.M. Lilley, *Nature* **292**, 380 (1981).
22. G.V. Glazko, L. Milanese and I.B. Rogozin, *J. Theor. Biol.* **192**, 475 (1998).
23. P.C. Hanawalt, *Environ. Health Perspect.* **76**, 9 (1987).
24. X. Veaute and R.P. Fuchs, *Science* **261**, 598 (1993).

25. T. Boulikas, Evolutionary consequences of nonrandom damage and repair of chromatin domains, *J. Mol. Evol.* **35**, 156 (1992).
26. K.H. Wolfe, P.M. Sharp and W.H. Li, *Nature* **337**, 283 (1989).
27. M.T. Webster, N.G. Smith and H. Ellegren, *Mol. Biol. Evol.* **20**, 278 (2003).
28. R. Adelman, R.L. Saul and B.N. Ames, *Proc. Natl. Acad. Sci. USA* **85**, 2706 (1988).
29. B.N. Ames, *Mutat. Res.* **214**, 41 (1989).
30. M.L. Michaels and J.H. Miller, *J. Bacteriol.* **174**, 6321 (1992).
31. R.G. Fowler and R.M. Schaaper, *FEMS Microbiol. Rev.* **21**, 43 (1997).
32. V.B. Berikov and I.B. Rogozin, *Bioinformatics* **15**, 553 (1999).
33. A.R. Oller and R.M. Schaaper, *Genetics* **138**, 263 (1994).
34. N.F. Cariello, W.W. Piegorsch, W.T. Adams and T.R. Skopek, *Carcinogenesis* **15**, 2281 (1994).
35. V.N. Babenko and I.B. Rogozin, *Biofizika* **44**, 632 (1999).
36. M.A. Pozdnyakov, I.B. Rogozin, V.N. Babenko and N.A. Kolchanov, *Dokl. Akad. Nauk.* **356**, 566 (1997).
37. Z. Zhao and E. Boerwinkle, *Genome Res.* **12**, 1679 (2002).
38. S. Anderson, A.T. Bankier, B.G. Barrell, M.H. de Bruijn, A.R. Coulson, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, P.H. Schreier, A.J. Smith, R. Staden and I.G. Young, *Nature* **290**, 457 (1981).
39. L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes and A.C. Wilson, *Science* **253**, 1503 (1991).
40. B.A. Malyarchuk and I.B. Rogozin, *Ann. Hum. Genet.* **68**, 324 (2004).
41. J.D. Roberts, K. Bebenek and T.A. Kunkel, *Science* **242**, 1171 (1988).
42. L.A. Loeb, J.M. Essigmann, F. Kazazi, J. Zhang, K.D. Rose and J.I. Mullins, *Proc. Natl. Acad. Sci. USA* **96**, 1492 (1999).
43. A.M. Sheehy, N.C. Gaddis, J.D. Choi and M.H. Malim, *Nature* **418**, 646 (2002).
44. R.S. Harris, S.K. Petersen-Mahrt and M.S. Neuberger, *Mol. Cell* **10**, 1247 (2002).
45. D. Lecossier, F. Bouchonnet, F. Clavel and A.J. Hance, *Science* **300**, 1112 (2003).
46. R.S. Harris, K.N. Bishop, A.M. Sheehy, H.M. Craig, S.K. Petersen-Mahrt, I.N. Watt, M.S. Neuberger and M.H. Malim, *Cell* **113**, 803 (2003).
47. R.S. Harris, A.M. Sheehy, H.M. Craig, M.H. Malim and M.S. Neuberger, *Nat. Immunol.* **4**, 641 (2003).
48. Y.I. Pavlov, I.B. Rogozin, A.P. Galkin, A.Y. Aksenova, F. Hanaoka, C. Rada and T.A. Kunkel, *Proc. Natl. Acad. Sci. USA* **99**, 9954 (2002).