doi: 10.1046/j.1529-8817.2004.00099.x

Mutagenesis by Transient Misalignment in the Human Mitochondrial DNA Control Region

B. A. Malyarchuk[1,*] and I. B. Rogozin[2,3]

[1]*Institute of Biological Problems of the North, Far-East Branch of the Russian Academy of Sciences, Portovaya str. 18, 68500 Magadan, Russia*

[2]*Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia*

[3]*National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda MD 20894, USA*

Summary

To study spontaneous base substitutions in human mitochondrial DNA (mtDNA), we reconstructed the mutation spectra of the hypervariable segments I and II (HVS I and II) using published data on polymorphisms from various human populations. Classification analysis revealed numerous mutation hotspots in HVS I and II mutation spectra. Statistical analysis suggested that strand dislocation mutagenesis, operating in monotonous runs of nucleotides, plays an important role in generating base substitutions in the mtDNA control region. The frequency of mutations compatible with the primer strand dislocation in the HVS I region was almost twice as high as that for template strand dislocation. Frequencies of mutations compatible with the primer and template strand dislocation models are almost equal in the HVS II region. Further analysis of strand dislocation models suggested that an excess of pyrimidine transitions in mutation spectra, reconstructed on the basis of the L-strand sequence, is caused by an excess of both L-strand pyrimidine transitions and H-strand purine transitions. In general, no significant bias toward parent H-strand-specific dislocation mutagenesis was found in the HVS I and II regions.

Keywords: mitochondrial DNA, spontaneous substitution, dislocation mutagenesis, context, mutation hotspot

Introduction

The origin of mitochondrial DNA (mtDNA) mutations, both inherited and somatic, is one of the most important questions in evolutionary and population mitochondrial genetics, as well as in molecular medicine, since mtDNA mutations are related to a variety of human degenerative diseases and cancer (Wallace, 1999; Copeland *et al.* 2002). The human mitochondrial genome is a small (16569 base pairs) double-stranded circular molecule and is strictly maternally inherited (Giles *et al.* 1980; Anderson *et al.* 1981). Human mtDNA evolves rapidly, at a rate 5-10 times higher than single-copy nuclear genes (Brown, 1980). Multiple copies of mtDNA are present in cells and within each mitochondrion. The copy number of mtDNA is typically 100-10000 copies/cell, depending on cell type (Wallace, 1999). Several possible factors that may cause the high mutation rate in mtDNA including inefficient DNA repair systems, a lack of DNA protective proteins, and continuous exposure to the mutagenic effects of reactive oxygen species (ROS) generated by oxidative phosphorylation (Richter *et al.* 1988; Copeland *et al.* 2002). However, recent studies have shown that mammalian mitochondria are well equipped to conduct base-excision repair (Bogenhagen, 1999; Dianov *et al.* 2001) and probably mismatch repair (Mason *et al.* 2003). A lack of histones covering the entire mtDNA molecule is probably equilibrated by the presence of the multifunctional protein TFAM (mtTFA, mitochondrial transcription factor A), which may cover the entire mtDNA, but preferentially binds to an active promoter region and cruciform DNA

*Corresponding author: Dr. Boris A. Malyarchuk, Genetics Laboratory, Institute of Biological Problems of the North, Portovaya str., 18, 685000 Magadan, Russia. Fax/Phone: 7 41322 34463. E-mail: malyar@ibpn.kolyma.ru

 © University College London 2004

structures (Kang & Hamasaki, 2002). Ribonucleotides covering the parent H-strand, which remains single-stranded for a long time during mtDNA replication and consequently might be damaged at a higher rate than the L-strand (Reyes *et al.* 1998), may also carry out a protective function, because it was recently found that replication intermediates contain large regions of RNA:DNA hybrid as a result of the incorporation of ribonucleotides on the lagging L-strand during mtDNA replication (Yang *et al.* 2002). Another factor explaining the high mutation rate in mtDNA is spontaneous errors arising during DNA replication. Although the fidelity of mtDNA polymerase is very high, it has been shown to cause frameshift errors in homopolymeric runs (Longley *et al.* 2001). It has further been that defects of nuclear genes responsible for mtDNA replication and maintenance cause the accumulation of mtDNA mutations (Copeland *et al.* 2003).

Most mtDNA variability studies have been examining 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 ($tRNA^{Phe}$) (Anderson *et al.* 1981). This region is highly polymorphic, and the majority of mutations are concentrated in two hypervariable segments, HVS I (positions 16024-16365) and HVS II (positions 73-340). Results of phylogenetic studies have suggested a complex pattern of mtDNA control region evolution. It was found that base composition in the HVS I and II regions is not uniform, transitions occur with higher frequencies compared to transversions, the number of pyrimidine transitions in the L-strand exceeds the number of purine transitions, and substitution rates vary among nucleotide positions (Hasegawa *et al.* 1993; Wakeley, 1993; Excoffier & Yang, 1999; Meyer *et al.* 1999; Heyer *et al.* 2001; Pesole & Saccone, 2001; Malyarchuk *et al.* 2002b).

Strand slippage in repetitive sequences may result in base substitutions by the transient misalignment dislocation mechanism. This model suggests that transient strand slippage in a monotonous run of nucleotides in the primer or template strand is followed by incorporation of the next correct nucleotide (Kunkel, 1985). Our analysis of phylogenetically reconstructed mutation spectra (distributions of mutations along analyzed sequences) of the mtDNA HVS I and II regions has suggested that dislocation mutagenesis plays an important role in generating base substitutions in mtDNA (Malyarchuk *et al.* 2002b). Additionally, it was found that next-nucleotide effects and dislocation mutagenesis may contribute to the formation of mtDNA mutations in patients with alterations of nucleoside metabolism (Nishigaki *et al.* 2003). Therefore the origin of mtDNA mutation hotspots may, to a great extent, depend on the contextual properties of the mtDNA. To study spontaneous base substitutions in human mtDNA, we reconstructed mutation spectra for the HVS I and II regions using published data on polymorphisms in various human populations and analyzed them by means of the dislocation mutagenesis model (Kunkel, 1985; Kunkel & Soni, 1988).

Materials and Methods

Reconstruction of Mutation Spectra

To determine mutations in the mtDNA control region, we have analysed different phylogenetic haplogroups of mtDNA revealed by means of median network analysis (Bandelt *et al.* 1995; http://fluxus-engineering.com for Network 3.1 program). We only used published population data comprising both the HVS I and/or HVS II nucleotide sequences and additional RFLP or coding-region information for each haplotype. Coding region variation was used to assign control region sequences to the phylogenetic haplogroups and place them in the reconstructed mtDNA phylogenetic tree (Macaulay *et al.* 1999; Maca-Meyer *et al.* 2001; Finnila *et al.* 2001; Herrnstadt *et al.* 2002; Kivisild *et al.* 2002; Salas *et al.* 2002; Yao *et al.* 2002). Using the human mtDNA nomenclature (Richards *et al.* 1998; Macaulay *et al.* 1999), each major clade (or mtDNA haplogroup) and nested subclade (or subhaplogroup) of the mtDNA tree was denoted with the corresponding Roman numerals.

The HVS I data set comprised 7482 sequences (between positions 16092-16365) belonging to 90 continental-specific mtDNA haplogroups and subgroups. This data set includes 3834 HVS I sequences from 28 West Eurasian haplogroups and subgroups: H, HV^*, pre-V, pre-HV, R^*, T1, T^*, J^*, J1a, J1b, J2, K, U^*, U1, U2, U3, U4, U5, U7, U8a, U8b, N1a, N1b, N1c, N^*, I, W, X (according to data of Richards *et al.*

© University College London 2004

2000); 801 sequences from 34 East Eurasian haplogroups and subgroups: C, Z, M8a, D* (including D4), D5, G2, G3, G4, E, M*, M7*, M7b, M7c, M9, M10, A, N9a, N2, N*, Y, R9a, R*, F*, F1a, F1b, F1c, F2, B*, B4*, B4a, B4b, B5*, B5a, B5b (according to data from Derbeneva *et al.* 2002a, 2002b; Kivisild *et al.* 2002; Yao *et al.* 2002; Derenko *et al.* 2003); and 2847 sequences from 28 African haplogroups and subgroups: L1a1, L1a2, L1b, L1c*, L1c1, L1c2, L1c3, L1d, L1e, L2a*, L2a1a, L2a1b, L2b, L2c, L2d1, L2d2, L3b1, L3b2 (including L3b*), L3d, L3e1, L3e2, L3e3, L3e4, L3f*, L3f1, L3g, U6, M1 (according to the data from Salas *et al.* 2002). Note that haplogroups U6 and M1 were included in the African-specific data set, because haplogroup U6 has predominantly been found in North Africans (Macaulay *et al.* 1999) and haplogroup M1 may have originated in East Africa (Quintana-Murci *et al.* 1999).

The HVS II data set was represented by 1703 individual sequences (between positions 72-297) belonging to 71 mtDNA haplogroups and subgroups. This data set includes 1002 HVS II sequences from 27 West Eurasian haplogroups and subgroups: H, HV*, pre-V, pre-HV, R*, T1, T*, J*, J1a, J1b, J2, K, U*, U1, U2, U3, U4, U5, U7, U8a, U8b, N1a, N1b, N1c, I, W, X (according to data of Hoffman *et al.* 1997; Finnila *et al.* 2001; Derbeneva *et al.* 2002a; 2002b; Malyarchuk *et al.* 2002a; Derenko *et al.* 2003); 496 sequences from 33 East Eurasian haplogroups and subgroups: C, Z, M8a, D* (including D4), D5, G2, G3, G4, E, M*, M7*, M7b, M7c, M9, M10, A, N9a, N*, Y, R9a, R*, F*, F1a, F1b, F1c, F2, B*, B4*, B4a, B4b, B5*, B5a, B5b (according to data of Finnila *et al.* 2001; Derbeneva *et al.* 2002a; 2002b; Kivisild *et al.* 2002; Malyarchuk *et al.* 2002a; Yao *et al.* 2002; Derenko *et al.* 2003); and 205 sequences from 11 African haplogroups and subgroups: L1a, L1b, L1c, L2a, L2b, L2c, L2d, L3*, L3e1, L3e2, L3e3 (according to data from Alves-Silva *et al.* 2000; Chen *et al.* 2000; Bandelt *et al.* 2001; Maca-Meyer *et al.* 2001; Torroni *et al.* 2001).

Mutations were inferred from the mtDNA median networks constructed, following the guidelines of Bandelt *et al.* (2000), and verified by comparison with the published networks. Parallel mutations in mtDNA were inferred by revealing variable positions in which identical mutations arose independently in different mitochondrial haplogroups or subgroups, as described by Macaulay *et al.* (1999), Finnila *et al.* (2001), Malyarchuk & Derenko (2001) and Herrnstadt *et al.* (2002). In all cases we have studied mutation spectra represented by nucleotide substitutions; therefore, nucleotide positions that showed point insertions or deletions were excluded from the analysis. All mutations are described using the L-strand orientation and numbered according to the mtDNA Cambridge Reference Sequence (CRS; Anderson *et al.* 1981; Andrews *et al.* 1999).

Hotspot Prediction

The general principle of mutation hotspot prediction in this study was based on a threshold (Sh) value for the number of mutations in a mutable site. All sites with the number of mutations greater than or equal to Sh were defined as hotspots. The threshold value and the resulting hotspot sites were defined for each mutation spectrum separately, using the CLUSTERM program (www.itb.cnr.it/webmutation/; Glazko *et al.* 1998). Briefly, this program decomposes a mutation spectrum into several homogeneous classes of sites, with each class approximated by a Poisson distribution. Variations in mutation frequencies among sites of the same class are random by definition (mutation probability is the same for all sites within a class), but differences between classes are statistically significant. A class (or classes) with the highest mutation frequency is(are) called a hotspot class(es). Each site has a probability P(C) assigned to a class C. Sites with $P(C_{hotspot}) \geq 0.95$ in hotspot class $C_{hotspot}$ are defined as hotspot sites. This approach ensures that the assignment is statistically significant and robust (Glazko *et al.* 1998; see Rogozin *et al.* 2001 for details).

Statistical Analysis of Dislocation Model

We performed statistical analyses to assess the likelihood of dislocation mutagenesis using a Monte Carlo procedure (Malyarchuk *et al.* 2002b). This approach takes into account the frequencies of substitutions for each nucleotide, the possibility of multiple mutations in a site, and the context of the mutating sites. The Monte Carlo simulation was run with weighted sites, with the weight W_j of a site j defined as the number of substitutions in this site which are compatible with the analysed

 © University College London 2004

dislocation model. For example, for the well-known hotspot of T-to-G errors (7 mutations) induced by the rat DNA polymerase β *in vitro* in the sequence 5′-GTTTT-3′ (the hotspot is underlined; Kunkel, 1985), $W_j = 7$. Weights W_j were summed for all sites in the analysed sequence, resulting in a total weight W. A distribution of total weights W_{random} was calculated for 10,000 sequences with randomly shuffled mutation distributions. Each of the resulting random mutation spectra contained the same number of mutations as the observed spectrum, with the same distribution of mutations over randomly chosen sites. The distribution of W_{random} was used to calculate the probability $P_{W \leq Wrandom}$. This probability is equal to the fraction of random spectra in which W_{random} is the same or greater than W. Low probability values ($P_{W \leq Wrandom} \leq 0.05$) indicate a good correspondence between the mutation data and the analyzed dislocation model.

Results

Substitution Frequencies and Strand Asymmetry

Reconstructed mutation spectra in the HVS I and II regions are shown in Figures 1 and 2, respectively. The first reconstructed spectrum includes 2212 substitutions in 202 variable nucleotide positions, while the second

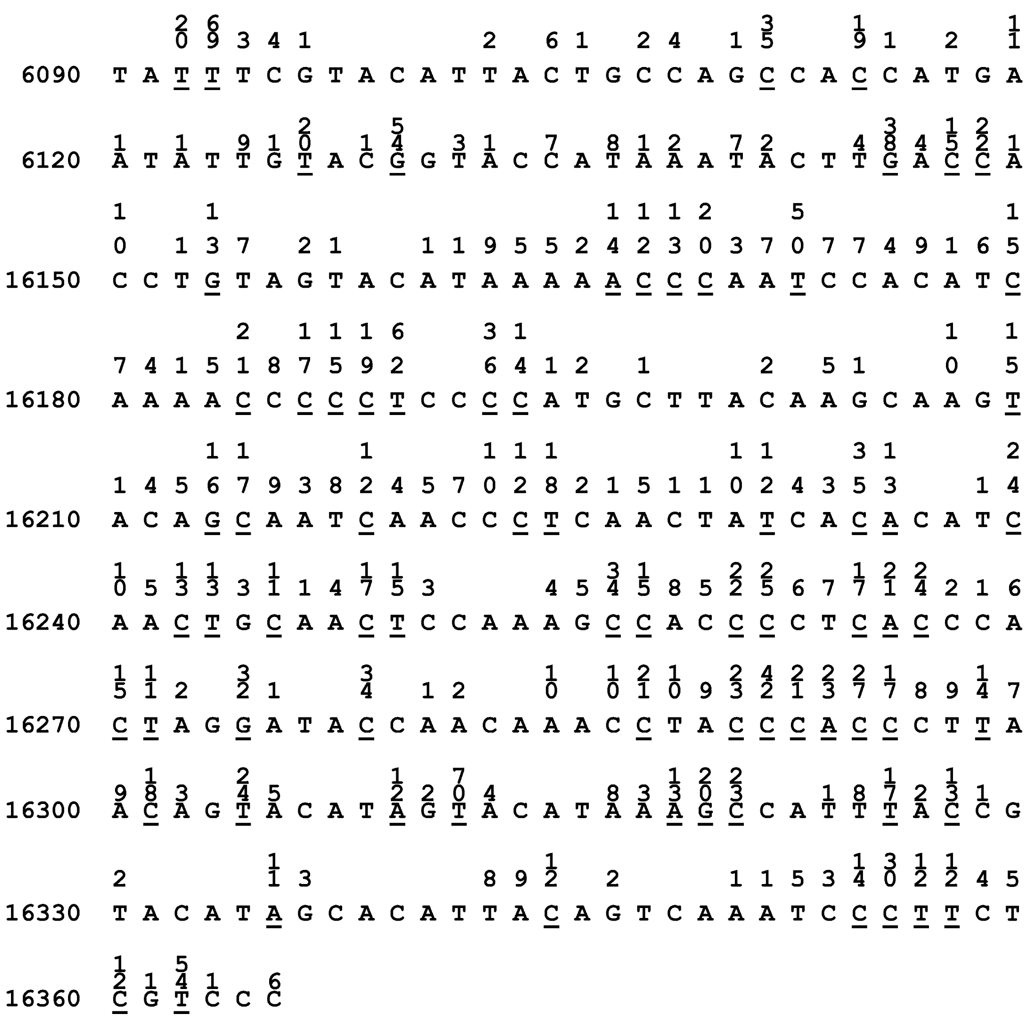

Figure 1 The reconstructed HVS I mutation spectrum; predicted hotspots are underlined, numbers above the sequence are the number of mutations (both transitions and transversions) at that position. Mutations are shown relative to the Cambridge reference sequence (Anderson *et al.* 1981).

© University College London 2004

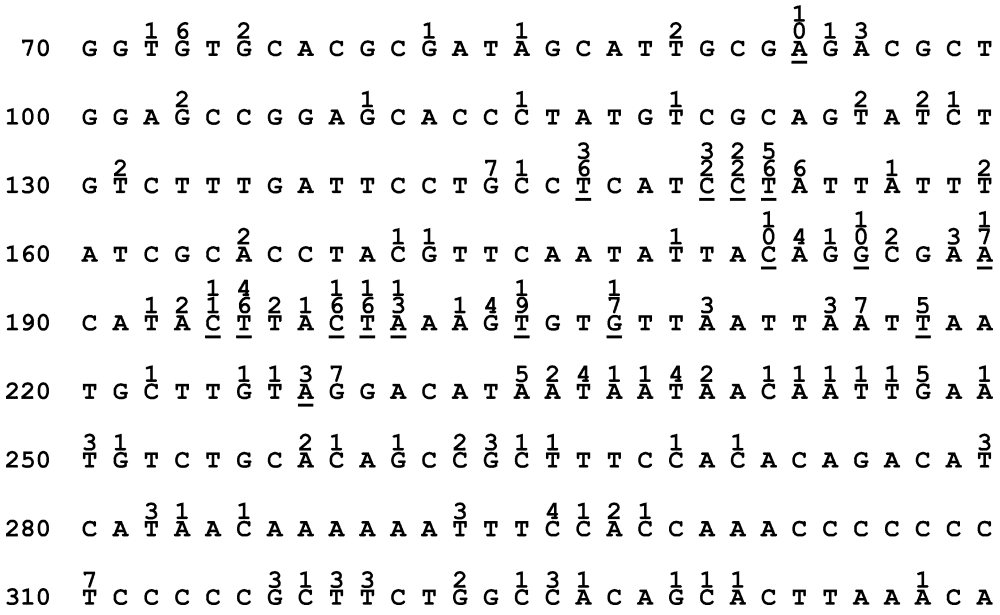

Figure 2 The reconstructed HVS II mutation spectrum; predicted hotspots are underlined, numbers above the sequence are the number of mutations at that position. Mutations are shown relative to the reference sequence, differing from the Cambridge mtDNA sequence at nucleotides 73 and 263.

Table 1 Frequencies of transitions and transversions in the HVS I and II regions

	HVS I	HVS II
Transitions (s)		
C → T	944	110
T → C	638	213
G → A	188	71
A → G	294	94
Transversions (v)		
A → C	32	5
C → A	45	0
T → G	5	1
G → T	0	1
A → T	25	0
T → A	3	4
G → C	7	1
C → G	31	0
s/v	14.0	40.7

spectrum was smaller, with 500 substitutions in 92 variable positions. The frequencies of the different types of base substitutions in both spectra are not equal; transitions constitute 93.3% and 97.6% of all mutations in the HVS I and II spectra, respectively (Table 1). Estimation of the pyrimidine-purine transition ratios shows that there is a strong bias toward transitions between pyrimidines both in the HVS I and II mutation spectra, with the values of the pyrimidine-purine transition parameter varying from 3.28 in HVS I to 1.96 in HVS II. Among transitions, the most frequent in the HVS I spectrum were C-to-T, while in the HVS II spectrum, the occurrence of T-to-C transitions was twice as high as C-to-T substitutions. For transversions, only the HVS I spectrum demonstrates a considerable amount of this type of substitution, with the majority of changes being from A to Y and from C to R.

There is a good correspondence between base composition and the frequency of variable nucleotide positions: the higher the content of a certain nucleotide, the higher the frequency of changed nucleotides observed (Table 2). The correlation between base content and frequency of variable positions was highly significant ($P < 0.01$) for both the HVS I and HVS II data sets ($r = 0.99$ and 0.96, respectively). Nevertheless, the average number of mutations (both transitions and transversions) per site varies considerably, with the lowest values observed for adenine, despite the high content of this base in the L-strand of HVS I and HVS II sequences. This conclusion is also valid when one compares the distributions of the variable nucleotide positions with the number of independent mutations occurring there (Table 1 and 2).

 © University College London 2004

Table 2 Base composition, frequency of variable nucleotide positions and number of mutations per site in the mutation spectra reconstructed on the basis of the L-strand of the mtDNA control region

	HVS I (L = 274)			HVS II (L = 226)		
Base	Base composition (%)	Frequency of variable nucleotide positions (%)	Number of mutations per variable nucleotide	Base composition (%)	Frequency of variable nucleotide positions (%)	Number of mutations per variable nucleotide
T	21.2	16.4	14.4	29.2	11.5	8.4
C	35.0	27.4	13.6	23.5	8.9	5.5
A	34.7	23.7	5.4	31.0	12.4	3.5
G	9.1	5.8	12.2	16.4	8.0	4.1

L is the length of the mtDNA fragment analyzed.

Mutation Hotspots

Analysis of the mutation spectrum in the HVS II region using CLUSTERM revealed four classes of sites. The first class includes obvious "cold" sites, with the number of substitutions varying from 0 to 3; the second class includes sites with the number of mutations from 0 to 10; the third class includes sites with the number of mutations from 7 to 20; the fourth class includes obvious hotspot sites where the number of mutations varied from 32 to 56. Differences between the observed and the expected distributions (a mixture of four Poisson distributions) were statistically insignificant. Since the members of the third class (P(site in 3) $\geq$ 0.95) were sites with more than 9 mutations, 10 was therefore chosen as the threshold value Sh for hotspot sites (these sites are underlined in Figure 2). Only 15 nucleotide positions out of 226 analyzed in the HVS II region were identified as hotspot sites using the classification analysis. Among them, the most frequent sites, which have undergone 32-56 changes, were 146, 150, 152 and 195. Sites 93, 151, 182, 185, 189, 194, 198, 199, 200, 204 and 207 were found to be very fast, with the means of the number of changes per site equal to 10-20.

Previous analysis of the HVS I mutation spectrum revealed four classes of sites, with 11 mutations used as the hotspot threshold (Malyarchuk *et al.* 2002b). The present study shows that 27% of sites in the HVS I region (74 out of 274 sites analyzed) should be classified as hotspots (these sites are underlined in Figure 1 and shown in Table 3). Among them, the fastest were sites 16093, 16129, 16172, 16189, 16291, 16311 and 16362, which have experienced more than 40 changes per site.

Table 3 The hotspot positions (>10 changes per site) in the HVS I region (within 16092-16365)

Nucleotide positions (+16000)
092, 093, 111, 114, 126, 129, 145, 147, 148, **153**, **166**, **167**, **168**, **169**, 172, **179**, **184**, 186, 187, **188**, 189, 192, 193, 209, 213, 214, **218**, 223, **224**, 231, 234, 235, 239, 242, **243**, 245, **248**, 249, 256, 257, 260, 261, 264, 265, 266, 270, **271**, 274, 278, 287, 290, 291, 292, 293, 294, 295, 298, 301, 304, 309, 311, **318**, 319, 320, 325, 327, 335, **344**, 354, 355, 356, 357, 360, 362

Positions shown in bold were not previously described as the speedy transitions according to Bandelt *et al.* (2002).

It is noteworthy that almost all of these sites have already been identified as fast sites in phylogenetic and familial studies (Hasegawa *et al.* 1993; Wakeley, 1993; Excoffier & Yang, 1999; Meyer *et al.* 1999; Heyer *et al.* 2001; Pesole & Saccone, 2001; Bandelt *et al.* 2002). Thus, hotspot sites in the HVS I are now well defined, and results of the present and other analyses give a reliable description of variability in the HVS I region.

Dislocation Mutagenesis

It has been shown that transient misalignment dislocation mutagenesis operating in monotonous runs of nucleotides might play an important role in generating base substitutions in the mtDNA control region, and define its contextual properties (Malyarchuk *et al.* 2002b). This model suggests that transient strand slippage in a homonucleotide run in the primer or template strand is followed by incorporation of the next correct nucleotide (Kunkel, 1985). Present analysis of the HVS I and II spectra revealed that many base substitutions

© University College London 2004

are consistent with the dislocation model. This model may explain the origin of 23.4% mutations (517 out of 2212) found in 34.7% of variable nucleotide positions in the HVS I (70 out of 202 sites) and of 49.4% mutations (247 out of 500) in 27.2% of variable positions (25 out of 92) in the HVS II region. For each hypervariable region, transitions predominate over transversions, being found with a ratio of 448:69 (s/v = 6.49) and 247:0 in HVS I and II, respectively. As for hotspot distribution, the model of dislocation mutagenesis can explain the origin of 16% of hotspot sites (with the number of changes per site >10) in the HVS I region, whereas in the HVS II 40% of hotspot sites (6 out of 15, including the most variable positions 146, 150, 152, and 195) may be caused by the strand dislocation model.

Recently, we have found that only the primer strand dislocation model has statistically significant support in both HVS spectra (Malyarchuk *et al.* 2002b). The present study partly supports this conclusion; however, important new details were revealed from analysis of the strand dislocation model itself. It is well established that mtDNA replication starts in the right domain of the control region (including the HVS II region) with the expansion of the displacement loop (D-loop), a stable triple-stranded structure (Clayton, 1982). During the D-loop formation, the parental H-strand is displaced by the nascent H-strand, remaining in a single-stranded state until L-strand DNA synthesis is initiated in the opposite direction to that of the H-strand. Thus, human mtDNA replicates by an unusual asynchronous transcription-primed mechanism that is initiated at the origin of H-strand replication (Shadel & Clayton, 1997). The mtDNA replication model suggests that the higher rate of transition between pyrimidines observed in the L-strand mutation spectra of the control region may be a result of both a high substitution rate of pyrimidine transitions on the L-strand, and purine transitions on the H-strand, or one may speculate that the observed high rate of pyrimidine transitions on the L-strand is a result of transitions between purines that occurred on the parental H-strand due to the probable higher mutability of purines on the single-stranded H-strand during mtDNA replication (Reyes *et al.* 1998; Tamura, 2000). However, this classical asymmetric replication model has recently been questioned by the finding that mammalian mtDNA replication is initially bidirectional, but after fork arrest near OriH, replication is restricted to one direction only (Bowmaker *et al.* 2003).

The problem of determination of the strand of origin for some of the mutations observed in the HVS I and II mutation spectra can be solved using the strand dislocation model. Figure 3 demonstrates this for the L- and H-strand mtDNA fragment (L: 5′-GATCCTA-3′, H: 5′-TAGGATC-3′) where monotonous dinucleotide CC/GG runs are located in the 5′ and 3′ direction with respect to T- and A-nucleotides on the L- and H-strands, respectively. Figure 3 shows that four different variants of primer and template strand dislocation at CC/GG runs during replication can lead to specific nucleotide substitutions at each position of the L-strand sequence 5′-TCCT-3′. As is seen, irrespective of which DNA strand is replicated, strand slippage in a homonucleotide run followed by base substitution occurs when homonucleotide runs are placed in a 3′ direction. However, because of the complementary nature of DNA, we can observe a different location for homonucleotide runs relative to the variable nucleotide observed on the L-strand: homonucleotide runs are located in a 3′ direction if strand dislocation originally occurred on the L-strand, and homonucleotide runs are located in a 5′ direction if strand dislocation actually occurred on the H-strand (Figure 3). As a result, a simple rule predicting the mutation change, depending on which DNA strand is the strand of origin for the mutation and which strand (primer or template) was dislocated during replication, follows from the data obtained:

$\underline{\mathbf{X}}$YY → $\underline{\mathbf{Y}}$YY (X→Y, L-strand, primer strand)
X$\underline{\mathbf{Y}}$Y → X$\underline{\mathbf{X}}$Y (Y→X, L-strand, template strand)
YY$\underline{\mathbf{X}}$ → YY$\underline{\mathbf{Y}}$ (X→Y, H-strand, primer strand)
Y$\underline{\mathbf{Y}}$X → Y$\underline{\mathbf{X}}$X (Y→X, H-strand, template strand).

To determine the strands on which any observed transitions or transversions originally occurred as a result of strand dislocation mutagenesis, we have investigated the distribution of trinucleotide sequences, containing monotonous dinucleotide runs, on the L-strand of the mtDNA reference sequence. For transitions, the distribution of eight types of trinucleotides (TCC, CTT, AGG, GAA, TTC, CCT, AAG, GGA) was analyzed; for transversions, the distribution of 16 different

 © University College London 2004

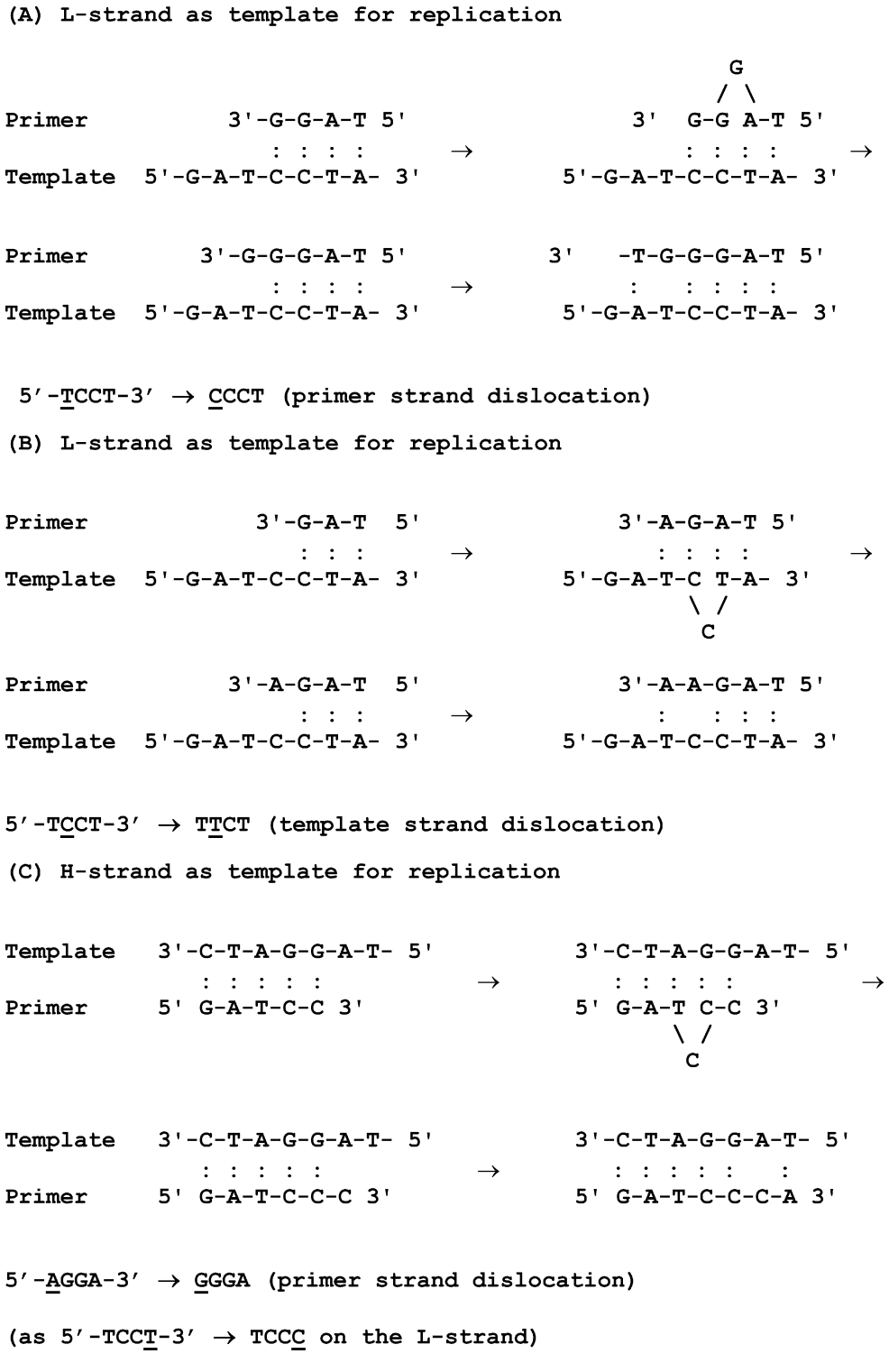

Figure 3 Four models of dislocation mutagenesis: A) primer strand dislocation during H-strand replication, B) template strand dislocation during H-strand replication, C) primer strand dislocation during L-strand replication, D) template strand dislocation during L-strand replication. Four-nucleotide subsequences of the template L-strand (A and B) and H-strand (C and D) are shown below schematic representations of dislocation models, dislocation mutations arising on the L- and H-strands are underlined.

© University College London 2004

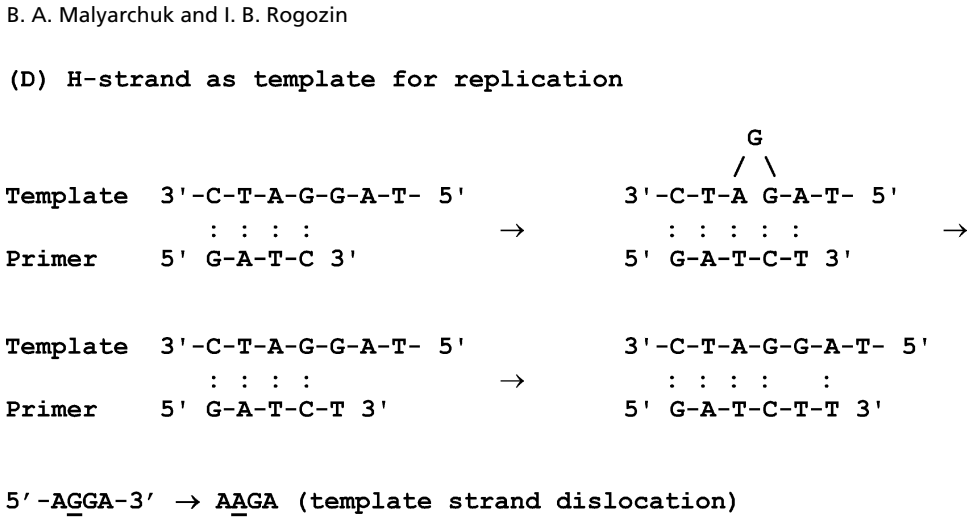

Figure 3 Continued.

	L-strand		H-strand	
HVS I Transition and strand dislocated	Expected variable positions	Observed variable positions (number of mutations)	Expected variable positions	Observed variable positions (number of mutations)
T->C (primer)	6	5 (184)	3	1 (1)
T->C (template)	5	2 (21)	6	4 (25)
C->T (primer)	5	3 (38)	6	3 (25)
C->T (template)	6	4 (13)	3	1 (32)
A->G (primer)	2	1 (2)	8	7 (118)
A->G (template)	2	1 (1)	4	2 (15)
G->A (primer)	2	0	4	2 (4)
G->A (template)	2	0	8	6 (88)
In total	30	16 (259)	42	26 (308)
Primer/template ratio	15/15	9/7 (224/35)	21/21	13/13 (148/160)
Py->Py/Pu->Pu ratio	22/8	14/2 (256/3)	18/24	9/17 (83/225)

Table 4 The observed and expected numbers of variable positions and dislocation mutations in the L- and H-strands of the HVS I region

trinucleotide sequences (GTT, TGG, ATT, TAA, ACC, CAA, GCC, CGG, TTG, GGT, CCA, AAC, CCG, GGC, TTA, AAT) was examined. Taking into account the rules predicting mutations in accordance with the strand dislocation model, one may compare the distribution of variable positions and mutations observed in the HVS I and II sequence data sets with the expected distributions.

Table 4 presents the observed and expected measurements for variable positions on each of the L and H strands for the HVS I data. The results suggest that more than half of the expected variable positions are actually observed on both mtDNA strands (53.3% on the L-strand and 61.9% on the H-strand). On the L-strand, pyrimidine transitions were most frequent, whereas on the H-strand purine transitions prevailed. It is important that there is strong statistical support for the primer strand dislocation model only for transitions occurring on the L-strand. Thus, for all 481 pyrimidine transitions observed on the L-strand (i.e., observed in the mutation spectra recorded relative to the L-strand), 256 (as pyrimidine transitions) originally occurred on the L-strand and 225 (as purine transitions) originated on the H-strand. Among 86 purine transitions observed on the L-strand, the overwhelming majority of them originally occurred on the H-strand as pyrimidine transitions (83 versus 3 purine transitions in the L-strand). For transversions, only 20.8% and 17.4% of the expected variable

 © University College London 2004

Table 5 The observed and expected numbers of variable positions and dislocation mutations in the L- and H-strands of the HVS II region

HVS II Transition and strand dislocated	L-strand		H-strand	
	Expected variable positions	Observed variable positions (number of mutations)	Expected variable positions	Observed variable positions (number of mutations)
T->C (primer)	6	1 (7)	4	0
T->C (template)	7	3 (48)	2	1 (1)
C->T (primer)	7	3 (12)	2	1 (4)
C->T (template)	6	2 (36)	4	0
A->G (primer)	2	2 (7)	9	4 (97)
A->G (template)	3	1 (3)	6	1 (3)
G->A (primer)	3	1 (5)	6	0
G->A (template)	2	2 (8)	9	2 (23)
In total	36	15 (126)	42	10 (128)
Primer/template ratio	18/18	7/8 (31/95)	21/21	5/5 (101/27)
Py->Py/Pu->Pu ratio	26/10	9/6 (103/23)	12/30	3/7 (5/123)

positions were found on the L and H strands, respectively. Among transversions, the most frequent on the L-strand were A-to-C and C-to-A substitutions (82% out of all transversions observed) and on the H-strand most common were G-to-T and T-to-G transversions (77.8%). Estimation of the primer/template ratios shows that only for the H-strand transversions is there a predominance of the template dislocation model over the primer model (primer/template ratio is 10/17); however, this excess is statistically insignificant ($P = 0.12$ by the binomial test).

Table 5 presents the observed and expected numbers of variable positions on the L and H strands for the HVS II sequence. These data suggest that less than half of the expected variable positions potentially involved in dislocation mutagenesis are actually observed on both mtDNA strands (41.7% on the L-strand and 23.8% on the H-strand). As in the case for the HVS I region, in HVS II most frequent were pyrimidine transitions on the L-strand and purine transitions on the H-strand. Strong statistical support for the primer strand dislocation model was found for transitions occurring on the H-strand, whereas template dislocation mutagenesis operates three times more frequently on the L-strand. Among all 226 pyrimidine transitions observed in the L-strand mutation spectrum, 103 (as pyrimidine transitions) originally occurred on the L-strand and 123 (as purine transitions) on the H-strand. Among 28 purine transitions observed, the majority of them occurred on the L-strand (23 versus 5 transitions on the H-strand). We did not find the transversions compatible with the strand dislocation model in the HVS II region, although the expected number of variable positions potentially leading to transversions was high (116 and 110 positions expected on the L and H strands, respectively).

In general, our results suggest that DNA strands dislocation during the replication appears to be an important mechanism for mutation in the mitochondrial genome. However, these mutational events may have a different origin depending on which DNA strand mutated. One of the interesting findings of the present study is the observation that there are some differences between the two mtDNA hypervariable regions, with respect to whether mutations occur on the primer or template strand during the replication. The frequency of mutations compatible with the primer strand dislocation in the HVS I region is almost twice as high as that for template strand dislocation (Table 4 and 5); no excess of mutations compatible with primer strand dislocation was found in the HVS II. However, if we apply the dislocation mutagenesis model to mtDNA L and H strand replication, the primer dislocation mutagenesis, leading to pyrimidine transitions in the parental L-strand, may operate during H-strand replication primarily in the HVS I region, whereas in the HVS II the most pronounced mutation process is template strand dislocation mutagenesis, also generating pyrimidine transitions. It is noteworthy that during the L-strand replication only primer strand dislocation operates on the parental H-strand in the HVS I and II regions, leading to purine transitions.

© University College London 2004

Discussion

It is well known that the hypervariable segments of the mtDNA control region are characterized by extreme site-specific rate heterogeneity, with a high substitution rate at certain nucleotide positions called mutation hotspots (Hasegawa *et al.* 1993; Wakeley, 1993; Excoffier & Yang, 1999; Meyer *et al.* 1999; Pesole & Saccone, 2001; Bandelt *et al.* 2002; Malyarchuk *et al.* 2002b; Meyer & von Haeseler, 2003). In the present study, we have further extended this concept, analyzing the phylogenetically reconstructed mutation spectra of the HVS I and II regions. The results of this study clearly demonstrate that an excess of pyrimidine transitions is a distinctive feature of both HVS I and II mutation spectra, and this is consistent with previous observations (Meyer *et al.* 1999; Tamura, 2000; Malyarchuk *et al.* 2002b). However, despite the predominance of pyrimidine transitions in both spectra, there is a good correspondence between the base composition and the frequency of variable nucleotide positions. These results suggest that the variable positions in the HVS I and II regions may be free from selective constraints. Nevertheless, despite the high content of A-bases in the L-strand of the HVS I and HVS II regions, the average numbers of independent mutations per A have the lowest values among all bases analyzed, so the lower frequency of mutations at A-bases may be attributable to some (possibly, structural) constraints. It should be noted that according to our preliminary results for the mtDNA coding-region data (represented by about 800 sequences published in Finnila *et al.* 2001; Maca-Meyer *et al.* 2001; Herrnstadt *et al.* 2002), the pattern of nucleotide substitutions is different between the control region and the coding region. In the mtDNA coding region only 4%-10% of the sites appear to be variable, and there is no excess of pyrimidine over purine transitions. Meanwhile, despite the low content of guanines in mtDNA as a whole, the highest numbers of mutations per site are observed at Gs in the protein-coding and rRNA-coding genes, suggesting that mutational pressure at the nucleotide level might have an important role in generating base substitutions in the mtDNA coding regions.

Thus, one of the interesting features of the mtDNA control region variability is a strong bias of nucleotide substitutions to pyrimidine transition changes. It has been suggested previously (Malyarchuk *et al.* 2002b), as well as in the present study, that DNA sequence context properties of the control region may influence the pattern of substitutions seen in the HVS I and II mutation spectra. Previously, contextual analysis of hotspots has demonstrated a complex influence of neighbouring bases on mutagenesis in the HVS I region. Statistical analysis of these hotspots has revealed two hotspot motifs, CC and KTNCNK in the HVS I mutation spectrum. However, we have found that these motifs do not correlate with the distribution of substitutions along HVS II. It seems that this might reflect some biological differences between mutation spectra in the HVS I and II regions. The most important finding of this and our previous study is that a statistically significant manifestation of dislocation mutagenesis for *in vivo* substitution spectra was found, indicating that such mutagenesis might be a general mechanism of substitutions in the human mtDNA control region. This model explains the origin of 23.4% and 49.4% of mutations observed in the HVS I and HVS II mutation spectra, respectively. Importantly, analysis of the strand dislocation model has allowed us to recognize the original DNA strands, L or H, on which the initial mutation events occurred. The data obtained have shown that the observed high rate of pyrimidine transitions in both HVS mutation spectra does not appear to be a result of the preferential purine transitions originating on the parental H-strand due, as previously speculated (Tamura, 2000), to the higher rate of purine transitions occurring on the single-stranded H-strand during mtDNA replication. We have found that both pyrimidine transitions on the L-strand and purine transitions on the H-strand give an approximately equal input to the excess of pyrimidine transitions observed in mutation spectra reconstructed on the basis of the L-strand sequence. It is possible also that these results are relevant to the model of mtDNA replication depicting an initially bidirectional replication, originating downstream of OriH (Bowmaker *et al.* 2003).

However, there should be other mechanisms of mutation, since the strand dislocation model allows us to explain the origin of only source of the mutations seen in the HVS I and II mutation spectra. Among other possible mechanisms, depurination can explain the higher mutation rate of purines on the single-stranded H-strand, since the rate of depurination of single-stranded

 © University College London 2004

DNA is four-times greater than that of double-stranded DNA (Lindahl, 1993). In addition, the repair of abasic sites requires double-stranded DNA to use the complementary strand as the template (Pinz & Bogenhagen, 1998). Base loss, giving rise to apurinic/apyrimidinic sites, may be spontaneous or induced by oxidative stress DNA damage (Loeb & Preston, 1986). Deamination of cytosine to uracil and adenine to hypoxanthine is another possible mechanism of mutation on the single-stranded H-strand (Reyes *et al.* 1998). This model suggests that one should expect more G-to-A and T-to-C transitions on the L-strand. The rates of these transitions are not so high in the HVS I (Table 1). Although the rate of T-to-C mutations is higher in the HVS II region, the majority of them (73%) are compatible with the strand dislocation model. However, the compositional correlation with the duration of single-stranded H strand during mtDNA replication is compatible with a gradient of deamination of cytosine to uracil and of adenine to hypoxanthine (Reyes *et al.* 1998). It has been found recently, by means of uracil-N glycosylase treatment of human DNA extracted from archaeological sites, that post-mortem damage of ancient mtDNA results mainly from the deamination of cytosine to uracil, with a bias toward L-strand-specific C-to-T changes in the HVS I region (Gilbert *et al.* 2003). Note, however, that *in vivo* steady-state levels of uracil (as well as hypoxanthine) in mtDNA have not yet been reported (Marcelino & Thilly, 1999). Moreover, relative to the nucleus, mitochondria contain a large excess of uracil-DNA glycosylase activity ($\sim$25% of the total cellular activity), scanning DNA for uracil residues (Bogenhagen, 1999). Thus, mitochondria may efficiently repair some DNA damage including uracil bases and abasic sites.

Another type of mtDNA damage may be due to oxidative stress, since about 2% of oxygen in mitochondria is converted to reactive oxygen species, which are able to cause endogenous oxidative damage to mtDNA, such as single- and double-strand DNA breaks, abasic sites and oxidized bases (Beckman & Ames, 1997). Of the oxidized bases, the most common and most important for inducing mutations is 8-oxoguanine that is characterized by a significantly higher incidence in mtDNA than in nuclear DNA (Richter *et al.* 1988). Nevertheless, there is strong evidence to suggest that mitochondria contain almost all components (e.g., 8-oxoguanine glycosylase, 8-oxoGTPase) of the system which responds to oxidative damage to guanosine, either as a free nucleotide or within DNA (Bogenhagen, 1999). In any case, as mtDNA is under strong oxidative stress, a significant excess of C-to-A and G-to-T transversions (when G in a C:G pair is oxidized to 8-oxoG) and A-to-C and T-to-G transversions (when 8-oxoGTP is used as a nucleotide substrate during replication) are expected to be the result of oxidative DNA damage (Cheng *et al.* 1992; Kang & Hamasaki, 2002). As seen in Table 1, only A-to-C and C-to-A transversions appear to be relatively frequent in the HVS I spectrum (accounting in total for 3.5% of mutations); thus, the 8-oxoG-induced mutations do not seem to be a major contributor to the mtDNA control region mutation spectra, especially taking into consideration that 70% of the aforementioned transversions may have arisen by strand dislocation mutagenesis.

Based on the present evidence, it appears that DNA strand dislocation during replication is one of the main mechanisms of mitochondrial mutagenesis. Dislocation mutagenesis in HVS I and II might be a result of errors produced by DNA polymerase γ during replication of human mtDNA. This conclusion is supported by observations that mitochondria in higher organisms (at least in vertebrates) appear to be deficient in the post-replication mismatch repair system responsible for recognizing mismatched bases and point insertions/deletions (Bogenhagen, 1999; Marcelino & Thilly, 1999). As recently reported for mismatch repair activity in rat liver mitochondrial lysate, it is suggested that this activity has not been retained to repair single base mismatches, but to resolve short palindrome-containing loop mismatches (Mason *et al.* 2003).

The control region of mtDNA that harbours not only the origin of H-strand replication but also the mitochondrial transcription promoters, three conserved-sequence blocks and termination-associated sequence, is the main control site for mtDNA replication and transcription (Foran *et al.* 1988; Shadel & Clayton, 1997). Thus, the control region may be under regulatory functional constraints and its functional organization may influence the pattern of nucleotide substitutions. This important problem appears to be largely unexplored, in part because not all functional segments have been identified in the HVS I and II regions. Nonetheless, some short nucleotide segments are highly conserved

© University College London 2004

among human sequences, and these may be novel functional elements in the HVS I and II regions (Tamura, 2000). The present study, based on extensive analysis of control region variability, has also identified short segments of low variation in the HVS I region, that might indicate that this region is functionally important (Figure 1). As seen in figure 1, the HVS I region contains sparse "cold" trinucleotides CAT (between positions 16099-16101, 16115-16117, 16159-16161, 16236-16238, 16306-16308, 16313-16315, 16321-16323, 16332-16334, 16339-16341) that are often observed as a parts of the sequence GTACAT (between positions 16096-16101, 16156-16161, 16303-16308, 16310-16315, 16329-16334, 16336-16341). Interestingly, this sequence sometimes includes a hotspot at T in the context of the motif KTNCNK. The function of these conserved CAT-sequences is unknown but one may speculate that their low variation is due to protection by putative protein-DNA associations.

The HVS II region is probably even more important for mtDNA replication and transcription processes because it contains all major regulatory elements. Previous studies have shown that the HVS II region is less variable than HVS I but substitution rate heterogeneity is more

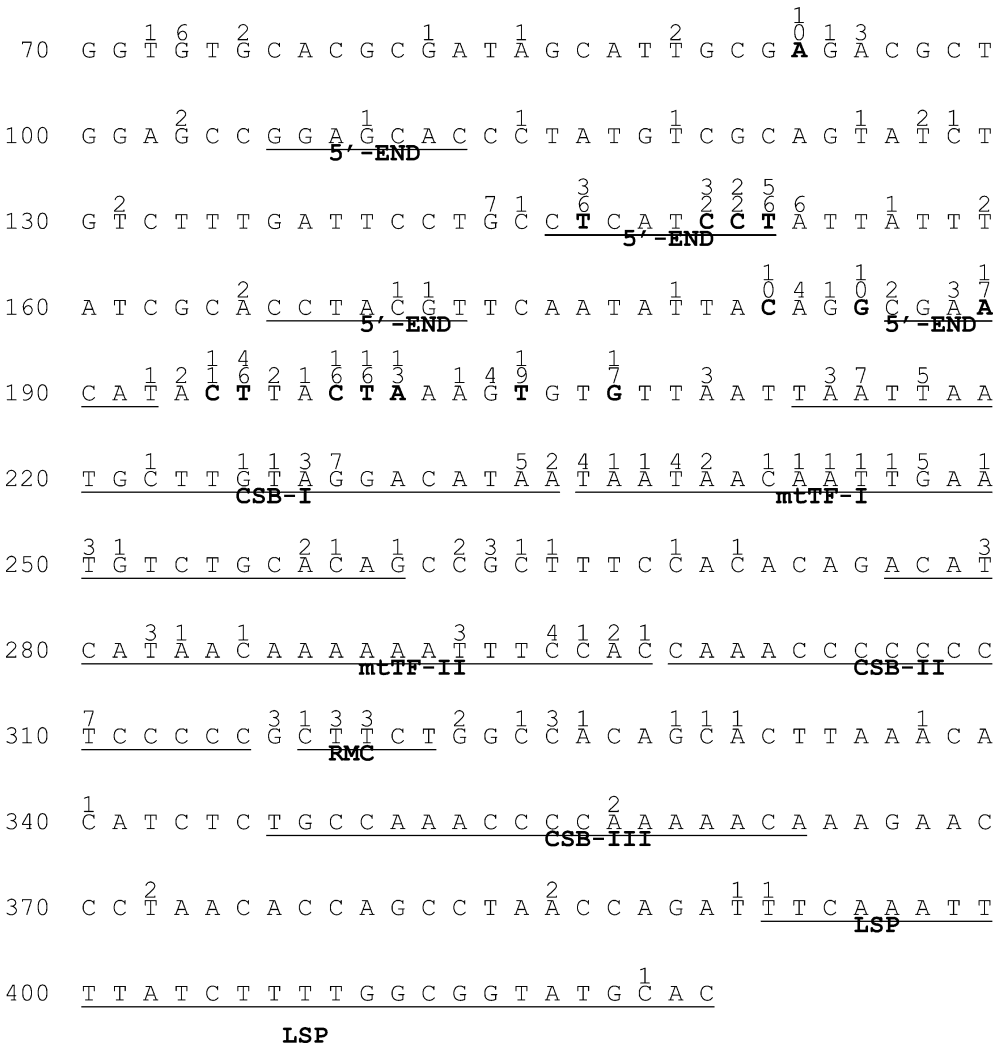

Figure 4 Nucleotide sequence of the L-strand HVS II region. OriH - region of the origin of H-strand synthesis (from nps 110 to 440); 5′-END = 5′-ends of H-strand synthesis; CSB-I, II, III - conserved sequence blocks; RMC - RNAse MRP cleaving site; mtTF - mitochondrial transcription-factor binding sites; LSP = L-strand promoter. Hotspot sites are shown in bold. Functional elements are underlined.

 © University College London 2004

pronounced in HVS II (Meyer *et al.* 1999). It has also been shown that the majority of positions in HVS II functional elements evolve much more slowly than average sites in HVS II, with some rare exceptions (Meyer *et al.* 1999). Our present study also demonstrates that, on average, HVS II appears to be less variable than HVS I, and the most frequent HVS II sites involve only four nucleotide positions (at positions 146, 150, 152 and 195). As seen in Figure 4, which presents the HVS II mutation spectrum relative to the distribution of known functional elements, only rare mutations occurred at regulatory elements, such as the three conserved-sequence blocks (CSBs), two mitochondrial transcription-factor binding sites (mtTFs), the RNase MRP cleaving site (RMC) and the L-strand transcription promoter (LSP) (Foran *et al.* 1988; Shadel & Clayton, 1997). Importantly, all hotspot sites revealed in HVS II are located in the central area, restricted by positions 146-207; this subregion maps only 5′-ends of the H-strand replication. It is noteworthy that not all of the 5′-end sites appear to be variable, since at sites located between positions 166-172, as well as between positions 106-112, most substitution rates are close to zero. *In organello* footprint analysis of human mtDNA suggested that mtTFA-DNA binding domains of about 30 base pairs occur at regular intervals in the control region (Ghivizzani *et al.* 1994). In the HVS II portion, downstream of CSB-1, there are at least three domains of protein-DNA interaction, located approximately between positions 91-114, 124-148, and 163-185. The majority of hotspot sites predicted in the HVS II mutation spectrum is located at the boundary on the bound or outside of the domains protected by proteins (for example, positions {93}, {146, 150, 151, 152} and {182, 185, 189, 194, 198, 199, 200}). This observation suggests that some DNA sites in the HVS II region may be protected from mutational changes by protein-DNA interactions. Therefore, further investigations are necessary to elucidate the variety of mechanisms of nucleotide substitutions in the mtDNA control region, as well as other parts of mtDNA.

Acknowledgements

This work was partially supported by RFBR (grants No. 01-01-00839, 02-04-48342, 02-04-49889, 03-04-48162) and by Far-East Branch of the Russian Academy of Sciences (grant No. 03-3-A-06-096). We thank M.V. Derenko and two anonymous reviewers for helpful comments on the manuscript.

References

Alves-Silva, J., da Silva Santos, M., Guimaraes, P. E. M., Ferreira, A. C. S., Bandelt, H.-J., Pena, S. D. J. & Prado, V. F. (2000) The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* **67**, 444–461.

Anderson, S., Bankier, A. T., Barrel, B. G., de Bruijn, M. H. L., Coulson, A. R. & Drouin, J. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147.

Bandelt, H.-J., Alves-Silva, J., Guimaraes, P. E. M., Santos, M. S., Brehm, A. & Pereira, L. *et al.* (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann Hum Genet* **65**, 549–563.

Bandelt, H.-J., Forster, P., Sykes, B. C. & Richards, M. B. (1995) Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743–753.

Bandelt, H.-J., Macaulay, V. & Richards, M. (2000) Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol Phylogenet Evol* **16**, 8–28.

Bandelt, H.-J., Quintana-Murci, L., Salas, A. & Macaulay, V. (2002) The fingerprint of phantom mutations in mtDNA data. *Am J Hum Genet* **71**, 1150–160.

Beckman, K. B. & Ames, B. N. (1997) Oxidative decay of DNA. *J Biol Chem* **272**, 19633–19636.

Bogenhagen, D. F. (1999) Repair of mtDNA in vertebrates. *Am J Hum Genet* **64**, 1276–1281.

Bowmaker, M., Yang, M. Y., Yasukawa, T., Reyes, A., Jacobs, H. T., Huberman, J. A. & Holt, I. J. (2003) Mammalian mitochondrial DNA replicates bidirectionally from an initiation zone. *J Biol Chem* **278**, 50961–50969.

Brown, W. M. (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc Natl Acad Sci USA* **77**, 3605–3609.

Chen, Y.-C., Olckers, A., Schurr, T. G., Kogelnik, A. M., Huoponen, K. & Wallace D. C. (2000) mtDNA variation in the South African Kung and Khwe - and their genetic relationships to other African populations. *Am J Hum Genet* **66**, 1362–1383.

Cheng, K. C., Cahill, D. S., Kasai, H., Nishimura, S. & Loeb, L. A. (1992) 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G→T and A→C substitutions. *J Biol Chem* **267**, 166–172.

© University College London 2004

Clayton, D. A. (1982) Replication of animal mitochondrial DNA. *Cell* **28**, 693–705.

Copeland, W. C., Ponamarev, M. V., Nguyen, D., Kunkel, T. A. & Longley, M. J. (2003) Mutations in DNA polymerase gamma cause error prone DNA synthesis in human mitochondrial disorders. *Acta Biochim Polonica* **50**, 155–167.

Copeland, W. C., Wachsman, J. T., Johnson, F. M. & Penta, J. S. (2002) Mitochondrial DNA alteration in cancer. *Cancer Invest* **20**, 557–569.

Dianov, G. L., Souza-Pinto, N., Nyaga, S. G., Thybo, T., Stevnsner, T. & Bohr, V. A. (2001) Base excision repair in nuclear and mitochondrial DNA. *Prog Nucleic Acid Res Mol Biol* **68**, 285–297.

Derbeneva, O. A., Starikovskaya, E. B., Volodko, N. V., Wallace, D. C. & Sukernik, R. I. (2002a) Mitochondrial DNA variation in Kets and Nganasans and the early peopling of Northern Eurasia. *Russ J Genet* **38**, 1554–1560.

Derbeneva, O. A., Starikovskaya, E. B., Wallace, D. C. & Sukernik, R. I. (2002b) Traces of Early Eurasians in the Mansi of Northwest Siberia revealed by mitochondrial DNA analysis. *Am J Hum Genet* **70**, 1009–1014.

Derenko, M. V., Grzybowski, T., Malyarchuk, B. A., Dambueva, I. K., Denisova, G. A. & Czarny, J. *et al*. (2003) Diversity of mitochondrial DNA lineages in South Siberia. *Ann Hum Genet* **67**, 391–411.

Excoffier, L. & Yang, Z. (1999) Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol Biol Evol* **16**, 1357–1368.

Finnila, S., Lehtonen, M. S. & Majamaa, K. (2001) Phylogenetic network for European mtDNA. *Am J Hum Genet* **68**, 1475–1484.

Foran, D. R., Hixson, J. E. & Brown, W. M. (1988) Comparison of ape and human sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis. *Nucl Acids Res* **16**, 5841–5861.

Ghivizzani, S. C., Madsen, C. S., Nelen, M. R., Ammini, C. V. & Hauswirth, W. W. (1994) In organello footprint analysis of human mitochondrial DNA: Human mitochondrial transcription factor A interactions at the origin of replication. *Mol Cell Biol* **14**, 7717–7730.

Gilbert, M. T. P., Hansen, A. J., Willerslev, E., Rudbeck, L., Barnes, I., Lynnerup, N. & Cooper, A. (2003) Characterization of genetic miscoding lesions caused by postmortem damage. *Am J Hum Genet* **72**, 48–61.

Giles, R. E., Blanc, H., Cann, H. M. & Wallace, D. C. (1980) Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci USA* **77**, 6715–6719.

Glazko, G. V., Milanesi, L. & Rogozin, I. B. (1998) The subclass approach for mutational spectrum analysis: application of the SEM algorithm. *J Theor Biol* **192**, 475–487.

Hasegawa, M., Di Rienzo, A., Kocher, T. & Wilson, A. (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. *J Mol Evol* **37**, 347–354.

Herrnstadt, C., Elson, J. L., Fahy, E., Preston, G., Turnbull, D. M. & Anderson, C. *et al*. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian and European haplogroups. *Am J Hum Genet* **70**, 1152–1171.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J. & Labuda, D. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet* **69**, 1113–1126.

Hofmann, S., Jaksch, M., Bezold, R., Mertens, S., Aholt, S., Paprotta, A. & Gerbitz, K. D. (1997) Population genetics and disease susceptibility: characterization of central European haplogroups by mtDNA gene mutations, correlations with D loop variants and association with disease. *Hum Mol Genet* **6**, 1835–1846.

Kang, D. & Hamasaki, N. (2002) Maintenance of mitochondrial DNA integrity: repair and degradation. *Curr Genet* **41**, 311–322

Kivisild, T., Tolk, H.-V., Parik, J., Wang, Y., Papiha, S. S., Bandelt, H.-J. & Villems, R. (2002) The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* **19**, 1737–1751.

Kunkel, T. A. (1985) The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J Biol Chem* **260**, 5787–5796.

Kunkel, T. A. & Soni, A. (1988) Mutagenesis by transient misalignment. *J Biol Chem* **263**, 14784–14789.

Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715.

Loeb, L. A. & Preston, B. D. (1986) Mutagenesis by apurinic/apyrimidinic sites. *Annu Rev Genet* **20**, 201–230.

Longley, M. J., Nguyen, D., Kunkel, T. A. & Copeland, W. C. (2001) The fidelity of human DNA polymerase gamma with and without exonucleolytic proofreading and the p55 accessory subunit. *J Biol Chem* **276**, 38555–38562.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F. & Guida, V. *et al*. (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* **64**, 232–249.

Maca-Meyer, N., Gonzalez, A. M., Larruga, J. M., Flores, C. & Cabrera, V. M. (2001) Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics* **2**, 13.

Malyarchuk, B. A. & Derenko, M. V. (2001) Variation of human mitochondrial DNA: Distribution of hot spots in hypervariable segment I of the major noncoding region. *Rus J Genet* **37**, 823–832.

Malyarchuk, B. A., Grzybowski, T., Derenko, M. V., Czarny, J., Wozniak, M. & Miścicka-Śliwka, D. (2002a) Mitochondrial DNA variability in Poles and Russians. *Ann Hum Genet* **66**, 261–283.

 © University College London 2004

Malyarchuk, B. A., Rogozin, I. B., Berikov, V. B. & Derenko, M. V. (2002b) Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. *Hum Genet* **111**, 46–53.

Marcelino, L. A. & Thilly, W. G. (1999) Mitochondrial mutagenesis in human cells and tissues. *Mutat Res* **434**, 177–203.

Mason, P. A., Matheson, E. C., Hall, A. G. & Lightowlers, R. N. (2003) Mismatch repair activity in mammalian mitochondria. *Nucl Acids Res* **31**, 1052–1058.

Meyer, S. & von Haeseler, A. (2003) Identifying site-specific substitution rates. *Mol Biol Evol* **20**, 182–189.

Meyer, S., Weiss, G. & von Haeseler, A. (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regionsv I and II of human mtDNA. *Genetics* **152**, 1103–1110.

Nishigaki, Y., Marti, R., Copeland, W. C. & Hirano, M. (2003) Site-specific somatic mitochondrial DNA point mutations in patients with thymidine phosphorylase deficiency. *J Clin Invest* **111**, 1913–1921.

Pesole, G. & Saccone, C. (2001) A novel method for estimating substitution rate variation among sites in a large dataset of homologous DNA sequence. *Genetics* **157**, 859–865.

Pinz, K. G. & Bogenhagen, D. F. (1998) Efficient repair of abasic sites in DNA by mitochondrial enzymes. *Mol Cell Biol* **18**, 1257–1265.

Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A. S. (1999) Genetic evidence for an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet* **23**, 437–441.

Richards, M. B., Macaulay, V. A., Bandelt, H.-J. & Sykes, B. C. (1998) Phylogeography of mitochondrial DNA in Western Europe. *Ann Hum Genet* **62**, 241–260.

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B. & Guida, V. *et al*. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67**, 1251–1276.

Richter, C., Park, J. W. & Ames, B. N. (1988) Normal oxidative damage to mitochondrial and nuclear DNA is extensive. *Proc Natl Acad Sci USA* **85**, 6465–6467.

Reyes, A., Gissi, C., Pesole, G. & Saccone, C. (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* **15**, 957–966.

Rogozin, I. B., Kondrashov, F. A. & Glazko, G. V. (2001) Use of mutation spectra analysis software. *Hum Mutat* **17**, 83–102.

Salas, A., Richards, M., De la Fe, T., Lareu, M.-V., Sobrino, B. & Sanchez-Diz, P. *et al*. (2002) The making of the African mtDNA landscape. *Am J Hum Genet* **71**, 1082–1111.

Shadel, G. S. & Clayton, D. A. (1997) Mitochondrial DNA maintenance in vertebrates. *Annu Rev Biochem* **66**, 409–435.

Tamura, K. (2000) On the estimation of the rate of nucleotide substitution for the control region of human mitochondrial DNA. *Gene* **259**, 189–197.

Torroni, A., Rengo, C., Guida, V., Cruciani, F., Sellitto, D. & Coppa A. *et al*. (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* **69**, 1348–1356.

Wakeley, J. (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J Mol Evol* **37**, 613–623.

Wallace, D. C. (1999) Mitochondrial diseases in mouse and man. *Science* **283**, 1482–1488.

Yang, M. Y., Bowmaker, M., Reyes, A., Vergani, L., Angeli, P. & Gringeri, E. *et al*. (2002) Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for a apparent strand-asymmetric DNA replication. *Cell* **111**, 495–505.

Yao, Y.-G., Kong, Q.-P., Bandelt, H.-J., Kivisild, T. & Zhang, Y.-P. (2002) Phylogenetic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* **70**, 635–651.

Received: 11 November 2003
Accepted: 26 January 2004

© University College London 2004