Hum Genet (2002) 111 : 46–53
DOI 10.1007/s00439-002-0740-4

ORIGINAL INVESTIGATION

Boris A. Malyarchuk · Igor B. Rogozin
Vladimir B. Berikov · Miroslava V. Derenko

Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region

Received: 30 January 2002 / Accepted: 16 April 2002 / Published online: 7 June 2002

Abstract Analysis of mutations in mitochondrial DNA is an important issue in population and evolutionary genetics. To study spontaneous base substitutions in human mitochondrial DNA we reconstructed the mutational spectra of the hypervariable segments I and II (HVS I and II) using published data on polymorphisms from various human populations. An excess of pyrimidine transitions was found both in HVS I and II regions. By means of classification analysis numerous mutational hotspots were revealed in these spectra. Context analysis of hotspots revealed a complex influence of neighboring bases on mutagenesis in the HVS I region. Further statistical analysis suggested that a transient misalignment dislocation mutagenesis operating in monotonous runs of nucleotides play an important role for generating base substitutions in mitochondrial DNA and define context properties of mtDNA. Our results suggest that dislocation mutagenesis in HVS I and II is a fingerprint of errors produced by DNA polymerase γ in the course of human mitochondrial DNA replication.

Introduction

The haploid mitochondrial genome is represented by 1,000–10,000 DNA molecules per cell and is strictly maternally inherited in humans (Giles et al. 1980). The mitochondrial DNA (mtDNA) evolves rapidly, at a rate five to ten times higher than single-copy nuclear genes (Brown 1980). The most variable part of mtDNA is a noncoding region (control region) which spans 1122 bases between the tRNA genes for proline (tRNA^{Pro}) and phenylalanine (tRNA^{Phe}; Anderson et al. 1981). The control region of mtDNA includes the origin of H strand replication, the promoters for H and L strand transcription, two transcription-factor binding sites, three conserved sequence blocks associated with the initiation of replication (CSB-1, CSB-2, CSB-3), and the D-loop strand-termination-associated sequences (TAS; Foran et al. 1988). However, despite its functional importance this region is highly polymorphic. The majority of mutations are concentrated in three hypervariable segments, HVS I (positions 16024–16365), HVS II (positions 73–340), and HVS III (positions 438–574; Lutz et al. 1998; Vigilant et al. 1991), the highest density of polymorphic positions was found in HVS I and II (Lutz et al. 1998).

Most of mtDNA variability studies have been based on sequence variation of the rapidly evolving HVS I region. Results of phylogenetic and statistical studies have suggested that different nucleotide positions in the mtDNA control region are characterized by unequal mutation rate (Excoffier and Yang 1999; Hasegawa et al. 1993; Heyer et al. 2001; Meyer et al. 1999; Wakeley 1993). Moreover, it has been shown that variations in HVS I do not appear to be clustered within this region (Wakeley 1993). Richards et al. (1998) suggested that the list of nucleotide positions in HVS I be divided into three classes or sites with fast, intermediate, and slow base substitution rates. The mechanism of the differences between rates of base substitutions is still unclear, but the influence of nucleotide context on the mutagenesis intensity provides a plausible explanation. Length polymorphisms in the polycytosine (poly-C) tract between nucleotide positions 16184–16193 is one of the best known case of molecular instability in HVS I which is caused by the T→C transition at the position 16189 (Bendall and Sykes 1995; Howell and Smejkal 2000). As a consequence of this type of instability many

B.A. Malyarchuk (✉) · M.V. Derenko
Institute of Biological Problems of the North,
Far-East Branch of the Russian Academy of Sciences,
Portovaya str. 18, Magadan 685000, Russia
e-mail: ibpn@online.magadan.su,
Tel.: +7-41322-31164, Fax: +7-41322-34463

I.B. Rogozin
Institute of Cytology and Genetics, Siberian Branch
of the Russian Academy of Sciences, Novosibirsk 630090, Russia

I.B. Rogozin
National Center for Biotechnology Information, NLM,
National Institutes of Health, Bethesda MD 20894, USA

V.B. Berikov
Institute of Mathematics, Siberian Branch
of the Russian Academy of Sciences, Novosibirsk 630090, Russia

distantly related sequences of HVS I may have identical types of poly-C tract.

Parallel and reverse mutations in HVS I sequences create numerous problems for the classification of mtDNAs (Macaulay et al. 1999; Malyarchuk and Derenko 1999; Richards et al. 1998, 2000; Torroni et al. 1996). In order to refine phylogenetic relationships between mtDNA control region sequences high-resolution restriction analysis of coding regions was used in the mtDNA analysis. These studies showed that mtDNAs can be classified into a number of monophyletic clusters (mtDNA haplogroups), defined by one or several restriction sites, and that the maternal genealogy has a marked continent-specific features (Chen et al. 2000; Macaulay et al. 1999; Schurr et al. 1999; Torroni et al. 1996; Wallace 1995). It was shown, for example, that virtually all West Eurasian mtDNAs belong to several haplogroups of mitochondrial sequences (H, V, HV*, U, J, T, R*, I, W, X), which are determined by certain restriction fragment length polymorphism (RFLP) and HVS I nucleotide motifs of cluster-diagnostic mutations (Macaulay et al. 1999; Richards et al. 2000; Torroni et al. 1996). The recent studies of the variation in complete mitochondrial genomes, which have allowed the highest possible level of phylogenetic resolution to be reached, confirmed the previously inferred phylogenetic relationships between haplogroups (Finnila et al. 2001; Maca-Meyer et al. 2001). It was shown that the frequency of parallel mutations in the control region was 31-fold higher than in the mtDNA coding region (Finnila et al. 2001). The list of parallel mutations in HVS I and HVS II published by Finnila et al. (2001) is in a good agreement with that suggested by others (Meyer et al. 1999; Richards et al. 1998; Wakeley 1993). Therefore to date there is a relatively large set of hypervariable positions in HVS I and II, which have been confirmed by several independent studies.

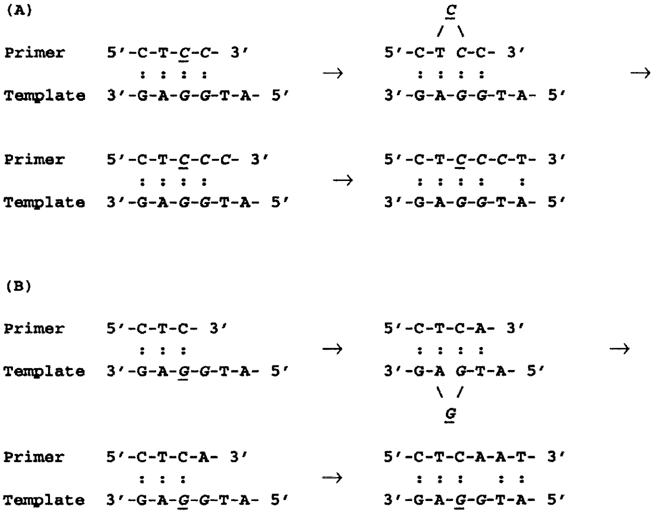

Fig. 1 Two models of dislocation mutagenesis. **A** The primer strand dislocation. **B** The template strand dislocation. *Underlined* Dislocated bases; *italicized* polytracts

To study spontaneous base substitutions in human mtDNA we reconstructed mutational spectra using published data on polymorphisms in various human populations. Analysis of two reconstructed spectra in HVS I and II regions suggested that the transient misalignment dislocation mutagenesis (Fig. 1; Kunkel 1985; Kunkel and Soni 1988) plays an important role for generating substitutions in these regions.

Materials and methods

Reconstruction of mutational spectra

To determine mutations in the mtDNA control region we analyzed different phylogenetic haplogroups of mtDNAs. We used only the published population data comprising the HVS I and/or HVS II nucleotide sequences and additional RFLP or coding region information for each haplotype. The HVS I data set was represented by 4072 West Eurasian sequences (positions 16092–16365 according to the numbering of Anderson et al. 1981), belonging to 34 haplogroups and subgroups: A, B, C, D, E, F, H, HV*, I, J, K, M*, M1, N1a, N1b, N1c, N*, pre-HV, R*, T, U*, U1-U8, V, W, X, Y, Z (Richards et al. 2000). Note that according to the human mtDNA nomenclature (Macaulay et al. 1999; Richards et al. 1998, 2000), each major clade (or mtDNA haplogroup) and nested subclades (or subhaplogroups) were denoted by corresponding Roman numerals. The HVS II data set was represented by 735 individual sequences (positions 72–297) from populations of Europe and South America (Alves-Silva et al. 2000; Finnila et al. 2001; Helgason et al. 2000). Mutations in HVS II were examined among 22 mtDNA haplogroups and subgroups: H, pre-V, V, HV*, J, T, X, I, W, K, U*, U2, U4-U8, A, B, C, D, Z. In all cases nucleotide positions that show point insertions or deletions were excluded from analysis. Similarly, transversions adjacent to the poly-C tract in positions 16184–16193, were ignored as probable artifacts of length variation in HVS I (Bendall and Sykes 1995). Parallel mutations in mtDNA were inferred by revealing variable positions in which identical mutations arose independently in different mitochondrial haplogroups of the previously reconstructed mtDNA phylogenetic tree (Finnila et al. 2001; Macaulay et al. 1999) as described by Macaulay et al. (1999), Finnila et al. (2001), and Malyarchuk and Derenko (2001).

Hotspot prediction

A 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 and resulting hotspot sites were defined for each mutational spectrum separately based on results of classification analysis (Glazko et al. 1998; Rogozin et al. 2001). For this purpose the CLUSTERM program was used (www.itba.mi.cnr.it/webmutation; Glazko et al. 1998). This program decomposes a mutation spectrum into several homogeneous classes of sites; each class is approximated by a Poisson distribution. Variations in mutation frequencies among sites of the same class are due to random reasons (since mutation probability is the same for all sites in one class), but differences between mutation frequencies among sites from different classes are statistically significant. A class, or classes, with the highest mutation frequency is called a hotspot class. A hotspot site is defined as a "permanent" member of the hotspot class C, meaning that this site has a probability of 0.95 or greater of being assigned to the hotspot class C [P(site in C) ≥0.95]. This guarantees that the assignment is statistically significant and robust (Glazko et al. 1998; Rogozin et al. 2001). A novel approach (Berikov 2002) confirmed the reliability of classification results described by Glazko et al. (1998).

Statistical analysis of dislocation model

We analyzed a statistical significance of the dislocation mutagenesis (Fig. 1) using a modification of a Monte Carlo procedure (the CONSEN program; Rogozin and Kolchanov. 1992). This approach takes into account frequencies of substitutions in A, T, G and C bases, a presence of several mutations in a site and context properties of a target sequence. A weight Wj of a site j is a number of substitutions in this site which are compatible with a studied dislocation model. The statistical weights were averaged for all sites in the target sequence resulting in the average weight W. A distribution of averaged statistical weights Wrandom was calculated for 10,000 groups of random sites using a computer random generator. Each group contained a number of mutations equal to the observed one with the same distribution of mutations throughout sites. Based on the distribution Wrandom a probability *P*(W≤Wrandom) was calculated. This probability is equal to the portion of groups of random mutations with Wrandom being the same as or higher then W. Small values (≤0.05) of the probability *P*(W≤Wrandom) indicates a significant correlation between a studied dislocation model and mutations.

Analysis of hotspot nucleotide context

In many cases mutational hotspots emerge due to neighboring nucleotides (Benzer 1961; Cooper and Youssoufian 1988; Coulondre et al. 1978; Horsfall et al. 1990; Krawczak et al. 1998; Rogozin and Kolchanov 1992; Zavolan and Kepler 2001). Hotspot context revealing in this case can be addressed using various methods. A commonly used approach is to calculate the number of times a given base is next to a mutated base, immediately in the 5′ or 3′ direction (positions –1 and +1). A significance of deviation from the expected can be estimated by using various statistical tests (Berikov and Rogozin 1999; Blake et al. 1992; Cariello 1994; Krawczak et al. 1998; Pozdnyakov et al. 1997; Rogozin and Kolchanov 1992). We used reconstruction of a hotspot consensus sequence using the binomial test (Pozdnyakov et al. 1997). In this method, a number N_{IJ} of a nucleotide I was calculated in each position J in a set of M aligned hotspot sequences. The probability $P(N_{IJ},M,F_I)$ to find N_{IJ} or more nucleotides I in a position J was calculated taking a frequency F_I of a nucleotide I in a target sequence as an expected number of the nucleotide I in the position J. A nucleotide with the lowest probability $P(N_{IJ},M,F_I)$ among all possible nucleotides in a position J was accepted as a consensus nucleotide for this position if $P(N_{IJ},M,F_I)$ for this nucleotide was below a threshold value *P**. For one-letter designations of the nucleotide groups that occur at the positions of aligned mutable sites, the commonly used nomenclature was used (A, T, G, C, W=A or T, S=G or C, R=A or G, Y=T or C, M=G or T, K=G or T, B=T or G or C, V=A or G or C, H=A or T or C, D=A or T or G, *n*=A or T or G or C). We analyzed predicted hotspots in A, T, G and C bases separately (below we call them A, T, G and C hotspot sites, respectively). Twenty bases surrounding each hotspot position were used for our analysis.

It is important to note that the estimate of $P(N_{IJ},M,F_I)$ cannot be used for rejecting or accepting statistical hypothesis due to multiplicity of binomial tests, moreover these test were strongly interdependent for each position. In order to estimate a true critical value of the P (N_{IJ},M,F_I) test we developed a resampling procedure. In this procedure N_{IJ} sites were randomly chosen from a target sequence and the minimal value $P_{mr}(N_{IJ},M,F_I)$ was calculated among all positions. This procedure was repeated 10,000 times, a critical value P* that separates the right critical region of the distribution $P_{mr}(N_{IJ}, M,F_I)$ at 5% level of significance was calculated. *P** was 0.005 and 0.008 for C and T sites, respectively.

We also used various implementations of regression analysis of mutational spectra (Berikov and Rogozin 1999; Rogozin and Kolchanov 1992) as well as analysis of correlations between hotspot consensus motifs and distribution of substitutions along a target sequence (Rogozin and Kolchanov 1992).

Results

Substitution frequencies and strand asymmetry

Reconstructed mutational spectra in HVS I and II regions are shown in Figs. 2 and 3, respectively. The first reconstructed spectrum includes 1051 substitutions in 276 sites, while the second spectrum was much smaller (115 substitutions in 231 sites). The frequencies of the different types of base substitutions in both spectra were not equal, transitions constitute 93% of all mutations in the HVS I spectrum (Table 1). This result is in good agreement with previous observations (Budowle et al. 1999; Lutz et al. 1998; Meyer et al. 1999; Tamura 2000; Vigilant et al. 1991). A strong strand bias is more expressed in the HVS I spectrum where transitions between pyrimidines constitute a majority of the mutations, while in the HVS II spectrum frequencies of G→A and C→T substitutions were almost equal (Table 1).

Mutational hotspots

Analysis of the mutational spectrum in the HVS I region using CLUSTERM revealed four classes of sites. The first class includes obvious “cold” sites with number of substitutions varying from 0 to 2, 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 5 to 17, the fourth class includes obvious hotspot sites (number of mutations varied from 16 to 32). Differences between the observed and the expected distributions (a mixture of four Poisson distributions) were statistically insignificant (*P*=0.63). It is important to note that hotspots are not equivalent to a class with the highest frequency of mutations even if two classes of sites are revealed. The problem of hotspot prediction becomes extremely complicated if any other number of classes is revealed by CLUSTERM

Table 1 Frequencies of transitions/transversions

	HVS I	HVS II
Transitions		
A→G	145	23
T→C	313	51
G→A	80	19
C→T	440	21
Transversions		
A→C	20	1
T→G	2	–
A→T	15	–
T→A	–	–
G→T	–	–
C→A	19	–
G→C	4	–
C→G	13	–

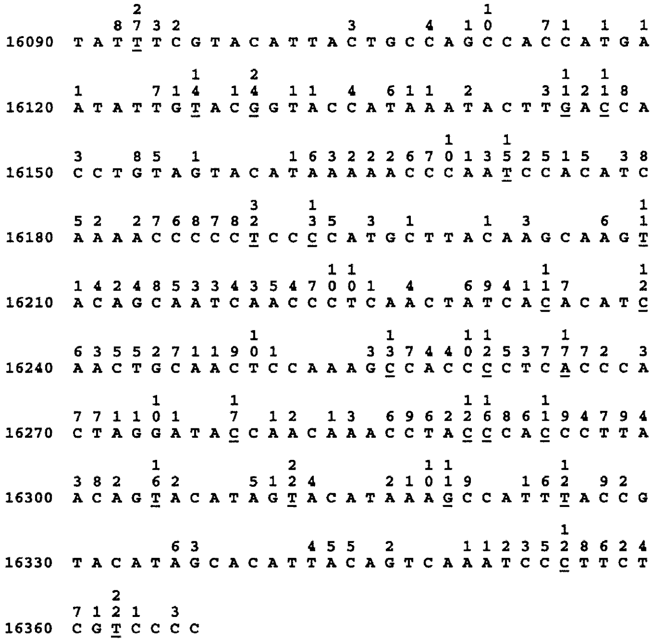

Fig. 2 The reconstructed HVS I mutational spectrum. *Underlined* Predicted hotspots; *numbers above sequence* number of mutations at the position

(Glazko et al. 1998; Rogozin et al. 2001). We suggested that the second class does not contain hotspot sites since several sites with no mutations were included, while hotspots may be present in the third class of sites. Resident members of the third class [*P*(site in 3)≥0.95] were sites with more than 10 mutations, thus 11 was chosen as the threshold value Sh for hotspot sites, these sites are underlined in Fig. 2. Of these 24 positions 20 (Table 2) were rapidly evolving according to the other studies (Gurven 2000; Hasegawa et al. 1993; Meyer et al. 1999; Wakeley 1993), while four remaining sites (hotspots at positions 16147, 16239, 16265, 16325) were not classified as fast in the aforementioned studies.

Analysis of the HVS II mutational spectrum revealed three classes of sites, differences between observed distribution and expected one (a mixture of three Poisson distributions) were statistically insignificant (*P*=0.34). The first class includes "cold" sites with number of mutations varying from 0 to 2 mutations, the second class includes sites with the number of mutations from 0 to 8, while the third class includes one obvious hotspot site with 14 mutations. Presence of true hotspot sites in the second class is problematic since one site with no mutations was included in this class. However, we chose Sh, assuming that the second class represents a mixture of two or more classes which were not recovered due to the small data volume, resident members of the second class were sites with 3 or more mutations which was used as the hotspot threshold (Sh=3; Fig. 3). All of these sites have already been identified as fast sites in phylogenetic analyses of the control region sequences and mtDNA coding region haplogroup-specific polymorphisms (Bandelt et al. 2000; Finnila et al. 2001; Macaulay et al. 1999). However, according to our results, the true hotspot was found only at the position 152, whereas other hypervariable sites were hidden among class 2 sites, which might represent a mixture of two or more mutational classes. Hotspot sites in HVS II are not well defined, and results of their analysis should be interpreted with caution. More precise picture awaits analysis of larger HVS II data sets.

Table 2 Positions of predicted HVS I hotspot sites

Number of mutations	Nucleotide positions
22–32	16093, 16129, 16189, 16311, 16362
11–17	16126, 16145, 16147, 16172, 16192, 16209, 16234, 16239, 16256, 16261, 16265, 16278, 16290, 16291, 16294, 16304, 16319, 16325, 16355

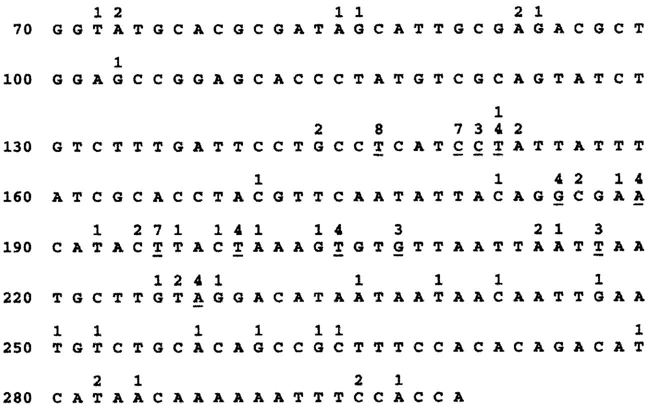

Fig. 3 The reconstructed HVS II mutational spectrum. *Underlined* Predicted hotspots; *numbers above sequence* number of mutations at the position

Dislocation mutagenesis

Strand slippage in repetitive sequences (e.g., monotonous runs of nucleotides or polytracts) may result in base substitutions by the transient misalignment dislocation mechanism (Fig. 1). This model suggests that transient strand slippage in a polytract in the primer or template strand is followed by incorporation of the next correct nucleotide (Kunkel 1985). In previous studies of rat DNA polymerase β in vitro (Kunkel 1985; Kunkel and Soni 1988), the template dislocation model (Fig. 1B) explained a hotspot for T→G substitutions at position 70 in the sequence 5′-G<u>T</u>TTT-3′ (the hotspot is underlined). Analysis of HVS I and II spectra revealed that many base substitution hotspots are consistent with the dislocation model (Table 3). Furthermore, the primer strand dislocation model (Fig. 1A) has a statistically significant support in both spectra [*P*(W≤Wrandom)=0.012 and 0.006, respectively]. The template strand dislocation mutagenesis (Fig. 1B) did not have a significant impact [*P*(W≤Wrandom)=0.323 and 0.168]; however, this does not exclude that this type of mutagenesis operates in sites with a specific nucleotide context [e.g., all four hotspot sites where the template dislocation might operate (Table 3) have a high A+T content; however, this was not implemented in our statistical test]. A strong statistical support for the primer strand dislocation model suggested that several hotspots are truly caused by dislocation mutagenesis (Table 3). Eight of 11 hotspots contain poly-C tracts (Table 3). Interestingly, polytracts in seven dislocation hotspot sites were located in 3′ direction with respect to hotspots (underlined in Table 3); this might reflect an early suggested strand biases of spontaneous mutations in mtDNA (Tamura 2000).

Table 3 Predicted hotspots of mutations in HVS I and II regions (*underlined* predicted hotspots, *italicized* polytracts)

Position	Number of mutations	Hotspot context	Dislocated strand
HVS I			
Dislocation mutagenesis			
16189	32	CCCCC T *CCCC*A	Primer
16362	22	TCTCG T *CCCC*A	Primer
16172	15	CCCAA T *CC*ACA	Primer
16355	12	AAT*CC* *C* *TT*CTC	Template/primer
16319	11	ATAAA G *CC*ATT	Primer
Other hotspot sites			
16265	17	CCCTC A CCCAC	–
16093	27	TGTAT T TCGTA	–
16311	22	CATAG T ACATA	–
16304	16	AACAG T ACATA	–
16126	14	TATTG T ACGGT	–
16325	12	CCATT T ACCGT	–
16209	11	GCAAG T ACAGC	–
Consensus		KTNCNK	–
16129	24	TGTAC G GTACC	–
16145	11	TACTT G ACCAC	–
16294	11	ACCCA C CCTTA	–
16278	17	GGATA C CAACA	–
16291	16	CCTAC C CACCC	–
16256	13	CAAAG C CACCC	–
16192	13	CCTCC C CATGC	–
16261	12	CCACC C CTCAC	–
16290	12	ACCTA C CCACC	–
16239	12	CACAT C AACTG	–
16234	11	TATCA C ACATC	–
16147	11	CTTGA C CACCT	–
Consensus		C C	–
HVS II			
Dislocation mutagenesis			
152	14	CAT*CC* T ATTAT	Primer
146	8	CTG*CC* T CATCC	Primer
195	7	CATAC *T* *T*ACTA	Template
150	7	CTCAT *C* *C*TATT	Template
227	4	CTTGT A *GG*ACA	Primer
151	3	TCAT*C* *C* TATTA	Template
Other hotspot sites			
189	4	GGCGA A CATAC	–
204	4	TAAAG T GTGTT	–
199	4	CTTAC T AAAGT	–
217	3	TTAAT T AATGC	–
185	4	TACAG G CGAAC	–
207	3	AGTGT G TTAAT	–

Nucleotide context of mutational hotspots

We removed all potential dislocation hotspots (Table 3) from the datasets since this type of mutagenesis may have distinct context features. We analyzed predicted hotspots in A, T, G, and C bases separately. The binomial test-based consensus approach revealed several conserved positions for substitutions in C and T sites (Table 3). A highly conserved C was found in the position +1 for C predicted hotspot sites from the HVS I spectrum (resulting consensus sequence is CC, hotspot position is underlined), this position in C hotspot sites has a probability $P(N_{IJ},M,F_I)=0.005$. Three conserved bases for T hotspot sites (K in the position –1, C in the position +2 and K in the position +4) had $P(N_{IJ},M,F_I)$ between 0.002 and 0.0007 which is less than P^* for these sites (0.008), a consensus sequence for T sites was KTNCNK.

It is important to note that both consensus sequences are short; thus they are frequent in sequences. We found a number of cold sites that match the CC consensus; this suggested that some additional context factors influence frequency of substitutions in CC sites. Regression analysis confirmed that context influences may be complex and is not described by CC and KTNCNK consensus sequences only (results not shown). Furthermore, no significant correlation has been revealed between CC and KTNCNK consensus sequences and the distribution of substitutions in the HVS II region.

Discussion

The existence of hypervariable sites in the human mtDNA HVS I and II regions is a well-established phenomenon that has been revealed by various analyses of mtDNA variability (for example, Excoffier and Yang 1999; Gurven 2000; Hasegawa et al. 1993; Heyer et al. 2001; Meyer et al. 1999; Stoneking 2000; Wakeley 1993). Based on different approaches, several discrete classes of HVS I and II sites were revealed; as many as eight (Meyer et al. 1999) and as few as three (Richards et al. 1998) classes. The list of HVS I mutational hotspots found in the present study is in a good concordance with nucleotide positions that have been described as rapidly evolving in the previous studies. The obtained mutational spectra for HVS I and II regions could contain rare errors due to unforeseen problems with the methodology for reconstructing mutations (e.g., undetectable multiple substitutions in a site or inaccurate predictions of ancestral haplotypes; Nei and Kumar 2000). However, these errors are not likely to systematically bias the reconstructed mutational spectra, and thus these spectra should be considered as reliable approximations of spontaneous substitutions in human mtDNA.

One interesting feature of the mtDNA variability is a strong bias of the nucleotide substitutions to the transition changes. It is well known that transitions make up the majority of the substitutions (more than 75%) in the mtDNA control region (Budowle et al. 1999; Lutz et al. 1998), with the average transition/transversion ratio estimated as approximately 15 (Vigilant et al. 1991). Of the transitions, the most prevalent are transitions between pyrimidines, and C→T substitution is the most frequent type of mutations. Transversions, point insertions, and deletions are observed with significantly lower frequency. It has been suggested that the excess of transitions may indicate that mispairing during replication is the major source of spontaneous mutations in mtDNA (Thomas and Beckenbach 1989). A significant excess of A:T→C:G and G:C→T:A transversions is expected to be a result of oxidative DNA damage (Cheng et al. 1992; Richter et al. 1988; Zorov 1996); however, it is not observed in the reconstructed HVS spectra. On the other hand, the results of our study demonstrated the excess of pyrimidine transitions both in HVS I and II regions. Interestingly, a similar trend was revealed when mutational hotspots were analyzed separately. For hotspots in HVS I the transition/transversion ratio is 35.8 and the pyrimidine/purine ratio is 5.63. HVS II sites with more than three independent mutations per site are characterized by a similar estimate of the pyrimidine/purine ratio 3.92. The problem of biases in the mtDNA mutational spectra might be a critical component to understanding of biological mechanisms of mutation (Gurven 2000; Tamura 2000; Wakeley 1993). It has been proposed that hypervariable sites are not mutational hotspots, but instead they are ancient mutations that were redistributed among mtDNA lineages via recombination (Eyre-Walker et al. 1999). However, a recent study performed through the analysis of linkage disequilibrium patterns in the mtDNA control region sequences has shown that many highly variable sites are mutational hotspots (Gurven 2000). Moreover, the examination of the evolutionary rates for sites at which new mtDNA mutations are observed has shown that both germline and somatic mutations occur preferentially at hypervariable sites (Stoneking 2000). The differences between mutation rates estimated from phylogenetic and family studies of mtDNA variability have received now a plausible explanation, it was found recently that nucleotide sites with fast mutation rates, which make up the minority of the variable positions, prevail in pedigree studies (Heyer et al. 2001). Mutational hotspots found in our analysis of phylogenetically reconstructed mutational spectra of mtDNA control region are strongly correlated with positions that were variable in the familial studies (summarized in Heyer et al. 2001).

The questions of why certain nucleotide positions in the mtDNA control region have high mutation rates, and whether mutation at one nucleotide site are influenced by other sites within the mtDNA appear to be important in understanding of the mitochondrial genome evolution (Howell and Smejkal 2000; Howell et al. 1996). To date there are several instances of molecular instability in the HVS I of the human mtDNA. Bendall and Sykes (1995) have identified instability that is associated with heteroplasmic length variation in the polycytosine tract between positions 16184–16193. This length variation may result from instability of the poly-C tract due to the loss of the 16189T variant. Interestingly, the additional mutations (at positions 16186 and 16192) that interrupt the poly-C tract appear to decrease or abolish the length variation (Bendall and Sykes 1995; Marchington et al. 1996).

The dislocation mutagenesis in polytracts was observed in vitro experiments (Kunkel 1985; Kunkel and Soni 1988; Longley et al. 2001). Mutational bias, favoring substitutions toward flanking bases, a phenomenon reminiscent of dislocation mutagenesis, was found in spectra of single-basepair substitutions in human genes (Krawczak et al. 1998). In the present study a statistically significant manifestation of the dislocation mutagenesis for in vivo substitution spectra was found. However, a large number of predicted hotspots were not compatible with the dislocation models. Statistical analysis of these hotspots revealed two hotspot motifs, CC and KTNCNK in the HVS I mutational spectrum. We found that these motifs are not correlated with the distribution of substitutions along HVS II. This might reflect on some biological differences between mutational spectra in these two regions or may be due to a smaller data volume in HVS II. However, possible involvement of the dislocation mutagenesis in generation of substitutions was found in both analyzed spectra, indicating that such mutagenesis might be a general mechanism of substitutions in human mtDNA.

The dislocation mutagenesis was also revealed for errors produced by DNA polymerase γ in vitro (Longley et al. 2001). Thus dislocation mutagenesis in HVS I and II might be a fingerprint of errors produced by DNA polymerase γ during replication of human mtDNA. An important role of DNA polymerase γ in mtDNA mutagenesis is confirmed by association between mutations in this polymerase and multiple mtDNA deletions (Ponamarev et al. 2002; Van Goethem et al. 2001). Revealed differences between the primer strand and the template strand dislocation models might be due to differences in replication of two DNA strands and/or a mutational specificity of DNA polymerase γ [e.g., some dislocation errors can be suppressed by proofreading depending on a sequence context (Longley et al. 2001)]. Since the primer dislocation model predicts the creation of longer polytracts, it is consistent with a high frequency of polytracts in HVS I and II (Figs. 2, 3). The dislocation mutagenesis may cause significant variations in mutability along a nucleotide sequence since regions with a high saturation of polytracts may be more mutable in comparison to regions with a lower frequency of polytracts. A polytract length is a result of several mutational processes including substitutions, deletion, and insertions (Kunkel 1985; Kunkel and Soni 1988; Longley et al. 2001; Pribnow et al. 1981), thus an accurate prediction of mutational frequency is a complicated task. A frequency of substitutions in polytracts might depend on nucleotide context (Longley et al. 2001); for example, poly-C tracts might be prone to dislocation mutagenesis. This is consistent with a presence of numerous poly-C tracts in HVS I and II (e.g., poly-C_{10} in positions 16184–16193) and in other regions of mtDNA (Howell and Smejkal 2000), suggesting that the primer strand dislocation is an important mechanism in defining the context properties of mtDNA.

Acknowledgements This work was partially supported by RFBR (grants no. 00-06-80448, 01-01-00839, 02-04-48342 and 02-04-49889). We thank B.A. Rogozin for helpful discussions and Y.I. Pavlov, F.A. Kondrashov, E.A. Vasunina, O.I. Sinitsina and an anonymous referee for helpful comments on the manuscript.

References

Alves-Silva J, da Silva Santos M, Guimaraes PEM, Ferreira ACS, Bandelt H-J, Pena SDJ, Prado VF (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67:444–461

Anderson S, Bankier AT, Barrel BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJM, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

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

Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 57:248–256

Benzer S (1961) On the topography of the genetic fine structure. Proc Natl Acad Sci USA 47:403–415

Berikov VB (2002) An approach to the evaluation of the performance of a discrete classifier. Pattern Recognit Lett 23:227–233

Berikov VB, Rogozin IB (1999) Regression trees for analysis of mutational spectra in nucleotide sequences. Bioinformatics 15: 553–562

Blake RD, Hess ST, Nicholson-Tuell J (1992) The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J Mol Evol 34:189–200

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

Budowle B, Wilson MR, DiZinno JA, Stauffer C, Fasano MA, Holland MM, Monson KL (1999) Mitochondrial DNA regions HV I and HV II population data. Forensic Sci Int 103:23–35

Cariello NF (1994) Software for the analysis of mutations at the human hprt gene. Mutat Res 312:173–185

Chen Y-C, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC (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 KC, Cahill DS, Kasai H, Nishimura S, Loeb LA (1992) 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G→T and A→C substitutions. J Biol Chem 267: 166–172

Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic disease. Hum Genet 78:151–155

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in Escherichia coli. Nature 274:775–780

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

Eyre-Walker A, Smith NH, Smith JM (1999) How clonal are human mitochondria? Proc R Soc Lond B Biol Sci 266:477–483

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

Foran DR, Hixson JE, Brown WM (1988) Comparison of ape and human sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis. Nucleic Acids Res 16:5841–5861

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci USA 77:6715–6719

Glazko GV, Milanesi L, Rogozin IB (1998) The subclass approach for mutational spectrum analysis: application of the SEM algorithm. J Theor Biol 192:475–487

Gurven M (2000) How can we distinguish between mutational "hot spots" and "old sites" in human mtDNA samples? Hum Biol 72:455–471

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

Helgason A, Sigurdardottir S, Gulcher JR, Ward R, Stefansson K (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. Am J Hum Genet 66:999–1016

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

Horsfall MJ, Gordon AJE, Burns PA, Zielenska M, van der Vliet GME, Glickman BW (1990) Mutational specificity of alkylating agents and the influence of DNA repair. Environ Mol Mutagen 15:107–122

Howell N, Smejkal CB (2000) Persistent heteroplasmy of a mutation in the human mtDNA control region: hypermutation as an apparent consequence of simple-repeat expansion/contraction. Am J Hum Genet 66:1589–1598

Howell N, Kubacka I, Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? Am J Hum Genet 59: 501–509

Krawczak M, Ball EV, Cooper DN (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63:474–488

Kunkel TA (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 TA, Soni A (1988) Mutagenesis by transient misalignment. J Biol Chem 263:14784–14789

Longley MJ, Nguyen D, Kunkel TA, Copeland WC (2001) The fidelity of human DNA polymerase gamma with and without exonucleolytic proofreading and the p55 accessory subunit. J Biol Chem 276:38555–38562

Lutz S, Weisser H-J, Heizmann J, Pollak S (1998) Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany. Int J Legal Med 111:67–77

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A (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 AM, Larruga JM, Flores C, Cabrera VM (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2:13

Malyarchuk BA, Derenko MV (1999) Molecular instability of the mitochondrial haplogroup T sequences at nucleotide positions 16292 and 16296. Ann Hum Genet 63:489–497

Malyarchuk BA, Derenko MV (2001) Variation of human mitochondrial DNA: distribution of hot spots in hypervariable segment I of the major noncoding region. Russ J Genet 37:823–832

Marchington DR, Poulton J, Sellar A, Holt IJ (1996) Do sequence variants in the major non-coding region of the mitochondrial genome influence mitochondrial mutations associated with disease? Hum Mol Genet 5:473–479

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

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, Oxford

Ponamarev MV, Longley MJ, Nguyen D, Kunkel TA, Copeland WC (2002) Active site mutation in DNA polymerase gamma associated with progressive external ophthalmoplegia causes error-prone DNA synthesis. J Biol Chem 277:15225–15228

Pozdnyakov MA, Rogozin IB, Babenko VN, Kolchanov NA (1997) Analysis of neighbor base influence on spontaneous mutations in human pseudogenes. Dokl Akad Nauk 356:566–568

Pribnow D, Sigurdson DC, Gold L, Singer BS, Napoli C, Brosius J, Dull TJ, Noller HF (1981) rII cistrons of bacteriophage T4. DNA sequence around the intercistronic divide and positions of genetic landmarks. J Mol Biol 149:337–376

Richards MB, Macaulay VA, Bandelt H-J, Sykes BC (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, Rengo C, Sellito D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Yu, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozzari R, Torroni A, Bandelt H-J (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251–1276

Richter C, Park JW, Ames BN (1988) Normal oxidative damage to mitochondrial and nuclear DNA is extensive. Proc Natl Acad Sci USA 85:6465–6467

Rogozin IB, Kolchanov NA (1992) Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochim Biophys Acta 1171:11–18

Rogozin IB, Kondrashov FA, Glazko GV (2001) Use of mutation spectra analysis software. Hum Mutat 17:83–102

Schurr TG, Sukernik RI, Starikovskaya YB, Wallace DC (1999) Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea – Bering Sea region during the Neolithic. Am J Phys Anthropol 108:1–39

Stoneking M (2000) Hypervariable sites in the mtDNA control region are mutational hotspots. Am J Hum Genet 67:1029–1032

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

Thomas WK, Beckenbach AT (1989) Variation in salmonid mitochondrial DNA: evolutionary constraint and mechanisms of substitution. J Mol Evol 29:233–242

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obidu D, Savontaus M-L, Wallace DC (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835–1850

Van Goethem G, Dermaut B, Lofgren A, Martin JJ, Van Broeckhoven C (2001) Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. Nat Genet 28:211–212

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of mitochondrial DNA. Science 253:1503–1507

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

Wallace DC (1995) Mitochondrial DNA variation in human evolution, degenerative disease and aging. Am J Hum Genet 57: 201–223

Zavolan M, Kepler TB (2001) Statistical inference of sequence-dependent mutation rates. Curr Opin Genet Dev 11:612–615

Zorov DB (1996) Mitochondrial damage as a source of disease and aging: a strategy of how to fight these. Biochim Biophys Acta 1275:10–15