

# Lineage-Specific Selection in Human mtDNA: Lack of Polymorphisms in a Segment of MTND5 Gene in Haplogroup J

Jukka S. Moilanen, Saara Finnilä, and Kari Majamaa

Biocenter and Department of Neurology, University of Oulu, Oulu, Finland

Human mitochondrial DNA (mtDNA) is a nonrecombining genome that codes for 13 subunits of the mitochondrial oxidative phosphorylation system, 2 rRNAs, and 22 tRNAs. Mutations have accumulated sequentially in mtDNA lineages that diverged tens of thousands of years ago. The genes in mtDNA are subject to different functional constraints and are therefore expected to evolve at different rates, but the rank order of these rates should be the same in all lineages of a phylogeny. Previous studies have indicated, however, that specific regions of mtDNA may have experienced different histories of selection in different lineages, possibly because of lineage-specific interactions or environmental factors such as climate. We report here on a survey for lineage-specific patterns of nucleotide polymorphism in human mtDNA. We calculated molecular polymorphism indices and neutrality tests for classes of functional sites and genes in 837 human mtDNA sequences, compared the results between continent-specific mtDNA lineages, and used two sliding window methods to identify differences in the patterns of polymorphism between haplogroups. A general correlation between nucleotide position and the level of nucleotide polymorphism was identified in the coding region of the mitochondrial genome. Nucleotide diversity in the protein-coding sequence of mtDNA was generally not much higher than that found for many genes in nuclear DNA. A comparison of nonsynonymous/synonymous rate ratios in the 13 protein-coding genes suggested differences in the relative levels of selection between haplogroups, including the European haplogroup clusters. Interestingly, a segment of the *MTND5* gene was found to be almost void of segregating sites and nonsynonymous mutations in haplogroup J, which has been associated with susceptibility to certain complex diseases. Our results suggest that there are haplogroup-specific differences in the intensity of selection against particular regions of the mitochondrial genome, indicating that some mutations may be non-neutral within specific phylogenetic lineages but neutral within others.

## Introduction

Human mitochondrial DNA (mtDNA) is a small (16.6-kb) circular genome that codes for 13 subunits of the mitochondrial oxidative phosphorylation system (*MTATP6*, *MTATP8*, *MTCO1*, *MTCO2*, *MTCO3*, *MTCYB*, *MTND1*, *MTND2*, *MTND3*, *MTND4*, *MTND4L*, *MTND5*, *MTND6*), 2 rRNAs (*MTRNR1*, *MTRNR2*), and 22 tRNAs. MtDNA is inherited maternally and it is not subject to significant recombination (Elson et al. 2001); therefore, mutations have accumulated sequentially in mtDNA lineages that diverged tens of thousands of years ago. Groups of related haplotypes (haplogroups) and clusters of these groups (haplogroup clusters) define branches of the phylogenetic tree for mtDNA, and many previous studies have indicated that the geographical distribution of haplogroups in aboriginal populations is continent-specific (Torroni et al. 1996). The African haplogroup cluster L is the most ancient of all clusters. Two lineages (M and N) diverged from L, presumably in northeastern Africa approximately 65,000 years ago, and the European haplogroup clusters HV, JT, KU, and IWX (Finnilä, Lehtonen, and Majamaa 2001) were subsequently derived primarily from N, whereas M and N contributed equally to the radiation of mtDNA into Asian-specific haplogroups A, C, D, G, Z, and Y. The American continent was populated from northeastern Asia by individuals with haplogroups A, B, C, and D (Wallace, Brown, and Lott 1999; Mishmar et al. 2003).

The oxidative phosphorylation system is located in

the inner mitochondrial membrane, and it consists of five protein complexes (I, II, III, IV, and V) in which mitochondrial- and nuclear-encoded subunits are in close proximity and subject to mutual interactions (Schmidt et al. 2001). Oxidative phosphorylation has a central role in cellular energy metabolism, because it is the final common pathway in the production of adenosine triphosphate (ATP) from glucose, and it is likely that the functionality of the various subunits is maintained by purifying selection. The genes in mtDNA are subject to different selective constraints (Pesole et al. 1999; Tourasse and Li 2000) and are therefore expected to evolve at different rates, but the rank order of these rates should be the same in all lineages of a phylogeny (Rand 2001). Analyses of *Drosophila* mtDNA have indicated, however, that specific regions of mtDNA have experienced different histories of selection in different lineages (Ballard 2000a, 2000b). Lineage-specific interactions have been suggested to explain this finding (Blier, Dufresne, and Burton 2001). Such interactions between gene products, either mitochondrial-nuclear or mitochondrial-mitochondrial, might expose a region of the mtDNA genome to unusual selection, which should result in a local decrease in polymorphism or a shift in the allele frequency spectrum that is not present in other lineages. Alternatively, environmental factors such as climate could also impose different selective forces in different lineages, and lineage-specific differences have recently been identified in the nonsynonymous/synonymous rate ratio between human mtDNA lineages from the tropical, temperate, and arctic zones (Mishmar et al. 2003).

We report here on a systematic search for patterns suggesting lineage-specific selective influences in 837 human mtDNA sequences (Ingman et al. 2000; Finnilä, Lehtonen, and Majamaa 2001; Maca-Meyer et al. 2001; Herrnstadt et al. 2002). We aligned the sequences and

Key words: human mitochondrial DNA, rates of evolution, selective constraints, structural/functional domains, phylogenetics.

E-mail: kari.majamaa@oulu.fi.

*Mol. Biol. Evol.* 20(12):2132–2142, 2003

DOI: 10.1093/molbev/msg230

*Molecular Biology and Evolution*, Vol. 20, No. 12,

© Society for Molecular Biology and Evolution 2003; all rights reserved.

calculated molecular polymorphism indices and neutrality tests for classes of functional sites and genes and compared the results between continent-specific mtDNA lineages. Furthermore, nonsynonymous/synonymous rate ratios for protein-coding genes were calculated, and a sliding window analysis of allele frequency distribution and a maximum-likelihood sliding window method were used to identify differences in the patterns of polymorphism between haplogroups. The results of the individual analyses suggest that there are haplogroup-specific differences in the intensity of selection against particular regions of the mitochondrial genome.

## Materials and Methods

### MtDNA Sequence Data Sets

For this study, 86 mtDNA sequences derived from diverse populations (Ingman et al. 2000; Maca-Meyer et al. 2001) were downloaded from GenBank in May 2003 (accession numbers AF346963–AF347015 and AF381981–AF382013). AF381992.2, AF381994.2, AF382008.2, and AF382009.2 had been updated, whereas the remaining sequences were initial versions. The sequences were assigned to haplogroup clusters according to phylogenetic networks based on coding region variation (Finnilä, Lehtonen, and Majamaa 2001; Herrnstadt et al. 2002). Nine Asian or Pacific sequences could not be confidently positioned on these networks but were included in the Asian group of sequences on the grounds of their geographic origin. 192 sequences were population controls from Finland (Finnilä, Lehtonen, and Majamaa 2001), but sequences belonging to haplogroup Z were assigned to the Asian group (Meinilä, Finnilä, and Majamaa 2001). 560 coding-region sequences (Herrnstadt et al. 2002) from population controls or patients with neurodegenerative disease in the U.K. and U.S.A. were downloaded from the MitoKor Web site in May 2003. Sequence 104 was excluded because of its HeLa origin, and sequence 327 was included in the total set but not assigned to any of the European haplogroup clusters HV, JT, KU, or IWX. The Finnish and GenBank sequences were complete, whereas the MitoKor sequences spanned only the coding region (nucleotide position [np] 577–16023). Of the sequences, 82 belonged to the African haplogroup cluster L and 108 to the various Asian haplogroups; 646 sequences belonged to the European haplogroup clusters HV (H 263, V 37), KU (K 60, U 82), JT (J 53, T 60), and IWX (I 29, W 46, X 16).

The sequences were aligned by assuming that indel events changing the length of a repeat sequence had occurred at the 3' end of the repeat sequence, and functional regions, genes, codon positions and nondegenerate, two-fold degenerate and fourfold degenerate sites (Li 1997, p. 87) were identified according to the MITOMAP reference sequence, a modified version of the 2001 Revised Cambridge Reference Sequence (Andrews et al. 1999) and mtDNA function locations (available at <http://www.mitomap.org>). The full alignment of the 837 sequences was 16,620 bp in length. Sites with alignment gaps representing indels were excluded from the analyses, making the length of the alignment 16,526 bp (coding

region 15,420, D-loop 1,106 bp). The sequence data were stored in a relational SQL database and partitioned into sequence data sets which were defined by haplogroups and classes of sites. For each sequence, the nucleotides belonging to a given class of sites were concatenated and extracted from the database by methods provided by the SQL query language and the programming language Perl.

The 192 Finnish sequences have been deposited in the GenBank database (accession numbers AY339402–AY339593).

### Diversity Indices and Neutrality Tests

The sequence data sets defined by haplogroups and classes of sites were analyzed in terms of nucleotide polymorphism indices and phylogenetic tests of neutrality. According to the coalescent theory, the population mutation parameter for mtDNA is  $\theta = 2N\mu$ , where  $N$  is the effective population size and  $\mu$  is the neutral mutation rate per generation (Fu 1997). The methods used for estimating  $\theta$  in a set of sequences include  $k$  or  $\theta_\pi$ , the average number of pairwise nucleotide differences (Tajima 1983), Watterson's estimate ( $\theta_s$ ), based on the number of polymorphic sites in the sample (Watterson 1975), and  $\theta_{\eta_s}$ , which is based on the number of singleton mutations ( $\eta_s$ ) in the sample (Fu and Li 1993). Under neutrality and certain demographic assumptions (Nielsen 2001) these three estimates should yield similar values (Simonsen, Churchill, and Aquadro 1995). However,  $\theta_s$  and  $\theta_{\eta_s}$  are affected by the presence of low-frequency alleles in the sample, whereas these have little impact on  $\theta_\pi$ , leading to differences between the estimates when selection is present (Fu and Li 1993). These differences form the basis of phylogenetic tests of neutrality, including Tajima's  $D$ , which is based on the difference between  $\theta_\pi$  and  $\theta_s$  and its variance (Tajima 1989), Fu and Li's  $F^*$ , which is based on the difference between  $\theta_\pi$  and  $\theta_{\eta_s}$  and its variance (Fu and Li 1993), and Fu's  $F_s$ , which is based on the probability of the observed number of haplotypes or more being observed under neutrality (Fu 1997). An excess of low-frequency alleles as compared with that expected under neutrality results in negative values for these tests, whereas a relative lack of low-frequency alleles results in a shift toward positive values.

Nucleotide diversity indices and neutrality tests, including  $\theta_\pi$ ,  $\theta_s$ ,  $\theta_{\eta_s}$ ,  $D$ ,  $F^*$ , and  $F_s$ , were calculated for each sequence data set. The three  $\theta$  estimators were standardized by the length of the sequence analyzed (Li 1997, pp. 237–242). The standard error of nucleotide diversity ( $\pi$ ;  $\theta_\pi/\text{site}$ ) was computed to include the variance over the stochastic process (Nei 1987, pp. 254–258), and both this and the standard error of  $\theta_s/\text{site}$  were computed assuming neutrality and population equilibrium (Li 1997, pp. 237–242). DnaSP 3.53 (Rozas and Rozas 1999), Arlequin 2.000 (available at <http://anthro.unige.ch/arlequin>), and dnastats 0.92 (available at <http://cc.oulu.fi/~jukkamoi/mtres/>) were used for the analyses with convergent results, except that DnaSP and Arlequin sometimes provided conflicting results for Fu's  $F_s$  test and ran into problems with the largest sequence data sets. In such instances the results obtained with dnastats were

preferred, as the source code is available for review. The statistical significance levels for D and F\* were obtained using the DnaSP default mode, which does not involve coalescent simulations.

#### Nonsynonymous/Synonymous Rate Ratios in Protein-Coding Genes

The ratio of the number of nonsynonymous mutations per nonsynonymous sites ( $d_N$ ) to the number of synonymous mutations per synonymous sites ( $d_S$ ) indicates the level of selection against nonsynonymous mutations relative to synonymous ones. In the presence of positive selection  $d_N$  should be higher than  $d_S$  (Nielsen 2001).  $d_N/(d_N + d_S)$  distributions in pairwise comparisons between sequences were calculated (Nei and Gojobori 1986) for the 13 protein-coding genes in the sequence sets defined by the continent-specific haplogroup clusters (African, Asian, European, HV, JT, KU, and IWX) using DnaSP, and these distributions were plotted and the significances of the differences between lineages in the location of the distribution were assessed using the Kruskal-Wallis rank-sum test as implemented in R 1.4.1 (Ihaka and Gentleman 1996).

#### Sliding Window Analysis of Allele Frequency Distribution

Neutrality tests are summaries of the allele frequency spectrum (Przeworski, Hudson, and Di Rienzo 2000), and as such they may be used to identify genome regions that show patterns of nucleotide substitution that are not consistent with the average pattern for the genome (Nielsen 2001; Rand 2001). Coding region sequence data sets were analyzed by a sliding window method as implemented in DnaSP and dnastats, and using the three  $\theta$  estimators and D, F\*, and Fs. By definition, the sequences in each data set shared their genetic history, because the sets were defined by haplogroups. Therefore, we assumed that lineage-specific local anomalies in the allele frequency spectrum would indicate selective forces specific to the corresponding lineage.

In the sliding window method, the relevant statistics are computed for a region (window) of nucleotide positions and plotted at the median position of that window. A low number of polymorphisms within the window results in high stochastic fluctuations or discontinuity for the diversity indices and neutrality tests, imposing a practical lower limit on the window size. Patterns narrower than the window become smoothed, and therefore the selection of the window size is a compromise between stochastic patterns and the desired resolution. In the present analysis a window size of 1,000 bp was used and the window was moved in 40-bp steps. This decision was based on preliminary analyses which suggested that markedly narrower window sizes would have caused discontinuity in the sets with the smallest numbers of sequences.

The hypothesis of correlation between sliding window estimates and nucleotide position was assessed by linear regression as implemented in R. The significance of each correlation was estimated from nonoverlapping

1,000-bp segments, because the points in each sliding window curve are autocorrelated because of the partial overlap of the windows.

As expectations and variances of summary statistics for allele frequency distribution may depend on the demographic model, small variations in these statistics should not be interpreted as evidence for selection (Nielsen 2001). Therefore, one of the most extreme lineage-specific local deviations in allele frequency distribution was analyzed in more detail to assess the possible causes of the observed deviation. A median network (Bandelt et al. 1995) was constructed manually for the respective gene, and segregating sites within and outside the observed pattern were identified. Amino acid translations of the sequences were obtained using the Bio::PrimarySeqI interface of Bioperl (Stajich et al. 2002) and nonsynonymous mutations were subsequently identified. Nonsynonymous and synonymous mutations were counted as differences relative to the reference sequence. Fisher's two-tailed exact test was used to assess the hypothesis that the frequencies of segregating sites and nonsynonymous and synonymous mutations within and outside the observed anomalous region were distributed evenly between haplogroups.

#### Maximum-Likelihood Sliding Window Analysis

The analysis of allele frequency distribution does not assess the significance of the patterns identified and, furthermore, the best resolution of the analysis is limited by the window size selected. Therefore, the coding region sequence data sets were also analyzed by a maximum-likelihood sliding window method which identifies regions which do not fit with a single phylogenetic topology and nucleotide substitution process along the entire sequence. This method has good spatial resolution and it also assesses the significance of findings, but it only detects fast-evolving regions; i.e., invariable regions subject to purifying selection may not be detected. Regions subject to an increased rate of evolution and recombination should be detectable, however. The method uses a maximum-likelihood phylogeny to calculate the likelihoods for each site in the sequence data. A measure (Q) of the average likelihood of the window with respect to the rest of the sequence is subsequently calculated for windows of varying sizes and positions. The maximum values of Q are associated with regions showing low likelihood, given the global maximum-likelihood model. Significance is tested assuming a normally distributed null distribution of maximized Q, which allows calculation of the Z-values corresponding to a given significance level using a Monte Carlo simulation and Bonferroni inequality (Grassly and Holmes 1997).

Sequence data sets consisting of tRNA and rRNA genes and nondegenerate, twofold degenerate and fourfold degenerate sites of protein-coding genes were analyzed separately, as they are presumably subject to different selective constraints and their transition/transversion rates may differ. Genes encoded by the L-strand (*MTND6* and eight tRNA genes) were excluded in order to avoid

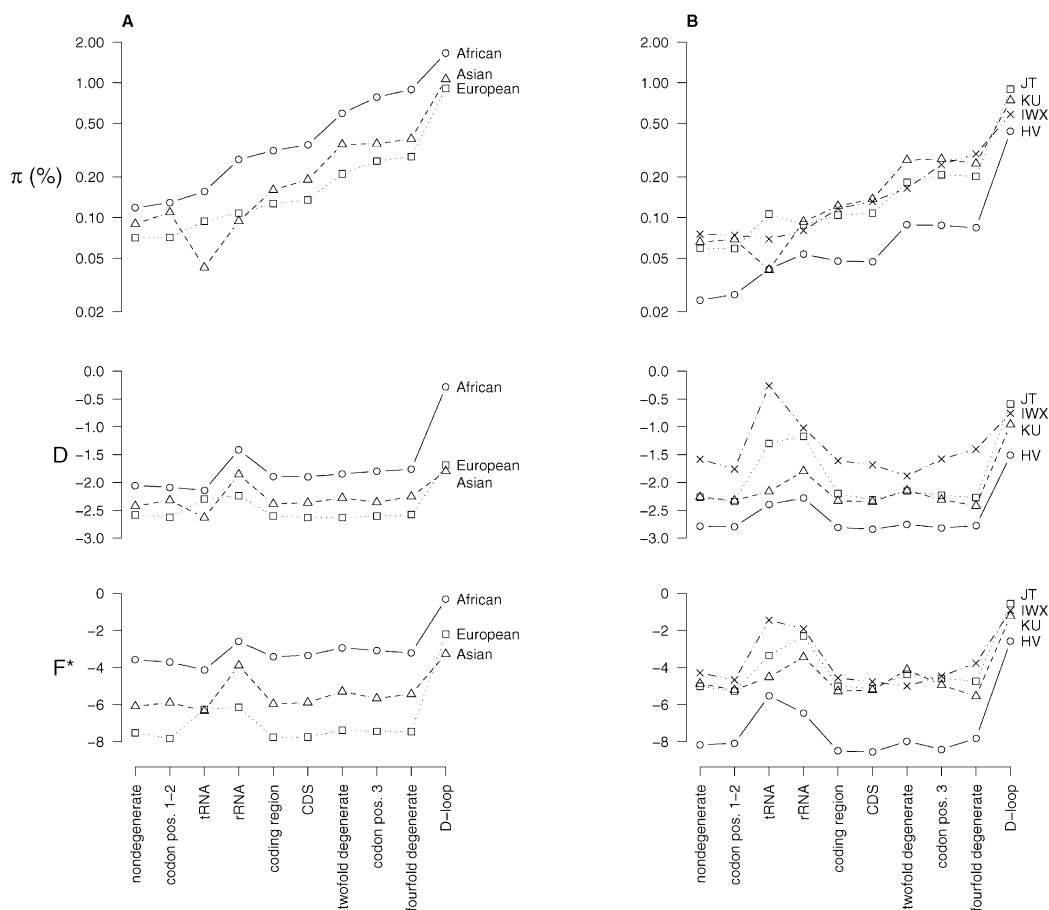


FIG. 1.—Comparison of nucleotide diversity ( $\pi$ ) and neutrality tests between classes of sites (x-axis) in African, Asian, and European mtDNA haplogroup clusters (A) and in the European haplogroup clusters HV, JT, KU, and IWX (B). Coding region (np 577–16023) was available for all sequences and it includes all classes of sites except the D-loop (np 16024–576), which was not present in the 559 MitoKor sequences. The number of D-loop sequences in each cluster was therefore lower than the number of coding region sequences. Numerical data and additional statistical parameters are available as online Supplementary Material.

pooling template and non-template DNA. Maximum-likelihood trees were reconstructed for each sequence data set using TREE-PUZZLE 5.0 (Strimmer and von Haeseler 1996), enforcing the HKY substitution model (Hasegawa, Kishino, and Yano 1985) with a uniform rate over all the sites. An African sequence (GenBank AF347015.1) was used as the root for each tree. The transition/transversion parameter (ti/tv) was estimated from the data. Multifurcations were resolved by adding zero-length branches where necessary to convert the trees into bifurcating ones, as required by the subsequent analysis. The sequence data, maximum-likelihood tree, and ti/tv parameter were used as inputs to PLATO 2.11 (Grassly and Holmes 1997). The HKY substitution model and rate homogeneity were assumed, and the number of trials in the Monte Carlo simulations was set at 250. The minimum window size was set at 10. The mutations responsible for each significant pattern found in the analysis of each sequence set were identified, and homologies in the other sets were screened for by considering each pair of mutations within the region, as the presence of the same  $\geq 2$  mutations in  $\geq 2$  lineages might indicate that the region has been subject to ancient recombination.

## Results and Discussion

### Diversity Indices and Neutrality Tests

Analysis of the set of all 837 sequences revealed a consistently high haplotype diversity. The probability that two randomly selected haplotypes would be different was  $\geq 0.9$  in all the subsections of the genome analyzed except the tRNA genes, where haplotype diversity was slightly below 0.8. Such a high diversity is unlikely to occur under neutrality, as indicated by the Fu's  $F$  test, which was  $< -200$  for all classes of sites in the total data set. The estimated minimum number of mutations ( $\eta$ ) was 1.01–1.07 times higher than the number of segregating sites ( $S$ ), because of the small proportion of sites harboring more than two alleles. The proportion of singleton mutations ( $\eta_S/\eta$ ) was 29%–57%, and the average number of pairwise differences ( $k$ ) varied from 1.5 (for tRNA genes) to 26.1 (for complete coding region sequences). A comparison of the three  $\theta$  estimators (standardized by the length of the sequence analyzed) within each sequence data set defined by classes of sites revealed an excess of private mutations (indicated by  $\theta_{\eta_S}/\text{site}$ ) over segregating sites (indicated by  $\theta_S/\text{site}$ ) and even more so over pairwise differences (indicated by  $\pi$ ). In consequence, Tajima's  $D$  and Fu and Li's  $F^*$  yielded highly negative values ( $\leq -2.4$

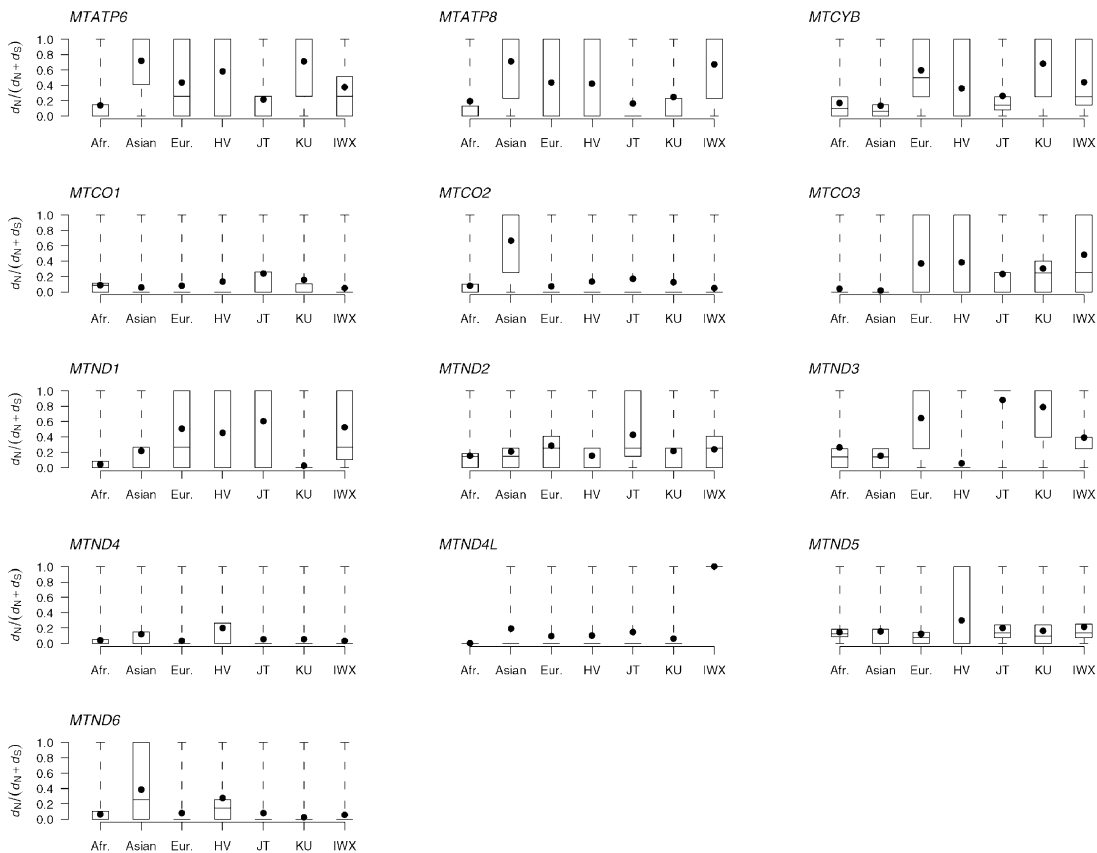


FIG. 2.—Distribution of the relative selective constraints  $d_N/(d_N + d_S)$  in the mtDNA sequence data sets defined by haplogroup clusters and protein-coding genes. The bottom, intermediate, and top horizontal lines in the boxes represent the 25th, 50th, and 75th percentile values, respectively. The dashed line indicates the range of the distribution, and the dot indicates the mean. The locations of distributions for African (Afr.), Asian, and European (Eur.) sets are different in all genes ( $P < 0.0001$  in each gene, Kruskal-Wallis rank sum test). The locations of distributions for HV, JT, KU, and WIX clusters are also different in all genes ( $P < 0.0001$  in each gene, Kruskal-Wallis rank sum test).

and  $\leq -5.8$ , respectively) suggesting selection against deleterious alleles, population expansion, or both (Gerber et al. 2001).

Nucleotide diversity ( $\pi$ ) does not depend on the length of the sequence analyzed or the sample size (Nei 1987, pp. 267–273), and therefore it can be used for direct comparison of the extents of polymorphism between classes of functional sites and lineages. In the total data set  $\pi$  was 0.087%–1.085%, and the rank order of functional classes of sites was in accordance with previous studies, as  $\pi$  increased in the order of nondegenerate sites/codon position 1–2 < tRNA < rRNA < coding sequence (CDS) < twofold degenerate sites < codon position 3 < fourfold degenerate sites < D-loop. The different levels of  $\pi$  between classes of sites are commonly assumed to arise from differential selection against these. This was also supported by the finding that the highest ratio of transitions to transversions was 41.5 (as estimated without correction for multiple hits) in twofold degenerate sites, in which all transitions are synonymous and transversions are non-synonymous (Li 1997, pp. 88, 177–184), whereas the lowest ratio was 8.6 in fourfold degenerate sites, where all mutations are synonymous. Our results therefore support the assumption that selection has a significant role in the evolution of mtDNA as a whole. Moreover, comparison

between our estimates and those reported for nuclear sequence variation (Przeworski, Hudson, and Di Rienzo 2000) indicated that  $\pi$  for mtDNA, excluding the D-loop, is generally not much higher than that found for many genes in nuclear DNA, despite the high mutation rate of mtDNA (Parsons et al. 1997), suggesting that selection limits the rate of fixation of mutations in the coding region of mtDNA.

Nucleotide diversity indices and neutrality tests were then analyzed in the major continent-specific haplogroup clusters to identify possible deviations from the general trends inferred from the total data set. A low haplotype diversity ( $0.424 \pm 0.061$ ) was identified in the Asian tRNA sequence data set, where the estimate for  $\pi$  (0.043%) was also lower than that for nondegenerate sites and lower than that in the African (0.156%) and European (0.094%) sets (fig. 1A). Furthermore, Tajima's D was highly significantly negative ( $P < 0.001$ ) in the Asian tRNA data set. This could imply that mutations within tRNA genes may have been more deleterious in Asian populations and that such mutations have not become fixed as easily as in other populations. Such an effect could result from environmental factors or mitochondrial-mitochondrial or mitochondrial-nuclear interactions (Blier, Dufresne, and Burton 2001) specific to the Asian mtDNA lineages.

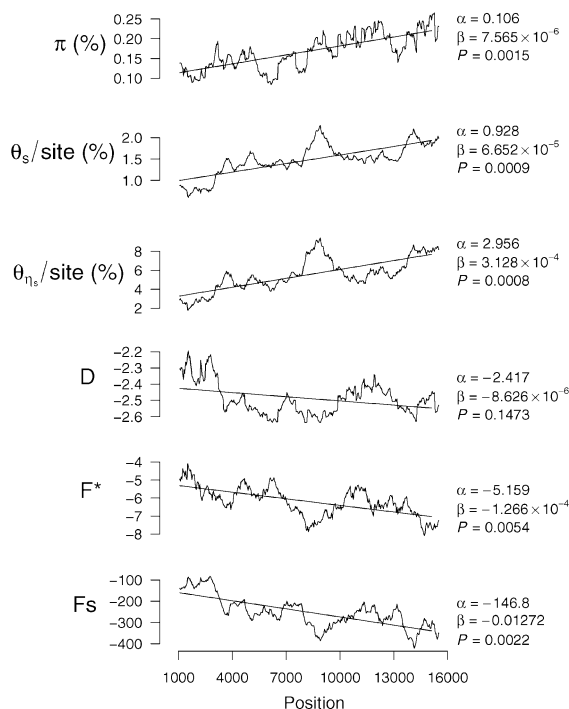


FIG. 3.—Sliding window analysis of allele frequency distribution in the set of all 837 sequences. A 1,000-bp window was moved along the mtDNA coding-region sequence (np 577–16023) in 40-bp steps and  $\pi$ ,  $\theta_s$ /site,  $\theta_{ns}$ /site, D,  $F^*$ , and  $F_s$  were computed for each window and plotted at its median position. Straight line indicates the regression line  $y = \alpha + \beta x$  fitted to the median positions of the non-overlapping 1,000-bp segments, and the parameters of the line and its  $P$ -value (from the  $F$  statistic) are shown on the right side of the figure.

Comparisons of the  $\theta$  estimators and neutrality tests indicated an excess of singleton mutations over segregating sites and pairwise differences in all European haplogroup clusters, but there were differences between the haplogroup clusters (fig. 1B). The HV cluster had a low  $\pi$  for all functional classes of sites and highly negative D and  $F^*$  because of the high number of segregating sites and singleton mutations, whereas the other clusters had generally higher  $\pi$ . In the IWX cluster the difference between the number of segregating sites and pairwise differences did not deviate markedly from that expected under neutrality in most functional classes of sites, but the difference between singleton mutations and pairwise differences did, resulting in nonsignificant values for D and significant values for  $F^*$ . Such differences between haplogroup clusters are likely to represent different population histories. A fairly low haplotype diversity and a low  $\pi$  in the tRNA genes as compared with that in nondegenerate sites were observed in haplogroup clusters KU and IWX, but this pattern was not as evident as in the Asian sequence data set. In the JT cluster the tRNA genes had slightly higher  $\pi$  than the rRNA genes, opposite to the situation in the other clusters (fig. 1B).

#### Nonsynonymous/Synonymous Rate Ratios in Protein-Coding Genes

The relative levels of selection against protein-coding genes in different haplogroup clusters were assessed by

computing  $d_N/(d_N + d_S)$  ratios (fig. 2). The locations of the  $d_N/(d_N + d_S)$  distributions for certain genes differed markedly between lineages. The European JT and African lineages appeared to be better conserved than the remaining ones with respect to *MTATP6* and *MTATP8*, whereas the Asian lineages had relatively high amino acid variation in these genes. In *MTCYB* and *MTCO3*, the European lineages, except JT, had a high  $d_N/(d_N + d_S)$  ratio relative to the African and Asian lineages. *MTCO1* was equally conserved in all the lineages, whereas *MTCO2* and *MTND6* were highly conserved in all lineages except the Asian ones. The JT cluster showed surprisingly high nonsynonymous variation in *MTND2*, JT and KU in *MTND3*, and HV in *MTND5*. Differences in the location of the distributions between lineages were highly significant because of the large numbers of sequences in the analyses.

Differences in the nonsynonymous/synonymous rate ratio between mtDNA lineages selected by reference to geographical origin have recently been identified for certain protein-coding genes in mtDNA (Mishmar et al. 2003), and the authors interpreted the observed differences as evidence of climatic selection. Their analysis was based on 104 sequences, with 32 belonging to the European haplogroup clusters. Our analysis included 53 of these sequences (AF346963–AF347015) and 784 additional sequences, and although some results were similar (e.g., higher amino acid variation in *MTCYB* in European than in African lineages), others were not (e.g., high variation in *MTCO3* in European lineages). Interestingly, we observed marked differences between the European haplogroup clusters, suggesting that gene-specific differences in the nonsynonymous/synonymous rate ratio also exist between lineages which have not been selected according to geographic origin. Such differences cannot therefore be explained solely by climatic selection, unless it is assumed that the European haplogroups have also evolved in geographic regions with different ambient temperatures. Lineage-specific interactions represent a viable alternative hypothesis for the observed differences.

#### Sliding Window Analysis of Allele Frequency Distribution

Sliding window analysis of diversity indices and neutrality tests revealed an unexpected correlation between nucleotide position and the three population mutation parameters and neutrality tests. The correlation was positive and significant for  $\theta_\pi$ ,  $\theta_s$ , and  $\theta_{ns}$  and negative and significant for  $F^*$  and  $F_s$  (fig. 3). Such a correlation could imply a different mutation rate or systematically different selection for genes at the 5' and 3' ends of the human mtDNA coding-region sequence. Different portions of mtDNA remain single-stranded and vulnerable to oxidative damage for different periods during the asymmetric replication of mtDNA, possibly leading to different mutation rates (Nedbal and Flynn 1998). This would result in a correlation that depends on the distance from the L-strand origin at np 5721–5798, but our results did not suggest such a correlation (fig. 3). Moreover, the maximum-likelihood sliding window analysis of fourfold degenerate sites (below) did not support the hypothesis that



FIG. 4.—Sliding window analysis of nucleotide diversity ( $\pi$ ) in the African, Asian, and European haplogroup clusters (A), and of  $\pi$  and Tajima's D in the European haplogroup clusters HV, JT, KU, and IWX (B). A 1,000-bp window was moved along the mtDNA coding-region sequence (np 577–16023) in 40-bp steps, and  $\pi$  and D were computed for each window and plotted at its median position. The bar at the top illustrates the locations of the various genes (shaded, rRNA; solid, tRNA; open, protein-coding). Horizontal bar within *MTND5* in B indicates the range of positions (12478–13611) spanned by the 1,000-bp window at the corresponding D peak in JT. Figures showing complete analyses for  $\pi$ ,  $\theta_s$ ,  $\theta_{ns}$ , D,  $F^*$ , and  $F_s$  are provided as online Supplementary Material.

significant differences in the silent mutation rate would exist between different regions of the genome. Analyses in other species might indicate whether the observed correlation is a phenomenon specific to human mtDNA or a more universal one.

Comparison of continent-specific sequence sets revealed that although  $\pi$  was generally highest in the African set (fig. 1), this was not equally true of all regions of mtDNA (fig. 4A). Higher  $\pi$  among the African sequences was most evident around np 2000–4000, np 5000–9000 and in *MTND5* and *MTND6*, whereas the Asian sequences showed a different pattern, as the highest  $\pi$  was found in *MTND2*, *MTCYB* and the central parts of the genome (np 8000–11000).

Comparison of the four European haplogroup clusters (fig. 4B) revealed more differences between lineages than the comparison of continent-specific sets. Nucleotide diversity was similar in the regions around *MTCO1* and *MTCO2* in all four clusters, but while  $\pi$  in the HV cluster was low between np 10000 and the 3' end, the other clusters had varying patterns of regions with high and low

$\pi$ . The patterns were common to two or three clusters in some regions (e.g., JT, KU, and IWX in *MTND5*); that is, the sliding-window curves were of a similar shape but were discordant in other regions (e.g., KU vs. JT and IWX in *MTND4*). Similar differences between clusters were also observed in  $\theta_s$ ,  $\theta_{ns}$ ,  $F^*$ , and  $F_s$ . These findings suggest that mutations may have accumulated in a similar manner in some lineages, even though they were separated thousands or tens of thousands of years ago, but differently in other lineages.

The sliding window analysis also revealed narrow regions in which the polymorphism indices and allele frequency distribution (as indicated by neutrality tests) differed from the surrounding regions of the genome and from the other haplogroup clusters. This was most conspicuous for np 12478–13611 (amino acids 48–425) within the *MTND5* gene in haplogroup cluster JT (fig. 4B), which was found to result from a lack of segregating sites and singleton mutations in this region, while the number of pairwise differences had not decreased. The median network of the 113 *MTND5* sequences from haplogroup

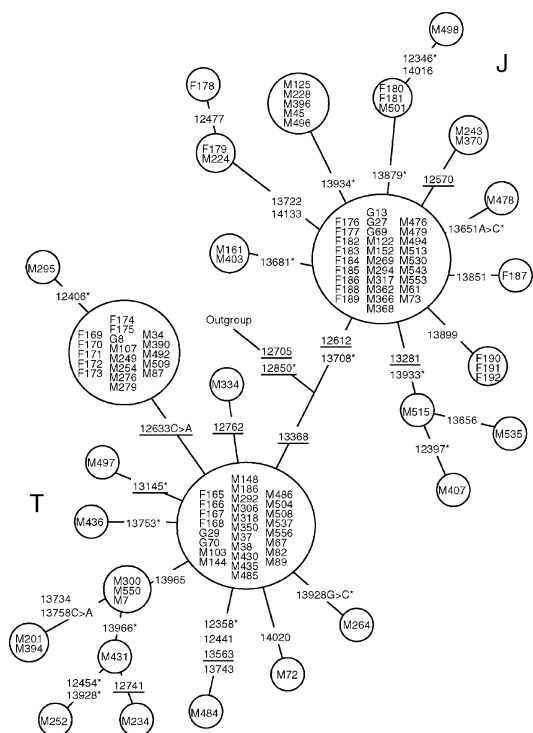


Fig. 5.—Median network of the 113 *MTND5* (np 12337–14148) sequences in the European haplogroups J and T. Numbers indicate transitions and numbers with letter suffixes indicate transversions. An asterisk indicates nonsynonymous mutation; Underline, np is between 12478 and 13611 (amino acids 48–425). Sequence identifiers are shown inside the nodes. F = Finnish sequences, M = MitoKor sequences, G = GenBank sequences (accession numbers: G8 = AF382006.1, G13 = AF382001.1, G27 = AF381987.1, G29 = AF381985.1, G69 = AF346983.1, G70 = AF346982.1). Outgroup, African sequence AF347015.1.

cluster JT (fig. 5) revealed a total of 36 segregating sites, 9 of which (25%) were within np 12478–13611. In the remaining 724 non-JT sequences there were a total of 209 segregating sites, 112 (54%) of which were within this region (JT vs. non-JT,  $P = 0.0019$ , Fisher's exact test). In haplogroup J, 2/16 segregating sites (12.5%) were within np 12478–13611 (J vs. non-JT,  $P = 0.0015$ ), whereas in haplogroup T 5/17 (29%) were within this region (T vs. non-JT,  $P = 0.076$ ). Significant differences were also observed for nonsynonymous mutations in JT and J (JT vs. non-JT, 1/16 vs. 25/66,  $P = 0.016$ ; J vs. non-JT, 0/8 vs. 25/66,  $P = 0.045$ ; T vs. non-JT, 1/8 vs. 25/66,  $P = 0.25$ ) but not for synonymous mutations (JT vs. non-JT, 8/21 vs. 89/147,  $P = 0.061$ ; J vs. non-JT, 3/10 vs. 89/147,  $P = 0.094$ ; T vs. non-JT, 5/11 vs. 89/147,  $P = 0.35$ ). Thus there were fewer segregating sites and nonsynonymous mutations than were to be expected within np 12478–13611 in haplogroup J.

*MTND5* is a hydrophobic polypeptide belonging to the membrane-spanning part of the complex I, and it is probably an important component of the proton translocation machinery. Amino acids 48–425 correspond to several transmembrane helices in the central parts of the subunit, including the large conserved domain between helices IX and XII, which is probably functionally

important and located on the inside of the membrane surface (Mathiesen and Hägerhäll 2002). The carboxy-terminal end of the subunit is nonconserved (Mathiesen and Hägerhäll 2002; Moilanen and Majamaa 2003). We suggest that amino acids 48–425 of *MTND5* might have been under stronger selective pressure in haplogroup J than in others, resulting in the observed lack of mutations. The mtDNA mutations specific to this lineage, or mutations in nuclear-encoded subunits interacting with the central parts of the *MTND5* subunit, if exposing that region of mtDNA to unusually strong selective pressure, could therefore explain our results. Interestingly, haplogroup J is defined by three amino acid replacements, two of which are within the subunits of complex I (*MTND1:Y304H* and *MTND5:A458T*). These mutations are therefore possible candidates for mutual interactions, especially the *MTND5* mutation 13708G > A, which is close to the aberrant segment we found. Given these findings, it is also intriguing that haplogroup J has been associated with susceptibility to certain complex diseases (Torrioni et al. 1997; Reynier et al. 1999; Wallace, Brown, and Lott 1999; Brown et al. 2002) and longevity (De Benedictis et al. 1999; Rose et al. 2001; Niemi et al. 2003).

#### Maximum-Likelihood Sliding Window Analysis

Several sequence data sets included at least one region which did not fit the phylogenetic topology and nucleotide substitution process estimated for the entire set (fig. 6). The largest number of regions with a significantly low likelihood was found in haplogroup cluster KU, where narrow regions within nondegenerate and twofold degenerate sites and rRNA and tRNA genes yielded significant  $Z$  values. The regions identified within nondegenerate and twofold degenerate sites were at different positions in the genome in different sequence sets, with the exception of the 3' end of *MTND5* in the African (np 13643–14148), JT (np 13927–13940), and KU (np 13928–14002) sets. Regions with significantly low likelihood in the African and KU sets also overlapped at the 5' end of *MTCYB*. The regions with low likelihood within rRNA genes were at different positions in different haplogroup clusters. From the tRNA genes, *MTTT* had a region with significantly low likelihood in the HV, KU and IWX sets, and the region also spanned the *MITH*, *MTTS2*, and *MTTL2* genes in the KU set. In contrast to nondegenerate and twofold degenerate sites, the variation at fourfold degenerate sites was explained well by a single phylogenetic topology and nucleotide substitutions process in each sequence set, with the exception of np 4850–4889 in the Asian set. The region in *MTND5* with the unusual pattern of polymorphism was not identified in this analysis, most likely because regions evolving at lower rates than the rest of the sequence do not tend to have a lower likelihood, as transition probabilities for a base staying the same never become particularly low (Grassly and Holmes 1997). There were not many regions which had a significantly low likelihood in the analysis of protein-coding sequences, and the regions did not generally span entire genes, suggesting that adaptive protein evolution has involved only certain regions of genes, if it has occurred at all.

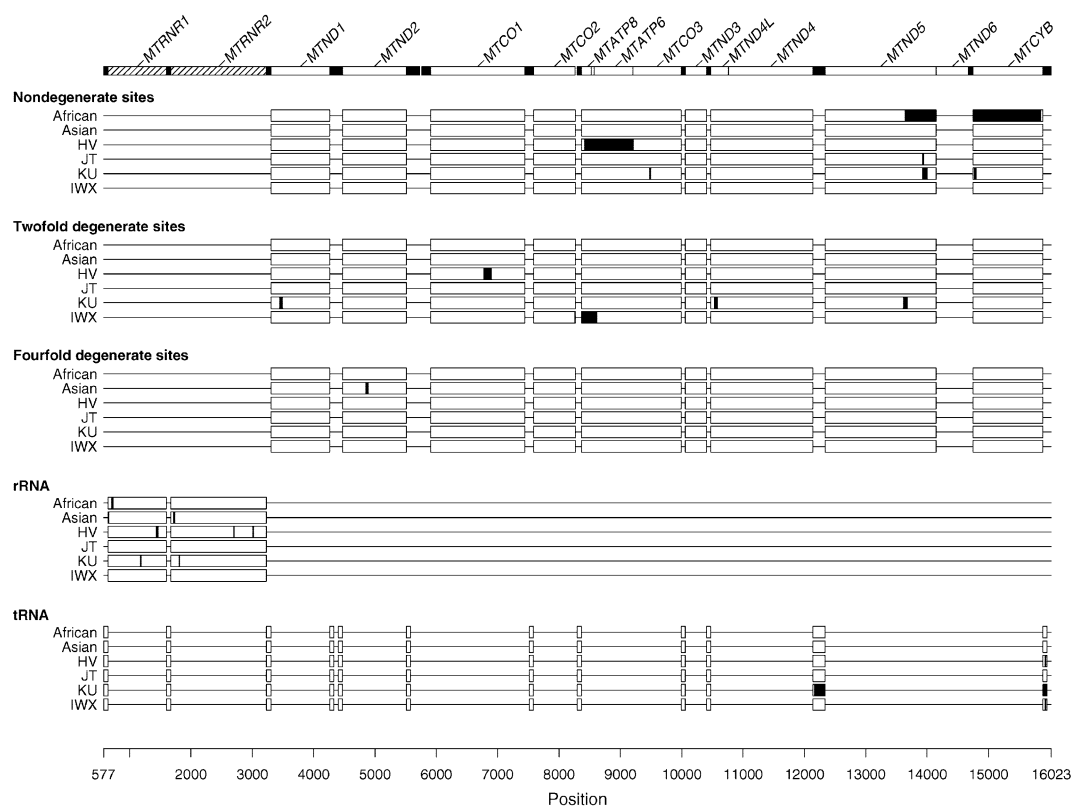


FIG. 6.—Maximum-likelihood sliding window analysis. African, Asian, and European HV, JT, KU, and IWX sequence data sets involving nondegenerate, twofold degenerate, and fourfold degenerate sites and rRNA and tRNA genes were analyzed using TREE-PUZZLE 5.0 and PLATO 2.11 to identify regions which do not fit with a single phylogenetic topology and nucleotide substitution process along the entire sequence. Genes encoded by the L-strand (*MTND6* and the tRNA genes *MTTQ*, *MTA*, *MTN*, *MTTC*, *MTTY*, *MTTS1*, *MTE*, and *MTTP*) were excluded in order to avoid pooling template and non-template DNA. The bar at the top illustrates the locations of the various genes (shaded, rRNA; solid, tRNA; open, protein-coding). Black in the remaining bars illustrates regions with significantly low likelihood in each set. Additional data on these regions are available as online Supplementary Material.

Regions with a low likelihood could also result from recombination (Grassly and Holmes 1997), but we found no homologous pairs of alleles within the identified regions which could have been consistent with ancient recombination events between the lineages.

### Conclusion

We analyzed an alignment of 837 human mtDNA sequences in terms of polymorphism indices and neutrality tests to identify lineage-specific local patterns of sequence variation within sequence sets defined by human mtDNA haplogroups. The general rank order of classes of sites was found to be consistent with earlier studies and with the assumption that nucleotide diversity indices reflect selection in within-genome comparisons, but there were also exceptions from these general rules in individual haplogroup clusters. Selection against mutations in tRNA genes, in particular, may have been different in different populations. The analysis of nonsynonymous/synonymous rate ratios in the 13 protein-coding genes revealed differences between the African, Asian, and, surprisingly, also between the European haplogroup clusters, suggesting that climate may not be the only explanation for such differences. The sliding window comparison of 1,000-bp segments of the coding region revealed a general corre-

lation between position in the genome and the diversity indices and neutrality tests, and greatly varying local patterns between haplogroup clusters. Interestingly, a segment of the *MTND5* gene was found to be almost invariable in haplogroup J, which has been associated with certain complex diseases, but not invariable in other haplogroups. Finally, the maximum-likelihood sliding window analysis also indicated that regions with the highest diversity differ between haplogroup clusters. Thus, several lines of evidence suggest that selective constraints against regions of the human mtDNA may be different in different lineages, and that lineage-specific interactions are a plausible explanation for this finding. The possibility of such interactions should be taken into account in evolutionary analyses and evolutionary models of mtDNA. In addition to such theoretical implications, the clinical consequences may also be significant, as the pathogenic potential of a mutation may depend markedly on the presence of other interacting mutations.

### Supplementary Material

Numerical data from the analysis of nucleotide diversity indices and neutrality tests by classes of sites, complete figures of the allele frequency distribution sliding

window analyses, and additional data on the regions with significantly low likelihood in the maximum-likelihood sliding window analysis are available as online Supplementary Material.

## Acknowledgments

This work was supported by grants from the Sigrid Juselius Foundation, the Maud Kuistila Memorial Foundation, and the Research Council for Health, Academy of Finland.

## Literature Cited

- Andrews, R. M., I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**:147.
- Ballard, J. W. O. 2000a. Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. *J. Mol. Evol.* **51**:48–63.
- . 2000b. Comparative genomics of mitochondrial DNA in *Drosophila simulans*. *J. Mol. Evol.* **51**:64–75.
- Bandelt, H. J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* **141**:743–753.
- Blier, P. U., F. Dufresne, and R. S. Burton. 2001. Natural selection and the evolution of mtDNA-encoded peptides: evidence for intergenomic co-adaptation. *Trends Genet.* **17**:400–406.
- Brown, M. D., E. Starikovskaya, O. Derbeneva, S. Hosseini, J. C. Allen, I. E. Mikhailovskaya, R. I. Sukernik, and D. C. Wallace. 2002. The role of mtDNA background in disease expression: a new primary LHON mutation associated with Western Eurasian haplogroup J. *Hum. Genet.* **110**:130–138.
- De Benedictis, G., G. Rose, G. Carrieri et al. (13 co-authors). 1999. Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans. *FASEB J.* **13**:1532–1536.
- Elson, J. L., R. M. Andrews, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. 2001. Analysis of European mtDNAs for recombination. *Am. J. Hum. Genet.* **68**:145–153.
- Finnilä, S., M. S. Lehtonen, and K. Majamaa. 2001. Phylogenetic network for European mtDNA. *Am. J. Hum. Genet.* **68**:1475–1484.
- Fu, Y. X., and W.-H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**:693–709.
- Fu, Y. X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**:915–925.
- Gerber, A. S., R. Loggins, S. Kumar, and T. E. Dowling. 2001. Does nonneutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? *Annu. Rev. Genet.* **35**:539–566.
- Grassly, N. C., and E. C. Holmes. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.* **14**:239–247.
- Hasegawa, M., H. Kishino, and T.-A. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- Herrnstadt, C., J. L. Elson, E. Fahy et al. (11 co-authors). 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.
- Thaka, R., and R. Gentleman. 1996. R: a language for data analysis and graphics. *J. Comp. Graph. Stat.* **5**:299–314.
- Ingman, M., H. Kaessmann, S. Pääbo, and U. Gyllensten. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**:708–713.
- Li, W.-H. 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, Mass.
- Maca-Meyer, N., A. M. González, J. M. Larruga, C. Flores, and V. M. Cabrera. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet.* **2**:13.
- Mathiesen, C., and C. Hägerhäll. 2002. Transmembrane topology of the NuoL, M and N subunits of NADH: quinone oxidoreductase and their homologues among membrane-bound hydrogenases and bona fide antiporters. *Biochim. Biophys. Acta* **1556**:121–132.
- Meinilä, M., S. Finnilä, and K. Majamaa. 2001. Evidence for mtDNA admixture between the Finns and the Saami. *Hum. Hered.* **52**:160–170.
- Mishmar, D., E. Ruiz-Pesini, P. Golik et al. (13 co-authors). 2003. Natural selection shaped regional mtDNA variation in humans. *Proc. Natl. Acad. Sci. USA* **100**:171–176.
- Möilänen, J. S., and K. Majamaa. 2003. Phylogenetic network and physicochemical properties of nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA. *Mol. Biol. Evol.* **20**:1195–1210.
- Nedbal, M. A., and J. J. Flynn. 1998. Do the combined effects of the asymmetric process of replication and DNA damage from oxygen radicals produce a mutation-rate signature in the mitochondrial genome? *Mol. Biol. Evol.* **15**:219–223.
- Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
- Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**:641–674.
- Niemi, A. K., A. Hervonen, M. Hurme, P. J. Karhunen, M. Jylhä, and K. Majamaa. 2003. Mitochondrial DNA polymorphisms associated with longevity in a Finnish population. *Hum. Genet.* **112**:29–33.
- Parsons, T. J., D. S. Muniec, K. Sullivan et al. (11 co-authors). 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.* **15**:363–368.
- Pesole, G., C. Gissi, A. De Chirico, and C. Saccone. 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. *J. Mol. Evol.* **48**:427–434.
- Przeworski, M., R. R. Hudson, and A. Di Rienzo. 2000. Adjusting the focus on human variation. *Trends Genet.* **16**:296–302.
- Rand, D. 2001. Mitochondrial genomics flies high. *Trends Ecol. Evol.* **16**:2–4.
- Reynier, P., I. Penisson-Besnier, C. Moreau, F. Savagner, B. Vielle, J. Emile, F. Dubas, and Y. Malthiery. 1999. mtDNA haplogroup J: a contributing factor of optic neuritis. *Eur. J. Hum. Genet.* **7**:404–406.
- Rose, G., G. Passarino, G. Carrieri, K. Altomare, V. Greco, S. Bertolini, M. Bonafe, C. Franceschi, and G. De Benedictis. 2001. Paradoxes in longevity: sequence analysis of mtDNA haplogroup J in centenarians. *Eur. J. Hum. Genet.* **9**:701–707.
- Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**:174–175.
- Schmidt, T. R., W. Wu, M. Goodman, and L. I. Grossman. 2001. Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome *c* oxidase. *Mol. Biol. Evol.* **18**:563–569.

- Simonsen, K. L., G. A. Churchill, and C. F. Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**:413–429.
- Stajich, J., D. Block, K. Boulez et al. (21 co-authors). 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**:1611–1618.
- Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**:964–969.
- Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**:437–460.
- . 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Torrioni, A., K. Huoponen, P. Francalacci, M. Petrozzi, L. Morelli, R. Scozzari, D. Obinu, M. L. Savontaus, and D. C. Wallace. 1996. Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**:1835–1850.
- Torrioni, A., M. Petrozzi, L. D'Urbano et al. (12 co-authors). 1997. Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. *Am. J. Hum. Genet.* **60**:1107–1121.
- Tourasse, N. J., and W.-H. Li. 2000. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* **17**:656–664.
- Wallace, D. C., M. D. Brown, and M. T. Lott. 1999. Mitochondrial DNA variation in human evolution and disease. *Gene* **238**:211–230.
- Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**:256–276.

Wolfgang Stephan, Associate Editor

Accepted July 29, 2003