Persistent Heteroplasmy of a Mutation in the Human mtDNA Control Region: Hypermutation as an Apparent Consequence of Simple-Repeat Expansion/Contraction

Neil Howell[1,2] and Christy Bogolin Smejkal[2]

[1]Biology Division 0656, Department of Radiation Oncology, and [2]Department of Human Biological Chemistry and Genetics, University of Texas Medical Branch, Galveston

In the genealogical and phylogenetic analyses that are reported here, we obtained evidence for an unusual pattern of mutation/reversion in the human mitochondrial genome. The cumulative results indicate that, when there is a T→C polymorphism at nt 16189 and a C→T substitution at nt 16192, there is an extremely high rate of reversion (hypermutation) at the latter site. The apparent reversion rate is sufficiently high that there is persistent heteroplasmy at nt 16192 in maternal lineages and at the phylogenetic level, a situation that is similar to that observed for the rapid expansion/contraction of simple repeats within the control region. This is the first specific instance in which the mutation frequency at one site in the D-loop is markedly influenced by the local sequence "context." The 16189 T→C polymorphism lengthens a $(C:G)_n$ simple repeat, which then undergoes expansion and contraction, probably through replication slippage. This proclivity toward expansion/contraction is more pronounced when there is a C residue, rather than a T, at nt 16192. The high T→C reversion frequency at nt 16192 apparently is the result of polymerase misincorporation or slippage during replication, the same mechanism that also causes the expansion/contraction of this simple-repeat sequence. In addition to the first analysis of this mitochondrial hypermutation process, these results also yield mechanistic insights into the expansion/contraction of simple-repeat sequences in mtDNA.

Introduction

The human mitochondrial genome (mtDNA) evolves rapidly, particularly so within the 1.1-kb noncoding control region or D-loop (Kocher and Wilson 1991). There is great uncertainty about both the overall rate of divergence within the control region and whether there is a simple mtDNA evolutionary "clock" (Horai et al. 1995; Bendall et al. 1996; Howell et al. 1996; Howell and Mackey 1997; Macauley et al. 1997; Parsons et al. 1997). An important phenomenon that limits the analysis of these problems is the marked site variability in the rate of mtDNA substitution. At one extreme, most of the sites within the control region do not vary among humans, but other sites are apparent "hotspots" at the other extreme of the mutational spectrum (Kocher and Wilson 1991; Hasegawa et al. 1993; Wakeley 1993; Excoffier and Yang 1999; Meyer et al. 1999). In addition to the controversy over rate, there also remains much uncertainty about other properties of the evolution of mtDNA, including an understanding of why certain nucleotides—but not others—have high divergence rates, the relationship between mutation/reversion at the molecular level and sequence divergence at the phylogenetic level (that is, the process of fixation and the role of selection in this process), and whether mutation or divergence rates at one site are influenced by other sites within the mtDNA.

The "standard" pathway of mtDNA evolution begins with a mutation in a single germline mtDNA molecule, a segregational stage during which the mutation is heteroplasmic within the mtDNA gene pool of a maternal lineage, establishment of the mutation in the homoplasmic state at the level of the individual and then within the lineage, and—perhaps—eventual fixation at the level of the population. There are numerous points in this pathway at which the newly arisen mutation can be "lost." For example, there is considerable interest in a developmental bottleneck (Lightowlers et al. 1997) in which the effective number of mtDNA transmission units is sharply reduced relative to the number of mtDNA molecules in either uninucleate somatic cells (thousands) or in the mature oocyte (⩾100,000). At each generation, as a result of the bottleneck, there can be rapid shifts in the proportion of mtDNA molecules that carry the mutation. Overall, the bottleneck should

Received December 22, 1999; accepted for publication March 3, 2000; electronically published April 10, 2000.

Address for correspondence and reprints: Dr. Neil Howell, Biology Division 0656, Department of Radiation Oncology, The University of Texas Medical Branch, Galveston, TX 77555-0656. E-mail: nhowell@utmb.edu

0002-9297/2000/6605-0014$02.00

reduce the number of mutations that become homoplasmic but increase the rate at which the homoplasmic state is reached (Howell et al. 1996).

We report here an unusual and complex departure from this typical pathway in which there is persistent heteroplasmy—at both the pedigree and phylogenetic levels—of a polymorphism within the mtDNA noncoding control region. On the basis of the experimental results, a mechanism is proposed that involves an extremely rapid rate of mutation at the molecular level (hypermutation) and that is mechanistically related to expansion/contraction of the surrounding simple-repeat sequence.

Experimental Procedures

DNA was isolated from the white blood cell (WBC)/platelet fraction of venous blood samples, obtained with informed consent, with standard procedures of SDS/proteinase K digestion, phenol extraction (followed by extraction with chloroform/isoamyl alcohol), and ethanol precipitation. DNA was pelleted by centrifugation, washed with 70% ethanol, and resuspended in buffer that contains 10 mM Tris-HCl and 1 mM EDTA (pH 7.5).

The mtDNA control region was amplified as four overlapping fragments of ∼350 bp in length that span the following nucleotide positions: 15,909–16,276; 16,216–16,569; 1–339; and 278–657 (Howell et al. 1995, 1996). With the exception of two experiments, described in the Results section, the PCR amplifications were carried out with *Taq* polymerase. In those two experiments, the amplifications were carried out with *Pfu* polymerase that has a proofreading exonuclease activity and, thus, a much lower error rate than does *Taq* polymerase.

The PCR primers were designed to contain *Sau*3A restriction sites, and the amplified fragments—after isolation and restriction enzyme cleavage—were ligated into *Bam*H1-cleaved M13 sequencing vectors. After bacterial transformation, the nucleotide sequence of cloned inserts in the recombinant vectors was determined using the standard ("manual") dideoxy chain termination method (Howell et al. 1995). The DNA sequence of both strands could be determined for the control-region fragment that spans nt 16189, and similar tract lengths were obtained irrespective of which mtDNA strand was sequenced. All nucleotide positions are numbered according to the revised Cambridge Reference Sequence (rCRS) (Anderson et al. 1981; Andrews et al. 1999). It is conventional that the human mtDNA sequence refers to that of the L-strand. To maintain consistency, therefore, we describe simple repeat sequences as those in the L-strand and we describe their length in terms of "residues." It should be kept in mind, however, that a C repeat of 6 residues is a repeat of 6 C:G base pairs within the mtDNA.

Results

Control-Region Heteroplasmy in an 11778 LHON Family

We elsewhere described (Howell et al. 1994) a matrilineal pedigree that was affected with Leber hereditary optic neuropathy (LHON), a late-onset form of bilateral optic atrophy in which the primary etiological event is a mutation in the mitochondrial genome (Howell 1999). The members of this pedigree were heteroplasmic in the WBC/platelet fraction of whole blood for the pathogenic LHON mutation at nt 11778. That is, there were two populations of mtDNA molecules in these individuals, one that carried the wild-type allele at nt 11778 and the other that harbored the mutant allele. The 11778 mutation load in this mixed cell population increases in a fairly uniform manner through the three generations, averaging 92% for the nine members of generation III (fig. 1 of Howell et al. 1994).

Subsequent analyses revealed that these family members are also heteroplasmic at nt 16192 of the control region (fig. 1). Among the different pedigree members, 0%–20% of the mtDNA molecules carried a C residue at this site, whereas the majority carried a T residue. The former is the wild-type nucleotide, as defined by its presence in the rCRS (Anderson et al. 1981; Andrews et al. 1999). This is the first instance known to us in which individuals are heteroplasmic at sites both in the coding region and in the control region. Our initial supposition was that either the mtDNA in this lineage had undergone a T→C mutation at nt 16192 that had not yet segregated to the homoplasmic state or that the homoplasmic 16192T polymorphism had recently undergone reversion. As will be shown below, however, the actual situation is much more complex than these "standard" mitochondrial genetic phenomena.

The heteroplasmy at nt 16192 is complicated by the simultaneous presence of a T→C polymorphism at nt 16189 in the mtDNA of this LHON pedigree. In the rCRS, there is a CCCCCTCCCC polypyrimidine tract that spans nt 16184–16193 of the L-strand (fig. 2). The length of this tract is stable; however, the length of the homopolymeric tract generally becomes unstable when the T:A base pair at nucleotide position 16189 is replaced by a C:G base pair. Individuals who carry this polymorphism are usually heteroplasmic and display tract lengths of 8–14 residues (Bendall and Sykes 1995; Marchington et al. 1997; and see "16192T/C Heteroplasmy in Other 16189C mtDNAs" below).

The 16189 polymorphism and the 16192 heteroplasmy in the mtDNA of this LHON matrilineal pedi-

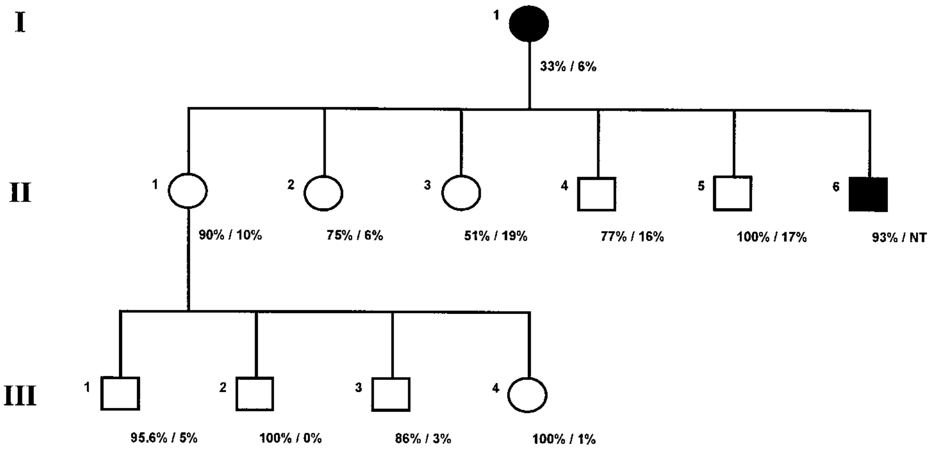

Figure 1 Matrilineal pedigree of the heteroplasmic 11778 LHON family. The first number for each family member is the proportion of the mutant allele at nt 11778, and the second number is the proportion of the mtDNA molecules with a C at nt 16192. Filled symbols represent family members that are visually affected. NT = not tested. A more complete pedigree, in terms of the analysis of the 11778 pathogenic mutation, was published in Howell et al. (1994), but there has been no change in the designations for individual family members.

gree should produce tract sequences of CCCCCCCCCC (16192 wild type) and CCCCCCCCTC (16192 polymorphism), but we observed that tract length was variable for *both* 16192 allelic variants. We analyzed a large number of clones from family member II-1 (DNA sample 0204), and the distribution of tract lengths is summarized in table 1. Tract length was variable in *both* the 16192C and 16192T mtDNA molecules, with modal lengths of 11 residues and 10 residues, respectively. Similar results were obtained for other pedigree members (table 1). All of the 16189C/16192T clones were of the type $(C)_n$TC; that is, the T at nt 16192 and the C at 16193 were invariant and length variation was limited to the C-tract. As will be discussed in "16192T/C Heteroplasmy in Other 16189C mtDNAs" below, tract length was less variable in the 16189C/16192T allelic variant than in 16189C/16192C mtDNAs.

The results in figure 1 indicate that the heteroplasmy at nt 16192 persisted through the three generations of this matrilineal pedigree, as did that at nt 11778. The 16192 allele proportions were not tightly correlated with those at nt 11778, however, as would be expected if both substitutions were simple mtDNA mutations that had arisen in single mtDNA molecules and were now showing segregation in the heteroplasmic state. For example, the proportion of the 11778 mutant allele was 33% in family member I-1, and the proportion of the 16192 C allele was 9% (fig. 1 and table 1). In contrast, the proportions in her daughter were 90% and 10%, respectively.

Bendall et al. (1996) have reported a similar observation to ours, although they did not recognize the complexities of the situation. They analyzed monozygotic twins who were homoplasmic for the 16189C polymorphism and heteroplasmic for the 16192T/C polymorphism; the polypyrimidine tract length was also variable for both allelic variants (see their fig. 3). It is highly unlikely that their subjects were maternal relatives of our 11778 LHON family, and we thus suspected that the heteroplasmy at nt 16192 signaled an unusual mtDNA mutational and/or segregational process. There are multiple explanations for such discordant segregation (including the possibility of intermolecular mtDNA recombination), and additional data were sought that would clarify this departure from the expected segregation pattern.

We have obtained control-region sequences for ~300 European mtDNAs. Although we have sequenced numerous mtDNAs that carried the 16189T/16192C (rCRS), 16189T/16192T, or 16189C/16192C combinations of alleles, we have not observed heteroplasmy at nt 16192 in any of these mtDNAs (see "Heteroplasmy and Tract-Length Variance in Other 16189C mtDNAs" below). Therefore, our first approach was to assay for heteroplasmy in other mtDNAs that carried the 16189C/16192T combination of polymorphisms.

16192T/C Heteroplasmy in Other 16189C mtDNAs

Nucleotide sequencing of segments of the mtDNA coding region, which total ~3 kbp, is routinely used in this laboratory to determine the presence of primary and secondary LHON mutations (reviewed in Howell 1999). Such sequence analysis of the mtDNA from this 11778 LHON pedigree revealed the presence of three informative polymorphisms: an A→G transition at nt 4732 that results in the substitution of serine for the asparagine at amino acid position 88 of the ND2 protein (ND2/

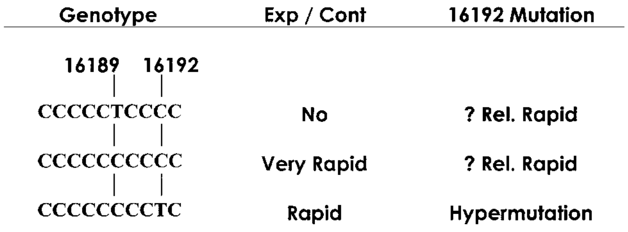

Figure 2 Expansion/contraction and hypermutation in a polypyrimidine tract in the noncoding control region (D-loop). "Exp/Cont" refers to expansion/contraction of the tract as a function of the genotype. Tract length is stable when there is a T at nt 16189, but it is unstable when there is a C. The C at nt 16192 has a higher-than-average substitution rate (Meyer et al. 1999). When there is a C at nt 16189 and a T at nt 16192, however, the rate of reversion is so rapid (hypermutation) that there is persistent heteroplasmy in individuals and maternal lineages, and which also persists at the phylogenetic level.

N88S); a T→C transition at nt 13617 (ND5/isoleucine 427 unchanged); and an A→G transition at nt 13637 (ND5/Q434R). Analysis of complete mtDNA sequences from 60 Europeans indicates that these polymorphisms occur in a subgroup of European haplogroup U mtDNAs (R. M. Andrews, P. F. Chinnery, D. M. Turnbull, N. Howell, unpublished data).

With this information, we reviewed our mtDNA database of ~300 normal controls and patients with known or suspected mitochondrial disorders to identify other lineages whose mtDNA harbored the 16189C/16192T combination of control-region polymorphisms or who are members of this haplogroup U subgroup. We identified three normal controls (DNA samples 0217, 0318, and 0805), a second 11778 LHON pedigree (represented by samples 0284, 0285, and 0286), and a 14484 LHON patient (DNA sample 0803) who is also affected with spondyloepiphyseal dysplasia (N. Howell, I. Kubacka, B. McDonough, A. B. Hodess, D. H. Harter, unpublished data). All of the mtDNAs from these individuals carried the 16189C/16192T control-region polymorphisms. With the exception of the 0217 mtDNA, all of these mtDNAs were members of haplogroup U on the basis of polymorphisms in the coding region (data not shown). The 0217 mtDNA, in contrast, was African in origin (data not shown), and thus one may be confident that the two control-region polymorphisms arose independently in this mtDNA. Finally, we included mtDNA samples from a MERRF family (myoclonic epilepsy with ragged red fibers) that carries the pathogenic mutation at nt 8344 (represented by sample 0159 in fig. 3). The mtDNA from this pedigree belongs to the haplogroup U subgroup, but the initial D-loop sequencing analysis showed the presence of the 16189C/16192C allele combination. We included this pedigree for studies that are described later.

The assays of 16192T/C heteroplasmy for these mtDNAs are summarized in table 2. The results, though complex, indicate the following. All three individuals in the second 11778 LHON family (0284, 0285, and 0286), the 14484 LHON patient (0803), and the normal control (0318) were clearly heteroplasmic at nt 16192 with the proportion of 16192C clones ranging from 6% to 29% of the total. The results for mtDNAs 0805 and 0217/African were different. For individual 0805, we were able to obtain DNA from both the WBC/platelet fraction of whole blood and from a skeletal-muscle biopsy. Among >200 clones, only one was 16192C. In a similar fashion, no 16192C clones were detected among the >100 that were analyzed from individual 0217's mtDNA.

These results indicate that 16192T/C heteroplasmy is not an invariant phenomenon among 16189C/16189T mtDNAs, but there are two noteworthy features of the "low" heteroplasmy mtDNAs. In the first place, the modal length of the simple-repeat sequence was 10 residues, rather than the 11 residues of the "high frequency of heteroplasmy" mtDNAs. Second, it also appeared that tract length instability was greater in the latter mtDNAs. Both properties are reflected in the mean and SD values

Table 1

Polypyrimidine Tract Lengths among Members of the Heteroplasmic 11778 LHON Family

Member[a] and Allele Status	No. of Clones of Tract Length (in residues)					Mean Tract Length ± SD[b]	Total Clones (%)
	⩽9	10	11	12	⩾13		
I-1:							
16192 T	1	85	11	3	0	10.2 ± .5	100 (91)
16192 C	1	7	0	1	1	ND	10 (9)
II-1:							
16192 T	3	67	72	10	2	10.6 ± .7	154 (90)
16192 C	2	13	3	1	0	ND	19 (10)
II-2:							
16192 T	0	14	16	2	0	10.6 ± .6	32 (94)
16192 C	0	1	1	0	0	ND	2 (6)
II-3:							
16192 T	1	18	4	2	0	10.3 ± .7	25 (81)
16192 C	0	4	2	0	0	ND	6 (19)
III-1:							
16192 T	0	18	79	6	0	10.9 ± .5	103 (95)
16192 C	0	2	3	0	0	ND	5 (5)
III-2:							
16192 T	0	4	59	11	2	11.1 ± .5	76 (100)
16192 C	0	0	0	0	0	ND	0 (0)
III-3:							
16192 T	0	4	20	3	1	11.0 ± .6	28 (97)
16192 C	0	1	0	0	0	ND	1 (3)
III-4:							
16192 T	0	13	88	8	0	11.0 ± .4	109 (99)
16192 C	1	0	0	0	0	ND	1 (1)

[a] The family members are designated as shown in figure 1.
[b] ND = not determined.

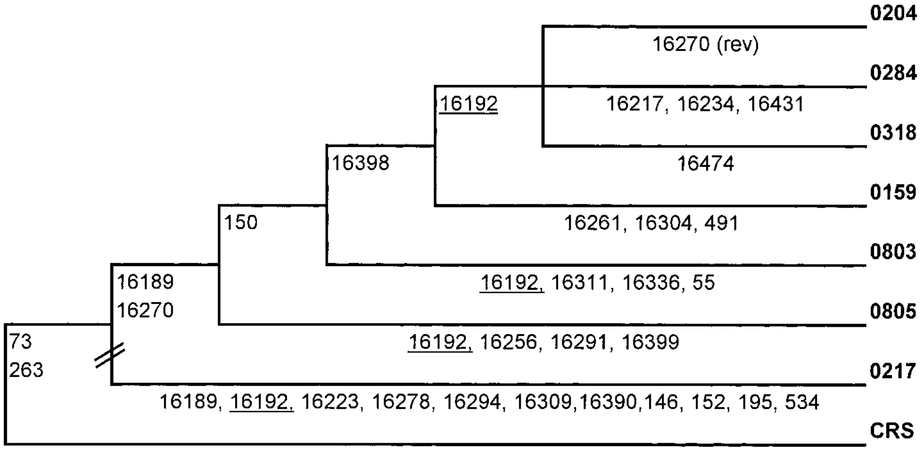

Figure 3 Cladogram of control-region sequences. The numbers beneath the lines indicate the nucleotide positions at which control-region sequence changes have occurred, relative to the rCRS. DNA sample 0204 was obtained from family member II-1 of the pedigree shown in figure 1. The notation 16270(rev) for mtDNA 0204 indicates that this nucleotide has apparently reverted back to the rCRS. Solely on the basis of the control-region sequences, there are multiple most parsimonious trees, but the use of coding region polymorphisms reduces the ambiguity and yields the phylogenetic tree shown here. Thus, mtDNAs 0204, 0284, and 0318 share polymorphisms at nt 4732, 13617, and 13637; mtDNA 0159 carries only the 13617 and 13637 polymorphisms, and mtDNA 0803 carries only the 13617 polymorphism.

(table 2). Thus, the mean tract length for the 0217 mtDNA was 10.0, whereas it varied from 10.2 to 11.1 for the mtDNAs with relatively high levels of 16192 heteroplasmy. The reduced variability in tract length is indicated by the smaller SD: 0.2 for the 0217 mtDNA, but 0.4–0.7 for the 16192 heteroplasmic mtDNAs. The use of these statistical parameters is justified, because the repeat-length distributions for the 16189C/16192T mtDNAs can be fit to simple Gaussian distributions (data not shown). As will be described further in the Discussion, we believe that the relationship between tract length variability and 16192 heteroplasmy reflects the same molecular mechanism.

We also determined the complete nucleotide sequence of the noncoding control region for these mtDNAs. There were multiple sequence differences among these different lineages, but a single-most-parsimonious tree could not be obtained. One possible phylogenetic tree is shown in figure 3. The principal source of ambiguity, when only control-region sequences are used, is the number of times that the 16192T polymorphism has arisen among these mtDNA members of haplogroup U. For example, as shown in figure 2, the polymorphism is shown to have occurred three times among the six mtDNAs. A *single* origin in the common ancestor of the haplogroup U mtDNAs is possible, however, if we include a reversion event (back to the C allele) in the 0159 mtDNA lineage or the segregational loss of 16192T mtDNA in a heteroplasmic maternal ancestor of 0159 (presumably at the stage of the developmental bottleneck, where segregation is often extremely rapid). We also considered the possibility that mtDNA 0159 was actually 16192T/C heteroplasmic but that the initial sequencing analyses were not sufficient to "capture" low-level heteroplasmy. Therefore, we reanalyzed this mtDNA as well as that from two maternal relatives (table 3). Some 300 total clones were sequenced, and all clones carried the 16192C allele. These additional data confirm that the 0159 mtDNA lineage is homoplasmic for the 16192C allele. Strong support for this tree is provided by the use of coding region polymorphisms (see fig. 3 legend).

These results illustrate a point we made previously: namely, there is no 16192 heteroplasmy in 16189C/16192C mtDNAs (see further discussion in the "Heteroplasmy and Tract-Length Variance in Other 16189C mtDNAs" section). The results in table 3 provide a further piece of information worth consideration. The tract lengths of the simple-repeat are more variable in the 16189C/16192C mtDNAs than are the lengths for the 16189C/16192T mtDNAs. The SDs for the length distributions for 16189C/16192C mtDNAs ranged from 0.9 to 1.1 (table 3), higher than those found for the 16189C/16192T mtDNAs (tables 1 and 2). This pattern strongly suggests that the truly homopolymeric sequence is more prone to expansion and contraction.

There is debate over the overall rate of mtDNA divergence in the control region and whether a single rate estimate can be used as a molecular clock to time the evolution and geographic movement of humans (Howell and Mackey 1997; Macauley et al. 1997; Parsons et al. 1997). Nevertheless, the control-region sequence differences among the haplogroup U mtDNAs (fig. 2) indicate that their last common ancestor arose hundreds—if not

Table 2

Polypyrimidine Tract Lengths in 16189C/16192T Individuals

Individual and Allele Status	No. of Clones of Tract Length (in residues)					Mean Tract Length ± SD	Total Clones (%)
	≤9	10	11	12	≥13		
0284:							
16192 T	0	9	16	1	1	10.7 ± .7	27 (90)
16192 C	0	1	2	0	0	ND	3 (10)
0285:[a]							
16192 T	0	1	14	2	0	11.1 ± .4	17 (71)
16192 C	0	2	0	5	0	ND	7 (29)
0286:[a]							
16192 T	0	6	21	1	1	10.9 ± .6	29 (94)
16192 C	0	2	0	0	0	ND	2 (6)
0318:							
16192 T	2	27	14	2	0	10.4 ± .6	45 (85)
16192 C	3	5	0	0	0	ND	8 (15)
0318P:[b]							
16192 T	0	20	21	2	0	10.6 ± .6	43 (90)
16192 C	1	3	1	0	0	ND	5 (10)
0803:							
16192 T	1	43	29	6	1	10.5 ± .7	80 (73)
16192 C	1	3	2	0	0	ND	6 (7)
0805/WBC:[c]							
16192 T	3	144	3	0	0	10.0 ± .2	150 (99)
16192 C	1	0	0	0	0	ND	1 (1)
0805/MUS:[c]							
16192 T	1	64	11	1	0	10.2 ± .4	77 (100)
16192 C	0	0	0	0	0	ND	0 (0)
0217:[d]							
16192 T	1	113	4	0	0	10.0 ± .2	118 (100)
16192 C	0	0	0	0	0	ND	0 (0)
0217P:[b]							
16192 T	0	66	1	0	0	10.0 ± .1	67 (100)
16192 C	0	0	0	0	0	ND	0 (0)

[a] These individuals are maternal nephews of NH0284.

[b] For 0318 and 0217, independent analyses were carried out using *Pfu* polymerase, which has a proofreading 3′→5′ exonuclease, rather than the relatively high-error-rate *Taq* polymerase which is used for all other analyses.

[c] For individual 0805, DNA was analyzed from both the WBC/platelet fraction of whole blood (WBC) and from skeletal muscle (MUS).

[d] The D-loop polymorphisms (see fig. 3) and coding region polymorphisms (data not shown) indicate that this is an African mtDNA. The other mtDNAs all belong to European haplogroup U.

thousands—of generations ago. Therefore, the persistent heteroplasmy at nt 16192 in these evolutionarily divergent lineages does not fit the standard model of a single mtDNA mutational event that becomes homoplasmic within a relatively small number of generations.

Heteroplasmy and Tract-Length Variance in Other 16189C mtDNAs

We have obtained preliminary data for other 16189C mtDNAs to determine whether heteroplasmy occurs at other sites within this simple-repeat tract. More than 100 total clones have been sequenced for six individuals whose mtDNA carries the 16189C/16186T combination of D-loop polymorphisms and the same number for five individuals whose mtDNA carries the 16189C/16187T combination. For both genotypes, there was no evidence of heteroplasmy at nt 16186 or nt 16187. All of the 16189C/16187T clones had the same tract length of 10 bp. Tract length, however, showed a low level of variability among the 16189C/16186T clones: one clone had a length of 9 bp, and one had a length of 11 bp. These results further support a relationship between heteroplasmy at nt 16192 and tract length expansion/contraction.

In addition to the results for the mtDNA from 0159 and her maternal relatives (table 3), we have obtained sequence for ~300 16189C/16192C clones from 17 unrelated individuals. Although there was tract length variation, none of these clones carried the 16192T allele. Taken together, the sequence analysis of >600 clones yields no evidence that the 16189C polymorphism—and the accompanying tract-length variation—measurably increases the C→T mutation frequency at nt 16192.

There is one other result from these analyses that merits brief mention. In the rCRS, the nt 16184–16193 polypyrimidine repeat is immediately preceded by a stretch of 4 A residues (nt 16180–16183). Our results indicate that the length of this A repeat also varies as a consequence of the 16189C polymorphism, and the length of this short repeat is—as a first approximation—related to the length of the polypyrimidine tract. Thus, in individual 0082, the modal length of the C repeat is 10 bp, and there are 4 As in all clones. In individuals 0078 and 0119, however, the model length is 11 or 12 Cs, and all clones (>70) have 3 As. Finally, in individual 0066, the modal length is 13 Cs, and all clones (>30) have 2 As. It thus appears that the overall length of the

Table 3

Polypyrimidine Tract Lengths in Individual 0159 and Her Offspring

Family Member and Allele Status	No. of Clones of Tract Length (in residues)					Mean Tract Length ± SD	Total Clones (%)
	≤9	10	11	12	≥13		
0159:							
16192 T	0	0	0	0	0	ND	0 (0)
16192 C	5	15	9	6	3	10.7 ± 1.1	38 (100)
0129:[a,b]							
16192 T	0	0	0	0	0	ND	0 (0)
16192 C	13	68	30	19	2	10.5 ± .9	132 (100)
0283:[b,c]							
16192 T	0	0	0	0	0	ND	0 (0)
16192 C	24	57	19	10	5	10.3 ± 1.0	115 (100)

[a] Daughter of NH0159.

[b] Data from WBC/platelet DNA and from fibroblasts have been pooled.

[c] Son of NH0159.

A+C repeat sequence is under some selective constraint. The same results were obtained by Bendall and Sykes (1995).

Sequence Analysis of 16189C Subclones

It is difficult to determine the true mutation frequency of an mtDNA site, in large part because of the large number of mtDNA molecules per cell. We thus attempted to obtain further information on the heteroplasmy at nt 16192 through the analysis of subclones. Single M13 clonal populations were used for subclone isolation after one passage in a bacterial host. The resulting subclones were then used for sequence analysis of the appropriate D-loop region (table 4). Two 16189C/16192C clones in which the tract length was 10 residues were analyzed. No C→T mutations were observed, although one of the clones gave rise to a low frequency of subclones in which a single–base pair contraction had occurred. A 16189C/16192T clone showed neither reversion at nt 16192 nor expansion/contraction of tract length; however, a clone of the same genotype but of greater tract length (13 residues) gave rise to both expanded and contracted subclones. (Again, there were no reversion events.)

Other Examples of Expansion/Contraction in Human mtDNA

Our sequencing analyses of the mtDNA control region have confirmed that there are other simple-repeat sequences that undergo expansion/contraction. Among the ~300 control-region sequences (almost entirely from individuals of European ethnicity) that we have determined, we have identified the following additional sites of expansion/contraction. The rCRS contains a (C_7)–T–(C_5) polypyrimidine tract starting at nt 303, although the (C_5) portion starting at nt 311 is (C_6) in all other control regions that we have analyzed (Andrews et al. 1999). Both C tracts are stable at those lengths, but a substantial proportion of individuals (~40%) have a C_8 tract at nt 303 that shows intraindividual heteroplasmy and expansion to longer tract lengths. In addition, we have observed one subgroup of haplogroup J mtDNAs in which there are two types of the nt 303 polypyrimidine tract. In one group of individuals, the sequence is (C_8)–T–(C_6), whereas in the other group, it is (C_7)–(T_2)–(C_6). For both repeat types, individuals are often heteroplasmic for expansion of the subtract starting nt 303, and there is thus no indication that the extra T residue affects expansion. Second, there is a C_6 tract starting at nt 568 in the rCRS. In all mtDNAs of European haplogroup I, and in a few non-I haplogroup mtDNAs, the tract length becomes longer and is always heteroplasmic within an individual (Howell et al. 1998; N. Howell, unpublished data; see also Torroni et al. 1994). The mean repeat length in individuals is 9–11 bp with a fairly marked dispersion about the mean. Finally, there is an rCRS $(CA)_5$ repeat that starts at nt 514. We have observed that the number of repeat units is usually five; it can also be four, six, or seven with intraindividual heteroplasmy at the longer repeat lengths. We have not observed that expansion/contraction of any of these simple repeats is associated with hypermutation at any nucleotide within the repeat (data not shown).

Table 4

Pyrimidine Tract Lengths of Subclones

		No. of Clones of Tract Length (in residues)						
Clone	Allele Status	9	10	11	12	13	14	Total Clones
1:								
Original	16192 T	0	0	0	0	0	0	0
16192 C/10	16192 C	3	57	0	0	0	0	60
10:								
Original	16192 T	0	59	0	0	0	0	59
16192 T/10	16192 C	0	0	0	0	0	0	0
16:								
Original	16192 T	0	60	0	0	0	0	60
16192 T/10	16192 C	0	0	0	0	0	0	0
56:								
Original	16192 T	0	0	0	6	52	1	59
16192 T/13	16192 C	0	0	0	0	0	0	0

The 16189 T→C polymorphism is one of the most rapidly occurring single-nucleotide substitutions in the human mitochondrial genome (Excoffier and Yang 1999; Meyer et al. 1999). Other sequence changes that generate longer simple-repeat tracts, however, do not seem to occur with measurable frequency. For example, a T→C substitution at nt 310 would generate a (C_{14}) tract, but we have not observed this polymorphism among the 200 control-region sequences that we have determined, even among the hundreds of cloned fragments that have been sequenced (data not shown). In a similar fashion, a T→C substitution at nt 3572 (ND1 gene) would generate a (C_{11}) tract, whereas a C→T substitution at nt 9485 would generate a (T_{10}) tract in COX3 gene. We have sequenced these regions from ~70 individuals, and these substitutions have not been detected (R. M. Andrews, P. F. Chinnery, D. M. Turnbull, N. Howell, unpublished data). Therefore, there is no general selection during mtDNA evolution for the generation of longer simple-repeat sequences. Instead, as would be predicted, they seem to be well tolerated in the noncoding control region but absent in the coding region.

Discussion

These results are the first reported example of persistent heteroplasmy, at both the genealogical and phylogenetic levels, for a human mtDNA substitution mutation. To explain these results, we hypothesize that, when there is a C residue at nt 16189 and a T residue at nt 16192, the reversion rate at the latter site is markedly elevated (hypermutation; see fig. 2). That is, when there is a C at 16189 and a T at 16192, there is a marked tendency for the latter nucleotide to "revert" to the C allele, presumably through nucleotide misincorporation during replication. The frequency of reversion at nt 16192 is sufficiently high that the usual pattern of rapid segregation of heteroplasmic mutations to the homoplasmic state—principally because of the developmental bottleneck—is overcome, as is also the case for expansions/contractions of simple-repeat sequences (for example, table 3; see also the discussion in Bendall and Sykes 1995). It cannot yet be ascertained whether this hypermutation process occurs only in the germline mtDNA or whether it also occurs in somatic tissues.

One important concern that must be addressed is the possibility that our results are an artifact that arises during PCR amplification and cloning of the mtDNA fragments. It is widely recognized that *Taq* polymerase has a relatively high error rate of both substitutions and single-bp frameshifts (for example, Tindall and Kunkel 1988); however, this possibility can be rejected for the following reasons. In the first place, we analyzed 16189C/16192T mtDNAs in which neither expansion/contraction nor 16192 hypermutation occurred (table 2). If there were some technical artifact, then one would expect that it would be consistent among all 16189C/16192T mtDNAs. Second, as noted in the last subsection of our Results, we do not detect hypermutation of any nucleotide in other simple-repeat sequences, including some that undergo expansion/contraction. Third, whereas expansion/contraction does occur on passage of recombinant M13 clones, the frequency of events is much lower than that among the mtDNA fragments (compare the results in tables 1–3 with those in table 4). Fourth, and most important, we obtained the same results after PCR amplification with *Pfu* polymerase, which has a much lower error rate. Both an mtDNA with a low frequency of hypermutation and one with a high frequency were tested (table 2).

The 16189C polymorphism generates a 10-bp simple-repeat sequence that undergoes rapid expansion and contraction (see also Bendall and Sykes 1995; Marchington et al. 1997). The variation in tract length apparently arises through replication slippage (Hauswirth et al. 1984), and we propose that this process is also responsible for the hypermutation at nt 16192. The hypermutation process appears to be very specific, and when there is a C at position 16189, there is no detectable tendency for any of the other C residues in the tract to mutate to other nucleotides. It is interesting to note that when there is a C at 16189 and a T at 16186 or at 16187, there is neither rapid expansion/contraction of the tract nor hypermutation of the T. In both cases, 16189C/16187T and 16189C/16186T, the length of the homopolymeric tract is shorter (by two and one residues, respectively) than that in 16189C/16192T mtDNAs.

The sequence analysis of subclones (table 4), although these results are only preliminary, may provide some important information on the expansion/contraction of simple mtDNA repeats. Hauswirth et al. (1984) studied a different repeat sequence in the bovine mtDNA control region. They observed that, when a recombinant clone was "passaged" multiple times in a bacterial host, length variation was regenerated (see their fig. 4). Our subclone analyses show a similar phenomenon, although we find that expansion/contraction was more frequent in a subclone with a "long" starting-length repeat. One may speculate that each length variant of the 16189C simple-repeat sequence has its own frequencies of expansion and contraction, with an optimum length, in terms of stability, near 10 residues. That suggestion would explain the relative stability of tract length, and the low level of 16192 heteroplasmy, in some of the individuals analyzed here (for example, 0217 in table 2). At the same time, it must be kept in mind that the initial 16189 T→C mutation should generate the most stable repeat sequence and that, as a result, one would not predict the distribution of tract lengths that are observed. Although bacterial systems are unlikely to be perfect mimics of the in situ mitochondrial genetic system, they may allow us to analyze the expansion/contraction process without the confounding effects of the developmental bottleneck.

The rapid expansion/contraction of the polypyrimidine tract in 16189C mtDNAs, coupled with the hypermutation at nt 16192, appears to constitute a *quasispecies*. The quasispecies concept has been developed to explain the complex population genetic behavior of the rapidly mutating RNA viruses (reviewed by Domingo and Holland [1988] and by Eigen and Biebricher [1988]). That is, these heteroplasmic mtDNA lineages (and it is probably safe to assume that the phenomenon extends to the cellular level) consist of multiple molecular species that are nonidentical but related. In the case described here, there are two subpopulations defined by the allele status at 16192 and, within each subpopulation, there are multiple tract-length variants. A key component of the quasispecies is that selection acts on the distribution as a whole, with varying selection on each genotype. Thus, our results (including those in table 4) suggest that tract lengths of 10–11 residues are relatively stable, whereas shorter and longer variants appear to be more prone to expansion and contraction, respectively. The relationship of the mtDNA hypermutation process described here to the quasispecies model will require further experimental and theoretical development, but the concept appears to be one that may be generally applicable to the rapidly evolving mitochondrial genome.

The present results indicate a context-specific hypermutation process in the human mtDNA control region. Context effects on ultraviolet mutagenesis in repair-deficient human cells have also been described, although the basis for such effects is not yet clear in that system (Levy et al. 1996). Expansion/contraction of simple repeats associated with hypermutation have also been described for the Friederich ataxia *X25* (Bidichandani et al. 1999) and the colorectal cancer *APC* (Laken et al. 1997) genes. It is not clear that hypermutation is a common occurrence in the mtDNA, although we have observed what appears to be the same phenomenon at nt 16093 of the coding region, but this hypermutation process is not associated with a simple-repeat sequence (N. Howell, I. Kubacka, D. A. Mackey, unpublished data).

As discussed in the Introduction, site variability in divergence rates has complicated the analysis of human mtDNA evolution. Even within the control region, sites vary from invariant to hypermutable as shown here. The rapid expansion/contraction of human mtDNA simple-repeat sequences makes them problematic for phylogenetic analyses, and they generally are omitted. Although omission of substitution hypermutable sites should also be done, the most important ramification of the present studies is that they further caution against overreliance on simple models of molecular evolution that assume *independence* of mtDNA mutations.

Acknowledgments

We thank the LHON family members and the other individuals who supplied DNA samples for their contribution to this research. Particular acknowledgement is also made of the efforts of Drs. Ivan Bodis-Wollner (State University of New York [SUNY] Health Science Center at Brooklyn) and Jerome Sherman (SUNY, New York). We also acknowledge the comments and suggestions of the two anonymous reviewers of a previous version of this manuscript. This research was supported by research grants from the National Science Foundation (BCS-9910871) and the John Sealy Endowment Fund that were awarded to N.H.

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin JJ, Eperon IC, et al (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge Reference Sequence for human mitochondrial DNA. Nat Genet 23:147

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

Bendall KE, Macauley VA, Baker JR, Sykes BC (1996) Heteroplasmic point mutations in the human mtDNA control region. Am J Hum Genet 59:1276–1287

Bidichandani SI, Purandare SM, Taylor EE, Gumin G, Machkas H, Harati Y, Gibbs RA, et al (1999) Somatic sequence variation at the Friedreich ataxia locus includes complete contraction of the expanded GAA triplet repeat, significant length variation in serially passaged lymphoblasts and enhanced mutagenesis in the flanking sequence. Hum Mol Genet 8:2425–2436

Domingo E, Holland JJ (1988) High error rates, population equilibrium, and evolution of RNA replication systems. In: Domingo E, Holland JJ, Ahlquist P (eds) RNA genetics. Vol 3: Variability of RNA genomes. CRC Press, Boca Raton, pp 3–36

Eigen M, Biebricher CK (1988) Sequence space and quasispecies distribution. In: Domingo E, Holland JJ, Ahlquist P (eds) RNA genetics. Vol 3: Variability of RNA genomes. CRC Press, Boca Raton, pp 211–245

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

Hasegawa M, Di Renzo A, Kocher TD, Wilson AC (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. J Mol Evol 37:347–354

Hauswirth WW, Van De Walle MJ, Laipis PJ, Olivo PD (1984) Heterogeneous mitochondrial D-loop sequences in bovine tissue. Cell 37:1001–1007

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci USA 92:532–536

Howell N (1999) Human mitochondrial diseases: Answering questions and questioning answers. Int Rev Cytol 186: 49–116

Howell N, Bogolin C, Jamieson R, Marenda DR, Mackey DA (1998) mtDNA mutations that cause optic neuropathy: how do we know? Am J Hum Genet 62:196–202

Howell N, Kubacka I, Halvorson S, Howell B, McCullough DA, Mackey D (1995) Phylogenetic analysis of the mitochondrial genomes from Leber hereditary optic neuropathy pedigrees. Genetics 140:285–302

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

Howell N, Mackey D (1997) Response to Macauley et al. Am J Hum Genet 61:986–990

Howell N, Xu M, Halvorson S, Bodis-Wollner I, Sherman J (1994) A heteroplasmic LHON family: tissue distribution and transmission of the 11778 mutation. Am J Hum Genet 55:203–206

Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in human and chimpanzees: control region and a protein-coding region. In: Osawa S, Honjo T (eds) Evolution of life: fossils, molecules, and culture. Springer, Tokyo, pp 391–413

Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, et al (1997) Familial colorectal cancer in Ashkenazim due to a hypermutable tract in *APC*. Nat Genet 17:79–83

Levy DD, Magee AD, Seidman MM (1996) Single nucleotide positions have proximal and distal influence on UV mutation hotspots and coldspots. J Mol Biol 258:251–260

Lightowlers RN, Chinnery PF, Turnbull DM, Howell N (1997) Mammalian mitochondrial genetics: heredity, heteroplasmy and disease. Trends Genet 13:450–455

Macauley VA, Richards MB, Forster P, Bendall KE, Watson E, Sykes BC, Bandelt H-J (1997) mtDNA mutation rates: no need to panic. Am J Hum Genet 61:983–986

Marchington DR, Hartshorne GM, Barlow D, Poulton J (1997) Homopolymeric tract heteroplasmy in mtDNA from tissues and single oocytes: support for a genetic bottleneck. Am J Hum Genet 60:408–416

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

Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, et al (1997) A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet 15:363–368

Tindall KR, Kunkel TA (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. Biochemistry 27: 6008–6013

Torroni A, Lott MT, Cabell MF, Chen Y-S, Lavergne L, Wallace DC (1994) mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to recurrent somatic duplication in the D-loop region. Am J Hum Genet 55:760–776

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