

The role of selection in the evolution of human mitochondrial genomes

Toomas Kivisild^{*,1,4,5}, Peidong Shen^{†,4}, Dennis Wall^{‡3}, Bao Do[†], Raphael Sung[†], Karen Davis[†], Giuseppe Passarino[§], Peter A. Underhill^{*}, Curt Scharfe[†], Antonio Torroni^{**}, Rosaria Scozzari^{††}, David Modiano^{‡‡}, Alfredo Coppa^{§§}, Peter de Knijff^{***}, Marcus Feldman[‡], Luca L. Cavalli-Sforza^{*}, Peter J. Oefner^{†,2,4}

*Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.

†Stanford Genome Technology Center, Palo Alto, California 94304, USA.

‡Department of Biological Sciences, Stanford University, Stanford, California 94305, USA.

§Dipartimento di Biologia Cellulare, Università della Calabria, Rende, Italy.

**Dipartimento di Genetica e Microbiologia, Università di Pavia, Pavia, Italy.

††Dipartimento di Genetica e Biologia Molecolare, Università "La Sapienza", Rome, Italy.

‡‡Dipartimento di Scienze di Sanità Pubblica, Sezione di Parassitologia, Università "La Sapienza", Rome, Italy.

§§Dipartimento di Biologia Animale e dell'Uomo, Università "La Sapienza", Rome, Italy.

***Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.

Present addresses:

¹Department of Evolutionary Biology, Tartu University and Estonian Biocenter, 51010 Tartu, Estonia.

² Institute of Functional Genomics, University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany.

³ Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA.

⁴Both authors contributed equally to the manuscript.

Complete coding region sequence information for 277 individual samples have been submitted to GenBank under accession numbers DQ112686-DQ112962.

Running title: Evolution of human mitochondrial genomes

Key words: mitochondrial DNA, human population genetics, natural selection, molecular clock, evolution

⁵ Corresponding authors: TK, Estonian Biocenter, Tartu, Estonia, 51010, e-mail tkivisil@ebc.ee, and P.J.O., Institute of Functional Genomics, University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany, e-mail: oefner@genome.stanford.edu.

ABSTRACT

High mutation rate in mammalian mitochondrial DNA generates a highly divergent pool of alleles even within species that have dispersed and expanded in size recently. Phylogenetic analysis of 277 human mitochondrial genomes revealed a significant ($p < 0.01$) excess of rRNA and non-synonymous base substitutions among hotspots of recurrent mutation. Most hotspots involved transitions from guanine to adenine that, together with thymine to cytosine transitions, illustrate the asymmetric bias in codon usage at synonymous sites on the heavy-strand DNA. The mitochondrion-encoded tRNA^{Thr} varied significantly more than any other tRNA gene. Threonine and valine codons were involved in 259 of the 414 amino acid replacements observed. The ratio of non-synonymous changes from and to threonine and valine differed significantly ($P = 0.003$) between populations with neutral (22/58) and those with significantly negative Tajima's D values (70/76), independent of their geographic location. In contrast to a recent suggestion that the excess of non-silent mutations is characteristic to Arctic populations implying their role in cold adaptation, we demonstrate that the surplus of non-synonymous mutations is a general feature of the young branches of the phylogenetic tree, affecting also those that are found only in Africa. We introduce a new calibration method of the mutation rate of synonymous transitions to estimate the coalescent times of mtDNA haplogroups.

INTRODUCTION

Mitochondrial DNA (mtDNA) encodes for 13 proteins, 2 ribosomal genes, and 22 tRNAs that are essential in the energy production of the human cell. Variation in its

sequence has provided significant insights into the maternal history of anatomically modern humans (DENARO *et al.* 1981; GILES *et al.* 1980), complementing the paternal legacy of the Y-chromosome (UNDERHILL *et al.* 2000). Studies based on restriction fragment length polymorphism (RFLP) of the coding and direct sequencing of the non-coding control region have formed the basis of a hierarchical classification of distinct geographic and ethnic affinities (CHEN *et al.* 1995; FORSTER *et al.* 2001; MACAULAY *et al.* 1999; TORRONI *et al.* 1996; TORRONI *et al.* 1993; WATSON *et al.* 1997). Studies addressing sequence variation in the mtDNA coding region have suggested that natural selection has significantly shaped the course of human mtDNA evolution (CANN *et al.* 1984; ELSON *et al.* 2004; INGMAN and GYLLENSTEN 2001; MISHMAR *et al.* 2003; MOILANEN *et al.* 2003; MOILANEN and MAJAMAA 2003; NACHMAN *et al.* 1996; RUIZ-PESINI *et al.* 2004). These studies have disagreed, however, upon whether the distribution of specific human mtDNA clades or haplogroups is due to an adaptation to different climates or if their distribution is a function of random genetic drift assisted by purifying selection that eliminates non-synonymous changes. In an attempt to clarify this disagreement and to study the mode of natural selection on mtDNA variation in human populations, we provide here a phylogenetic analysis of a global sample of mtDNAs and investigate the position, chemical nature and geographic distribution of recurrent and frequent mutations in the coding region.

MATERIAL AND METHODS

DNA samples: The ascertainment set comprised 277 individuals from the five

continents, including 129 African (10 Biaka Pygmy, 15 Mbuti Pygmy, 2 Lisongo, 6 San, 2 Mandenka, 4 Ethiopian Jew, 9 Sudanese, 1 Eritrean, 1 Ghanan, 3 Herero, 1 Ovambo, 1 Pedi, 1 Sotho, 2 Tswana, 2 Zulu, 10 Fulbe, 10 Mossi, 10 Rimaibe, 1 Berta, 1 Tuareg, 37 Dominican), 43 Asian (1 Arab, 1 Kazak, 1 Druze, 4 Bedouin, 1 Sepharadim, 1 Yemenite Jew, 2 Pathan, 5 Sindhi, 2 Burushaski, 1 Baluchi, 1 Brahui, 2 Makran, 2 Hazara, 1 Tamil, 2 Cambodian, 1 Hmong, 1 Atayal, 1 Ami, 4 Han Chinese, 5 Japanese, 4 Korean), 76 European (5 Northern European, 12 Italian, 1 Greek, 2 Finn, 2 Ashkenazi, 1 Georgian, 17 Hungarian, 3 Icelander, 3 Czech, 1 Sardinian, 5 Basque, 1 Iberian, 23 Dutch), 13 Oceanian (4 New Guinean, 3 Melanesian, 6 Australian Aborigines), and 16 Native American (1 Auca, 1 Guarani, 5 Brazilian Indian, 3 Colombian Indian, 2 Mayan, 1 Piman, 1 Muskogee, 1 Navaho, 1 Quechua) genomic DNA samples (Supplemental Table 1). A subset of 103 sequences of these has been reported elsewhere (SHEN *et al.* 2005). DNA was extracted using QIAamp DNA Blood Kit (QIAGEN Inc., Valencia, CA). Immortalized cell lines have been established for all individuals with the exception of the 17 Hungarians, the 23 Dutch, the 37 Dominicans, the 10 Mossi, the 10 Rimaibe, and the 10 Fulbe.

PCR and DNA sequencing: The 41 primer pairs used for bidirectional sequencing of mtDNA nucleotides 435 to 16023, the PCR conditions and determined complete coding region sequence information for 277 individual samples are available at http://insertion.stanford.edu/primers_mitogenome.html. Amplicons were purified with Qiagen (Valencia, CA) QIAquick spin columns and sequenced with the Applied Biosystems (Foster City, CA) Dye Terminator Cycle Sequencing Kit and a model 3700

DNA sequencer.

Phylogenetic and statistical analyses: An unrooted tree from a median-joining network (BANDELT *et al.* 1999) was drawn and labeled following existing mtDNA haplogroup nomenclature (KIVISILD *et al.* 2004; KIVISILD *et al.* 2002; KONG *et al.* 2003; KONG *et al.* 2004; MACAULAY *et al.* 1999; SALAS *et al.* 2002; SHEN *et al.* 2005; TORRONI *et al.* 1996; TORRONI *et al.* 2001; YAO *et al.* 2002). The tree was rooted using nuclear inserts of mtDNA retrieved from human genomic sequence and the consensus sequence of the three chimpanzee mitochondrial genomes. The accession numbers, mtDNA positional range and identity measures of the genomic contigs containing the inserts that were used for rooting are as follows: NT_006713.14 (bps 341-2697; ID 94%); NT_009237.17 (bps 521-2976; ID 94%); NT_006316.15 (bps 2899-3050; ID 94%); NT_077913.3 (bps 3914-9756; ID 98%); and NT_034772.5 (bps 10269-15487; ID 94%). The GenBank accession numbers of the two *Pan troglodytes* and one *Pan paniscus* sequences that were used are D38113, X93335, and D38116, respectively. Haplogroup divergence estimates ρ and their error ranges were calculated as averages of the distances from the tips to the most recent common ancestor of the haplogroup (FORSTER *et al.* 1996; SAILLARD *et al.* 2000). Two separate measures of non-synonymous (N) to synonymous (S) substitution ratios were used: first, the M_N/M_S ratio estimates the number of mutational changes inferred from the phylogenetic tree (Figure 1), whereas the $d_N/(d_S+\text{constant})$ refers as in (MISHMAR *et al.* 2003) to the ratio of average pair-wise distances of N and S changes in the given sample. Statistical significance was determined from binomial or χ^2 probabilities. Disease implicated substitutions were excluded from these analyses. For

interspecies comparisons, mammalian mtDNA sequences were retrieved from the Mitochondriome web site

(http://bighost.area.ba.cnr.it/mitochondriome/Mt_chordata.htm).

Tests for positive selection

Seven primate taxa, namely *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla*, *Papio hamadryas*, *Hylobates lar*, *Pongo pygmaeus*, and *Macaca sylvanus* were chosen from Genbank (gi|17981852, gi|5835121, gi|5835149, gi|5835638, gi|5835820, gi|5835163, gi|14010693) and aligned using clustalW (THOMPSON *et al.* 1994), to test for the historic occurrence of positive, directional selection on the 13 coding regions of the primate mitochondrion using the program codeml of the PAML package (YANG 2002). In these tests, maximum likelihood ratios of non-synonymous to synonymous mutations (ω , omega) exceeding 1 are consistent with the hypothesis of positive selection, while values close to 1 indicate selective neutrality, and values converging on 0 suggest strong purifying selection. We conducted both lineage and site specific tests. For the lineage-specific tests, we used a model in which all lineages have the same ω (hereafter referred to as M0) and compared that with a model in which ω is estimated for each lineage (hereafter referred to as M1). To test for the action of selection among amino acid sites within a specific lineage, we compared a model that allows for heterogeneity in ω among sites, but not among lineages, with a model that allows for variation in ω along a predefined lineage (as in (YANG and NIELSEN 2002)). We assumed the following unrooted phylogeny (troglodytes, (((macaca, papio), hylobates), pongo), gorilla), troglodytes), human). However, results of our analyses were robust to minor fluctuations

in the tree.

RESULTS

The deepest splits of the phylogeny constructed from 277 mtDNA complete coding region sequences (Figure 1) were sustained by African mtDNAs which belonged to previously defined haplogroups L0-L5 (KIVISILD *et al.* 2004; MISHMAR *et al.* 2003; SALAS *et al.* 2002; SHEN *et al.* 2005; TORRONI *et al.* 2001). A number of new sub-clades were identified among these that are illustrated in Figure 1. Haplogroup sharing between distinct geographic regions was generally low. All European sequences could be assigned to clades N1, X, W, HV, TJ, and U (FINNILÄ *et al.* 2001; HERRNSTADT *et al.* 2002; MACAULAY *et al.* 1999; TORRONI *et al.* 1996). Asian, Amerindian, Oceanian, and Australian Aborigine sequences belonged to region specific haplogroups nested within macro-clades M and N (FRIEDLAENDER *et al.* 2005; KIVISILD *et al.* 2002; KONG *et al.* 2003; KONG *et al.* 2004; YAO *et al.* 2002). All Australian Aborigine M sequences (two from this study and one from Ingman *et al.* (INGMAN *et al.* 2000) share six mutations that define a new haplogroup M42. The majority of Australian N and R sequences (INGMAN and GYLLENSTEN 2003) belong to clades S and P defined by transitions at np 8404 and 15607, respectively (FORSTER *et al.* 2001; FRIEDLAENDER *et al.* 2005).

The most parsimonious root of the mtDNA tree using nuclear inserts of mtDNA and the chimpanzee consensus sequence as outgroups, appeared between haplogroup L0 and the rest of the phylogeny (Figure 1). Extensive interspecies homoplasy and mutational saturation was highlighted by the fact that for more than one third (417/1292)

of the variable sites, regardless of their phylogenetic position on the tree, the derived allele among humans corresponded to the chimpanzee allele. In agreement with non-coding region information (AQUADRO and GREENBERG 1983), a high ratio (21.5 on average, 34.8 in synonymous positions) of transitions to transversions was observed in the coding region (577-16023).

Interspecies calibration of the molecular clock over the complete mtDNA sequence (INGMAN *et al.* 2000; MISHMAR *et al.* 2003) is problematic because of saturation of transitions at silent positions and the effect of selection on the fixation rate of amino acid replacement mutations (HO *et al.* 2005). Assuming 6 million years for the human-chimp species split (GOODMAN *et al.* 1998) and 6.5 million years for the MRCA of their mtDNA lineages (MISHMAR *et al.* 2003) we estimated the average transversion rate at synonymous and rRNA positions as 2.1×10^{-9} and 4.1×10^{-10} per year per position, respectively. Using the observed relative rates of different substitution types in humans (Table 1) the average transition rate at 4,212 synonymous positions is 3.5×10^{-8} (S.D. 0.1×10^{-8}) per year per position. Over all genes in mtDNA this would be equivalent to accumulation of one synonymous transition per 6,764 (S.D. 140) years on average. The coalescent date of the human mitochondrial DNA tree using this rate is 160,000 (S.D. 22,000) years. This coalescent date is broadly consistent with the dates of the *Homo sapiens* fossils recognized so far from Ethiopia (CLARK *et al.* 2003; MCDUGALL *et al.* 2005; WHITE *et al.* 2003). The most recent common ancestor of all the Eurasian, American, Australian, Papua New Guinean and African lineages in clade L3 dates to $65,000 \pm 8,000$ years while the average coalescent time of the three basic non-African

founding haplogroups M, N, and R is 45,000 years. These estimates, bracketing the time period for the recent out-of-Africa migration (STRINGER and ANDREWS 1988), are younger than those based on calibrations involving all coding region sites (INGMAN *et al.* 2000; MISHMAR *et al.* 2003) but are still in agreement with the earliest archaeological signs of anatomically modern humans outside Africa (MELLARS 2004). The differences with the date estimates of previous studies are most likely due to the overrepresentation of possibly slightly deleterious non-synonymous mutations in the younger branches of the tree (ELSON *et al.* 2004) that introduces a bias to the coalescent approach if all the sites of the coding region are used.

Of the 1,788 mutations depicted in the tree, 1,758 occurred at 1,292 variable sites in the coding region, between nps 577-16023. Consistent with previous reports (ELSON *et al.* 2004; MISHMAR *et al.* 2003; MOILANEN and MAJAMAA 2003), there was a significant excess of synonymous mutations in all genes coded by mtDNA, especially among those positions that defined the deeper branches of the tree (Tables 2 and 3). In contrast to Ruiz-Pesini *et al.* (RUIZ-PESINI *et al.* 2004), we did not observe any significant regional (climatic) differences in the rate of non-synonymous changes for mtDNA haplogroups. This discrepancy likely results from the fact that Ruiz-Pesini *et al.* compared region-specific haplogroups of different diversity levels - e.g. the "old" paragroup L in Africans versus "young" Arctic haplogroups (Table 4). Populations of Asian, European, and West African origin showed significantly negative Tajima's D values (Table 3), consistent with either selection, population growth, and/or population subdivision (RAY *et al.* 2003). That population substructure accounts at least for part of the deviation from neutrality is

obvious from the observation that it decreases upon partitioning of West Africans into a sample from Burkina Faso and one from the Dominican Republican.

Significant ($p < 0.05$) mutational bias towards specific (NNG to NNA and NNU to NNC) codon usage was observed in 27 out of 32 pairs of codons that differed by a transition in the third codon position (Table 5). This relative preference of G to A and T to C mutations (per existing nucleotide pool in the light strand) extends over the non-silent positions and is characteristic of the non-coding D-loop region (AQUADRO and GREENBERG 1983; MALYARCHUK and ROGOZIN 2004). The general strand bias, known, however, to be reversed in some Metazoan genera, can be related to asymmetric mutational constraints involving deaminations of A and C nucleotides during the replication and/or transcription processes (HASSANIN *et al.* 2005). Importantly, the ND6 gene, encoded by the heavy strand showed the opposite mutational bias suggesting that the differences of codon usage in human mtDNA might be primarily a function of strand asymmetry rather than differences in the tRNA pools as generally expected (TANAKA and OZAWA 1994). Notably, 16/18 nucleotide positions (Table 6) that had undergone five or more recurrent changes involved the transition of guanine to adenine in the light strand.

Of the 1,292 variable sites, 288 (22.2%) had mutated recurrently. Unexpectedly, the hotspots that had mutated ≥ 5 times were predominantly within mitochondrial rRNA ($p < 5 \times 10^{-15}$) and showed a significantly higher ratio of non-synonymous to silent mutations (90:32 hits, respectively) than polymorphic sites with lower recurrence (608:1004) (Table 6). Finally, these hotspots of mutational activity included positions

where the human derived allele predominates in the mammalian consensus sequences (e.g. nps 709, 3010, 10398, and 13928), implying the effect of site-specific positive selection. Among the six non-synonymous substitutions that have recurred ≥ 4 times, five involved threonine ($p < 6.1 \times 10^{-7}$). Overall, threonine and valine codons were involved in 259 of the 414 amino acid replacements observed on the tree.

Lineage-specific tests failed to detect significant positive selection along any unique lineage in the 7 taxon phylogeny of primates. A model fixing a single ratio of ω to all lineages (M0) could not be rejected in favor of a model of different ω 's on specified lineages (M1). The ω estimated across all lineages in the phylogeny was 0.35. A test of the previous model against a model enforcing neutral selection, where ω is expected to be equal to 1, showed that these data do not deviate significantly from neutrality (M0 was rejected in favor of model where $\omega=1$; $p \approx 0$, $df=1$). Further tests for lineage specific variation in ω , including a model that assigned a different omega to the human lineage from the remaining primates, fit the data worse than M1. However, site specific model testing revealed significant positive selection across regions of the primate mitochondrion. A model enforcing a single ω ratio on all codon sites was rejected in favor of a model allowing for 3 ratios across sites with 3 site classes ($p \approx 0$, $df = 5$). The 3-ratio model identified 16 codon sites to be under significant (posterior probabilities > 0.95 ; $dN/dS = 2.02$) positive selection (Table 7). Among these, four codon sites appeared to be among the non-synonymous sites with recurrent mutation (particularly #114 in ND3 gene, np 10398 with 7 recurrences) in human-human comparisons (Table 6).

A majority of the mitochondrial disease related mutations have been detected in the tRNA genes, and they mainly affect the secondary structure of the molecule ((MCFARLAND *et al.* 2004) and references therein). Our global survey of natural variation in the tRNA genes showed that there was a seven-fold excess of tRNA^{Thr} mutations (N=28) over other tRNA genes ($p < 10^{-19}$). This finding would suggest, at first glance, that this gene might have become nonfunctional in mitochondrion and that its encoded tRNA needs to be imported from the nucleus. Evidence suggesting nuclear tRNA import into mitochondria in marsupials has been obtained previously (DÖRNER *et al.* 2001). However, plotting the observed mutations in the tRNA^{Thr} gene against the mammalian consensus sequences (HELM *et al.* 2000) showed that none of the mutations we have observed in 277 humans fell within the 100% conserved regions of the tRNA (Figure 2). Most pathological mutations affecting tRNAs cluster to highly conserved regions (MCFARLAND *et al.* 2004), as illustrated in the case of tRNA^{Leu} in Figure 2. Four private mutations changed the nucleotide that is more than 90% conserved in mammalian tRNA^{Thr}, while most of the frequent and recurrent mutations in the data set affected the minor fraction of the sites that are not highly conserved in mammalian species. This argues against the proposition that human mitochondrial tRNA^{Thr} has lost function. A large fraction (12/28) of the mutations affecting tRNA^{Thr} occurred at three positions, two of which have a different allele in consensus human as compared to the 31 mammalian species analyzed by Helm et al. (HELM *et al.* 2000). Similarly, a mutational hotspot at position 5821 in tRNA^{Cys} showed a majority allele in human different from that found in consensus mammalian (Figure 2). Surprisingly, in the latter tRNA we observed two parallel mutations at position 5814 that have been previously reported pathogenic. Yet,

because this position is not highly conserved in other mammalian species, its pathological role has to be questioned. No other tRNA site that has been confirmed as associated with mitochondrial disease was found to be variable in our data set. The only mutational hotspot (12172) affecting the human allele that matches the allele conserved in >90% of mammalian tRNA-s was found in tRNA^{His}.

A comparison of the amino acid substitutions in the mtDNA encoded proteins in humans, primates, carnivores, and artiodactyls, revealed that substitutions between threonine and alanine are significantly over-represented in humans while changes between methionine and leucine are most common in other mammalian species (Table 8). The direction of threonine and valine substitution with other amino acids was significantly different between populations with neutral and significantly negative Tajima's D values, respectively (Table 3), and between haplogroups: in H1 sequences sampled broadly from Europe and Near East, 7 of 11 non-synonymous mutations resulted in the replacement of threonine and valine with alanine and isoleucine, while only three mutations resulted in a change towards threonine or valine (Figure 1). In contrast to this pattern, in haplogroup V sequences from Finland (FINNILÄ *et al.* 2001), where populations continued to rely largely on hunting and fishing for subsistence even after the first contacts with farmers, 6 of 7 replacement polymorphisms resulted in a change to threonine and valine, and none in the replacement of the latter two amino acids ($p < 0.01$). Similarly, L3 sequences of West African origin showed significantly lower ($p < 0.001$) ratio of gains to losses of threonine and valine residues (13/14) than haplogroup L0 sequences from East and South Africa (22/2; Figure 1). East Asian sequences showed an

increase of valine codons (9 mutations to and 2 from valine codons), but also a significant ($p < 0.01$) decrease of threonine, with 11 mutations from and 5 to threonine codons, while Native Americans had 1 mutation from and 7 to threonine, respectively. Over all haplogroups and genes, the direction of amino acid change was significantly ($p < 0.02$) biased toward replacement of isoleucine and methionine to valine, even when considering the transitional preferences observed in mitochondrial D-loop (TAMURA and NEI 1993). The strand-specific mutational biases are unlikely to explain this pattern because of the observed excess of mutations involving valine codons (8/16 in our data set and 13/16 in the list of ND6 polymorphic sites in MITOMAP) in the ND6 gene that is encoded by the opposite strand.

DISCUSSION

In phylogenetic analysis of human mitochondrial DNA coding region sequences, two different spheres of character evolution can be distinguished (HO *et al.* 2005; PENNY 2005). Firstly, within our species, at the population level, relatively low level of parallel mutations – as compared to the mtDNA control region based phylogenies – enables the reconstruction of the unrooted tree from individual sequences without significant ambiguity. This tree is determined by a substantial fraction of amino acid replacement mutations whose proportion to synonymous substitutions increases from the average of 0.37 in “older” clades to 0.62 in the “younger” ones. In the second sphere, high level of homoplasmy with chimpanzee, affecting at least one third of the variable sites in humans, complicates detailed phylogenetic analyses at the interspecies level. Approximately 930 synonymous mutations that can be observed between a human and a chimpanzee mtDNA

represent only the visible component of variation between the species while the effective ratio of non-synonymous to silent mutations is expectedly significantly less than the observed value of 0.2 due to the hidden load of synonymous mutations. These differences between the two spheres imply that even among the substitutions that define the deepest branches of the human mtDNA tree there is a significant excess of non-synonymous mutations that have not yet been eliminated by purifying selection – assuming, of course, that they are generally deleterious, after all.

More than half of the amino acid replacements observed in the human mtDNA tree involved threonine and valine codons. Adaptive correlation with the elevated mutability in the mitochondrion-encoded tRNA^{Thr} could be, in principle, considered as one explanation for the excess of mutations involving threonine codons. However, none of the highly conserved sites in the tRNA^{Thr} gene was found to be different in humans from the consensus mammalian and, instead, the excessive variability in this gene could be ascribed largely to the presence of three hotspot positions. Furthermore, no such general molecular phenomenon or the characteristic G to A and T to C mutational bias on the light strand of mtDNA would explain the pattern of differences of amino acid replacement directions that were observed between human populations.

One factor that could, theoretically at least, explain the different amino acid replacement patterns observed between populations and between humans and other mammals is diet. Threonine and valine, essential amino acids that must be taken in the diet, are abundant in meats, fish, peanuts, lentils, and cottage cheese, but deficient in most

grains. Alternatively, or in combination with dietary restriction, other constraints of selection on slightly deleterious positions during the phases of population expansion and contraction may be involved. Because of the specific compositional bias in mtDNA induced by characteristic mutational preferences different from those observed in the nuclear genome, additional inter- and intraspecies comparisons of mtDNA encoded amino acid replacement patterns should be examined to gain deeper insights into the non-synonymous character evolution in metazoan mitochondria, particularly including taxa with shifted strand symmetry (HASSANIN *et al.* 2005).

Tests of neutrality based on the comparisons of the ratio of non-synonymous and synonymous mutations across all sites can detect only major effects of purifying (K_N/K_S approaches 0) or directional selection (K_N/K_S is significantly >1), which affect simultaneously a large number of codon positions. Consistent with previous studies (CANN *et al.* 1984; ELSON *et al.* 2004; INGMAN and GYLLENSTEN 2001; MISHMAR *et al.* 2003; MOILANEN *et al.* 2003; MOILANEN and MAJAMAA 2003; NACHMAN *et al.* 1996; RUIZ-PESINI *et al.* 2004) human mtDNA encoded proteins did not provide evidence of directional selection. However, several hotspots of mutational activity included non-silent substitutions susceptible to site-specific positive selection. Comparing the mtDNA protein encoding genes from several primates (*Macaca*, *Papio*, *Hylobates*, *Pongo*, *Gorilla*, and *Pan*) with human, we discovered significant positive selection in several regions, generally non-matching, however, with the codons displaying high K_N/K_S ratio in human-human comparisons. This difference might be explained by the dynamic polarity of the amino acid replacements at the intra- and interspecies levels whereby the

constraint of selection is determined in each lineage by the ancestral state of each codon position.

In conclusion, we have provided new evidence for non-random processes affecting the evolution of the human mtDNA-encoded proteins. The potential role of selection in affecting fixation probabilities at different non-silent positions undermines the appropriateness of using the average mitochondrial clock over all sites in dating events in human population history. Despite the evidence of departures from neutrality and high levels of homoplasy at the interspecies level, the phylogenetic approach for analyzing mtDNA sequence data at the intraspecies level remains viable because the reconstruction of the basic branches is robust and the excess of non-synonymous substitutions affects mainly the terminal branches of the tree.

ACKNOWLEDGEMENTS

We thank Richard Villems for useful comments. This work was supported by NIH grants GM28428, GM63883, and GM55273, EC DG Research grant ICA1CT20070006, Progetto CNR- MIUR Genomica Funzionale-Legge 449/97, and Telethon-Italy E.0890. We thank Rita Horvath for providing DNA samples.

LITERATURE CITED

- ACHILLI, A., C. RENGO, C. MAGRI, V. BATTAGLIA, A. OLIVIERI *et al.*, 2004 The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am J Hum Genet* **75**: 910-918.
- ANDREWS, R. M., I. KUBACKA, P. F. CHINNERY, R. N. LIGHTOWLERS, D. M. TURNBULL *et al.*, 1999 Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**: 147.
- AQUADRO, C. F., and B. D. GREENBERG, 1983 Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. *Genetics* **103**: 287-312.
- BANDELT, H. J., J. ALVES-SILVA, P. E. GUIMARAES, M. S. SANTOS, A. BREHM *et al.*, 2001 Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann Hum Genet* **65**: 549-563.
- BANDELT, H.-J., P. FORSTER and A. RÖHL, 1999 Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37-48.
- CANN, R. L., W. M. BROWN and A. C. WILSON, 1984 Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. *Genetics* **106**: 479-499.
- CHEN, Y. S., A. TORRONI, L. EXCOFFIER, A. S. SANTACHIARA-BENERECETTI and D. C. WALLACE, 1995 Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet* **57**: 133-149.
- CLARK, J. D., Y. BEYENE, G. WOLDEGABRIEL, W. K. HART, P. R. RENNE *et al.*, 2003 Stratigraphic, chronological and behavioural contexts of Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* **423**: 747-752.
- DENARO, M., H. BLANC, M. J. JOHNSON, K. H. CHEN, E. WILMSEN *et al.*, 1981 Ethnic variation in *Hpa* I endonuclease cleavage patterns of human mitochondrial DNA. *Proc Natl Acad Sci U S A* **78**: 5768-5772.
- DÖRNER, M., M. ALTMANN, S. PÄÄBO and M. MORL, 2001 Evidence for import of a lysyl-tRNA into marsupial mitochondria. *Mol Biol Cell* **12**: 2688-2698.
- ELSON, J. L., D. M. TURNBULL and N. HOWELL, 2004 Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. *Am J Hum Genet* **74**: 229-238.
- FINNILÄ, S., M. S. LEHTONEN and K. MAJAMAA, 2001 Phylogenetic network for European mtDNA. *Am J Hum Genet* **68**: 1475-1484.
- FORSTER, P., R. HARDING, A. TORRONI and H.-J. BANDELT, 1996 Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet* **59**: 935-945.
- FORSTER, P., A. TORRONI, C. RENFREW and A. RÖHL, 2001 Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol Biol Evol* **18**: 1864-1881.
- FRIEDLAENDER, J., T. SCHURR, F. GENTZ, G. KOKI, F. FRIEDLAENDER *et al.*, 2005 Expanding southwest Pacific mitochondrial haplogroups P and Q. *Mol Biol Evol* **22**: 1506-1517.
- GILES, R. E., H. BLANC, H. M. CANN and D. C. WALLACE, 1980 Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A* **77**: 6715-6719.

- GOODMAN, M., C. A. PORTER, J. CZELUSNIAK, S. L. PAGE, H. SCHNEIDER *et al.*, 1998
Toward a phylogenetic classification of Primates based on DNA evidence
complemented by fossil evidence. *Mol Phylogenet Evol* **9**: 585-598.
- HASSANIN, A., N. LEGER and J. DEUTSCH, 2005 Evidence for multiple reversals of
asymmetric mutational constraints during the evolution of the mitochondrial
genome of metazoa, and consequences for phylogenetic inferences. *Syst Biol* **54**:
277-298.
- HELM, M., H. BRULE, D. FRIEDE, R. GIEGE, D. PUTZ *et al.*, 2000 Search for characteristic
structural features of mammalian mitochondrial tRNAs. *RNA* **6**: 1356-1379.
- HERRNSTADT, C., J. L. ELSON, E. FAHY, G. PRESTON, D. M. TURNBULL *et al.*, 2002
Reduced-median-network analysis of complete mitochondrial DNA coding-region
sequences for the major African, Asian, and European haplogroups. *Am J Hum
Genet* **70**: 1152-1171.
- HO, S. Y., M. J. PHILLIPS, A. COOPER and A. J. DRUMMOND, 2005 Time Dependency of
Molecular Rate Estimates and Systematic Overestimation of Recent Divergence
Times. *Mol Biol Evol* **22**: 1561-1568.
- INGMAN, M., and U. GYLLENSTEN, 2001 Analysis of the complete human mtDNA
genome: methodology and inferences for human evolution. *J Hered* **92**: 454-461.
- INGMAN, M., and U. GYLLENSTEN, 2003 Mitochondrial genome variation and
evolutionary history of Australian and New Guinean aborigines. *Genome Res* **13**:
1600-1606.
- INGMAN, M., H. KAESSMANN, S. PÄÄBO and U. GYLLENSTEN, 2000 Mitochondrial
genome variation and the origin of modern humans. *Nature* **408**: 708-713.
- KIVISILD, T., M. REIDLA, E. METSPALU, A. ROSA, A. BREHM *et al.*, 2004 Ethiopian
mitochondrial DNA heritage: tracking gene flow across and around the gate of
tears. *Am J Hum Genet* **75**: 752-770.
- KIVISILD, T., H.-V. TOLK, J. PARIK, Y. WANG, S. S. PAPIHA *et al.*, 2002 The emerging
limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* **19**: 1737-1751
(erratum 1720:1162).
- KONG, Q.-P., Y.-G. YAO, C. SUN, H.-J. BANDELT, C.-L. ZHU *et al.*, 2003 Phylogeny of
East Asian mitochondrial DNA lineages inferred from complete sequences. *Am J
Hum Genet* **73**: 671-676.
- KONG, Q. P., Y. G. YAO, C. SUN, C. L. ZHU, L. ZHONG *et al.*, 2004 Phylogeographic
analysis of mitochondrial DNA haplogroup F2 in China reveals T12338C in the
initiation codon of the ND5 gene not to be pathogenic. *J Hum Genet* **49**: 414-423.
- MACAULAY, V. A., M. B. RICHARDS, E. HICKEY, E. VEGA, F. CRUCIANI *et al.*, 1999 The
emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences
and RFLPs. *Am J Hum Genet* **64**: 232-249.
- MALYARCHUK, B. A., and I. B. ROGOZIN, 2004 Mutagenesis by transient misalignment in
the human mitochondrial DNA control region. *Ann Hum Genet* **68**: 324-339.
- MCDUGALL, I., F. H. BROWN and J. G. FLEAGLE, 2005 Stratigraphic placement and age
of modern humans from Kibish, Ethiopia. *Nature* **433**: 733-736.
- McFARLAND, R., J. L. ELSON, R. W. TAYLOR, N. HOWELL and D. M. TURNBULL, 2004
Assigning pathogenicity to mitochondrial tRNA mutations: when "definitely
maybe" is not good enough. *Trends Genet* **20**: 591-596.

- MELLARS, P., 2004 Neanderthals and the modern human colonization of Europe. *Nature* **432**: 461-465.
- MISHMAR, D., E. RUIZ-PESINI, P. GOLIK, V. MACAULAY, A. G. CLARK *et al.*, 2003 Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* **100**: 171-176.
- MOILANEN, J. S., S. FINNILA and K. MAJAMAA, 2003 Lineage-specific selection in human mtDNA: lack of polymorphisms in a segment of MTND5 gene in haplogroup J. *Mol Biol Evol* **20**: 2132-2142.
- MOILANEN, 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.
- NACHMAN, M. W., W. M. BROWN, M. STONEKING and C. F. AQUADRO, 1996 Nonneutral mitochondrial DNA variation in humans and chimpanzees. *Genetics* **142**: 953-963.
- PALANICHAMY, M., C. SUN, S. AGRAWAL, H.-J. BANDELT, Q.-P. KONG *et al.*, 2004 Phylogeny of mtDNA macrohaplogroup N in India based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet* **75**: 966-978.
- PENNY, D., 2005 Evolutionary biology: relativity for molecular clocks. *Nature* **436**: 183-184.
- QUINTANA-MURCI, L., R. CHAIX, S. WELLS, D. BEHAR, H. SAYAR *et al.*, 2004 Where West meets East: The complex mtDNA landscape of the Southwest and Central Asian corridor. *Am J Hum Genet* **74**: 827-845.
- RAY, N., M. CURRAT and L. EXCOFFIER, 2003 Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol* **20**: 76-86.
- RUIZ-PESINI, E., D. MISHMAR, M. BRANDON, V. PROCACCIO and D. C. WALLACE, 2004 Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**: 223-226.
- SAILLARD, J., P. FORSTER, N. LYNNERUP, H.-J. BANDELT and S. NURBY, 2000 mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* **67**: 718-726.
- SALAS, A., M. RICHARDS, T. DE LA FE, M. V. LAREU, B. SOBRINO *et al.*, 2002 The making of the African mtDNA landscape. *Am J Hum Genet* **71**: 1082-1111.
- SHEN, P., A. E. HIRSH, T. KIVISILD, B. DO, S. SONG *et al.*, 2005 Population genetic implications from 103 pairs of globally representative Y-chromosome and mitochondrial DNA sequences. *Am J Hum Genet* **in press**.
- SHEN, P., T. LAVI, T. KIVISILD, V. CHOU, D. SENGUN *et al.*, 2004 Reconstruction of patri- and matri-lineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Hum Mut* **24**: 248-260.
- STRINGER, C. B., and P. ANDREWS, 1988 Genetic and fossil evidence for the origin of modern humans. *Science* **239**: 1263-1268.
- TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**: 512-526.

- TANAKA, M., V. M. CABRERA, A. M. GONZALEZ, J. M. LARRUGA, T. TAKEYASU *et al.*, 2004 Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res* **14**: 1832–1850.
- TANAKA, M., and T. OZAWA, 1994 Strand asymmetry in human mitochondrial DNA mutations. *Genomics* **22**: 327-335.
- THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673-4680.
- TORRONI, A., K. HUOPONEN, P. FRANCALACCI, M. PETROZZI, L. MORELLI *et al.*, 1996 Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**: 1835-1850.
- TORRONI, A., C. RENGO, V. GUIDA, F. CRUCIANI, D. SELBITTO *et al.*, 2001 Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* **69**: 1348-1356.
- TORRONI, A., T. G. SCHURR, M. F. CABELL, M. D. BROWN, J. V. NEEL *et al.*, 1993 Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet* **53**: 563-590.
- UNDERHILL, P. A., P. SHEN, A. A. LIN, L. JIN, G. PASSARINO *et al.*, 2000 Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**: 358-361.
- WATSON, E., P. FORSTER, M. RICHARDS and H. J. BANDELT, 1997 Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* **61**: 691-704.
- WHITE, T. D., B. ASFAW, D. DEGUSTA, H. GILBERT, G. D. RICHARDS *et al.*, 2003 Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* **423**: 742-747.
- YANG, Z., 2002 Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics* **162**: 1811-1823.
- YANG, Z., and R. NIELSEN, 2002 Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**: 908-917.
- YAO, Y.-G., Q.-P. KONG, H.-J. BANDELT, T. KIVISILD and Y.-P. ZHANG, 2002 Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* **70**: 635-651.

Figure legends

Figure 1. Maximum parsimony tree of 277 human mtDNA coding region sequences. Locations of branch-defining mutations are listed relative to the revised reference sequence (ANDREWS *et al.* 1999). Transversions are specified in capital letters, del indicates deletion, + insertion. Amino acid replacements are specified in parentheses with threonine and valine affecting changes highlighted in red and blue, respectively. ~t indicates change in tRNA, ~r in rRNA gene. Haplogroup labels follow existing classification (ACHILLI *et al.* 2004; BANDELT *et al.* 2001; CHEN *et al.* 1995; FORSTER *et al.* 2001; FRIEDLAENDER *et al.* 2005; KIVISILD *et al.* 2004; KIVISILD *et al.* 2002; KONG *et al.* 2003; MACAULAY *et al.* 1999; PALANICHAMY *et al.* 2004; SHEN *et al.* 2004; TANAKA *et al.* 2004; TORRONI *et al.* 1996; TORRONI *et al.* 2001; TORRONI *et al.* 1993; WATSON *et al.* 1997). Substitutions at nps 709, 3693, and 4715 in haplogroups B4, L2, and Z, respectively, are reconstructed less parsimoniously, yet in accordance with the phylogenetic relations suggested by additional data (FINNILÄ *et al.* 2001; KIVISILD *et al.* 2002; KONG *et al.* 2003; TORRONI *et al.* 2001). Coalescent estimates of haplogroups, shown in thousands of years in italics beside clade labels, are based on the average number of synonymous transitions to the root of the clade. The tree was rooted using as outgroup the nuclear inserts of mtDNA and the majority consensus of complete sequences of two *Pan troglodytes* and one *Pan paniscus* (for details see Material and Methods). The numbers in triangles indicate the additional mutational steps required in comparison to the optimally rooted tree using only the chimpanzee outgroup with the red font indicating the root that was rejected ($p < 0.01$) under the assumption of equal

evolutionary rate over branches. The positions at which the derived character states in humans and the chimpanzee consensus match are shown above the triangles. Disease implicated substitutions confirmed by two independent studies (www.mitomap.org) are indicated by #. Ad, Af, Am, As, Eu and Oc correspond to African descent, African, Amerindian, Asian, European and Oceanian, respectively. The continent of origin is distinguished for each individual sample by box color and sub-continental affiliation with letters below the boxes where N, E, W, S, C, SW refer to north, east, west, south, central, and southwest parts of the continents, respectively; BP and MP refer to Biaka and Mbuti Pygmies, respectively, A – Australian, P – Papuan, M – Melanesian, H – Hungarian LHON patients, G - Georgian. The two African mtDNAs (L0a2 and L1b1) found in West Asia (Pakistan) are likely due to recent admixture as indicated by their low genetic distance to related African samples presented in the tree and by the fact that they were detected in a Makrani and a Sindhi. According to Quintana-Murci *et al.* (QUINTANA-MURCI *et al.* 2004) the Makrani harbor an extremely high frequency (39%) of African haplogroups L3d, L3b, L2a, and L1a, most likely as a result of the forced migration of slave women from Africa that began in the 7th century and increased considerably during the Omani Empire. European gene flow, on the other hand, may account for the presence of European K and X sequences in one Aboriginal Australian and one Native American (Muskogee), respectively.

Figure 2. Natural and pathological variation in human mitochondrial species of tRNAs.

Mutations observed in the current study are shown in yellow, those with a status of confirmed pathological variation (www.mitomap.org) in red triangles.

Nucleotide positions that are 100% and 90% conserved in mammalian species (HELM *et al.* 2000), are highlighted in green and blue rectangles, respectively.

Nucleotide positions at which the human and mammalian consenses differ from each other are shown in red font colour.

Table 1. Distribution of mtDNA mutations by recurrence.

	non-syn	rRNA	tRNA	syn
Length in base pairs	8812	2513	1486	4212 ^a
Number of observed substitutions (per site)	413 (0.047)	173 (0.069)	110 (0.074)	1037 (0.246)
Transition/transversion ratio	12.4	23.7	12.8	34.8 ^b
Invariable sites	8506	2404	1409	3427
Sites with single hit	241	80	57	617
Sites with 1 recurrent hit	47	21	12	111
Sites with 2 recurrent hits	9	2	4	40
Sites with 3 recurrent hits	3	0	3	12
Sites with ≥ 4 recurrent hits	6	6	1	5
No of variable sites (proportion)	306 (0.035)	109 (0.043)	77 (0.052)	785 (0.307)

^a – including 2039 sites that are allowed to carry synonymous transversions

^b – effectively, Ts/Tv=16.8, when taking into account the number of sites that are allowed to vary.

Table 2. Distribution of mutations as a function of derived-allele frequency in 277 mtDNAs.

frequency	non-syn	syn	non-syn/syn	tRNA	rRNA	rRNA/syn
<1%	221	514	0.43	49	63	0.12
1-5%	67	197	0.34	24	28	0.14
5-10%	14	39	0.38	1	8	0.21
>10%	5	32	0.16*	3	10	0.31*

* p<0.05

Table 3. Ratios of numbers and rates of non-synonymous over synonymous sites, nucleotide diversity, Tajima's D, Fu and Li's D and F, and number of changes from and to valine and threonine for the 13 protein-coding mitochondrial genes according to continental affiliation.

Continent	Africa	S Africa	W Africa	B.F. ^a	D.R. ^b	Pygmy	E Africa	Non-African	SW Asia	Oceania	E Asia	America	S Europe	N Europe	Global	
No. of samples	129	16	70	33	37	25	15	148	23	12	18	15	20	38	277	
<i>ND1</i>																
M_N/M_S	13/46	3/14	11/26	5/15	8/20	1/9	1/11	15/39	3/8	2/6	2/9	4/4	2/5	3/8	19/74	
$d_N/(d_S+const.)$	0.058	0.052	0.073	0.073	0.071	0.037	0.014	0.116	0.102	0.080	0.028	0.193	0.089	0.119	0.074	
<i>ND2</i>																
M_N/M_S	17/42	2/10	13/25	6/12	11/20	3/10	3/13	19/53	8/10	4/5	5/16	3/6	1/8	4/16	30/84	
$d_N/(d_S+const.)$	0.129	0.079	0.118	0.086	0.140	0.108	0.064	0.138 ^c	0.196	0.239	0.107	0.079	0.056	0.121	0.136	
<i>COI</i>																
M_N/M_S	11/58	3/14	9/37	4/15	8/33	2/19	3/15	10/60	2/17	0/4	0/11	1/7	1/9	5/15	18/100	
$d_N/(d_S+const.)$	0.022	0.050	0.060	0.060	0.060	0.042	0.055	0.028 ^c	0.030	0	0	0.018	0.024	0.06	0.043	
<i>COII</i>																
M_N/M_S	7/31	0/10	4/16	1/8	3/15	3/9	0/5	6/27	0/5	1/4	2/5	2/2	1/3	0/9	10/49	
$d_N/(d_S+const.)$	0.045	0	0.035	0.008	0.053	0.093	0	0.025 ^c	0	0.044	0.070	0.080	0.030	0	0.073	
<i>ATP8</i>																
M_N/M_S	8/9	5/6	4/8	0/5	4/7	1/3	1/5	6/8	2/1	0/1	1/1	1/0	0/0	1/2	12/14	
$d_N/(d_S+const.)$	0.050	0.102	0.049	0	0.014	0.039	0.020	0.061 ^c	0.055	0	0.101	0.168	0	0.008	0.054	
<i>ATP6</i>																
M_N/M_S	25/23	5/4	15/16	9/5	10/13	4/9	7/3	23/23	8/6	3/7	7/5	4/0	2/2	6/6	42/39	
$d_N/(d_S+const.)$	0.174	0.191	0.187	0.212	0.168	0.079	0.300	0.236 ^c	0.245	0.134	0.352	0.367	0.066	0.148	0.259	
<i>COIII</i>																
M_N/M_S	8/38	2/10	3/20	0/11	3/18	3/14	2/14	11/37	2/9	0/8	2/10	0/3	2/6	3/11	16/59	
$d_N/(d_S+const.)$	0.024	0.055	0.014	0	0.025	0.027	0.019	0.062 ^c	0.033	0	0.037	0	0.057	0.124	0.037	
<i>ND3</i>																
M_N/M_S	5/9	1/1	3/6	2/4	3/4	2/2	2/2	5/10	1/3	1/3	1/5	2/1	1/1	2/1	7/18	
$d_N/(d_S+const.)$	0.118	0.037	0.075	0.058	0.088	0.197	0.103	0.109 ^c	0.087	0.055	0.045	0.103	0.045	0.108	0.157	
<i>ND4L</i>																
M_N/M_S	1/10	0/3	1/9	0/6	1/7	0/4	0/4	1/10	0/1	0/1	0/3	0/0	0/2	1/3	2/16	
$d_N/(d_S+const.)$	0.004	0	0.009	0	0.015	0	0	0.005	0	0	0	0	0	0.018	0.005	

Table 3. continued

<i>ND4</i>																
M_N/M_S	9/67	2/14	6/46	0/20	6/38	3/19	0/20	14/56	4/15	2/8	3/13	3/5	1/12	2/16	21/103	
$d_N/(d_S+const.)$	0.023	0.030	0.015	0	0.026	0.034	0	0.029 ^c	0.049	0.041	0.037	0.101	0.027	0.016	0.024	
<i>ND5</i>																
M_N/M_S	35/75	5/17	25/38	12/21	21/26	9/23	11/25	38/60	8/18	7/6	8/10	4/10	3/12	8/17	56/112	
$d_N/(d_S+const.)$	0.181	0.154	0.193	0.161	0.220	0.106	0.220	0.098 ^c	0.094	0.171	0.135	0.084	0.046	0.086	0.145	

<i>ND6</i>															
M_N/M_S	9/26	2/7	7/16	5/8	3/16	1/7	0/8	7/23	1/7	1/2	2/7	1/4	1/4	1/4	14/42
$d_N/(d_S+const.)$	0.059	0.060	0.061	0.079	0.047	0.054	0	0.033 ^c	0.014	0.048	0.067	0.064	0.031	0.015	0.047
<i>Cytb</i>															
M_N/M_S	31/41	5/12	18/31	12/14	12/26	3/8	8/8	29/41	5/11	1/10	8/11	2/9	5/4	8/10	47/65
$d_N/(d_S+const.)$	0.145	0.080	0.163	0.213	0.126	0.111	0.193	0.199 ^c	0.138	0.035	0.127	0.054	0.240	0.351	0.166
No. recurrent synonymous sites	81	2	28	5	14	4	2	71	8	3	4	2	1	1	241
No. recurrent non-synonymous sites	35	2	14	5	6	1	3	30	3	0	2	0	0	3	106
π ($\times 10^{-3}$)	3.79	3.59	3.22	2.69	3.68	4.26	3.16	1.85	1.78	1.93	1.99	1.66	1.01	1.41	3.06
\pm S.D.	± 0.12	± 0.23	± 0.17	± 0.18	± 0.23	± 0.16	± 0.37	± 0.08	± 0.14	± 0.23	± 0.15	± 0.19	± 0.18	± 0.09	± 0.10
θ ($\times 10^{-3}$)	10.55	4.12	7.54	4.34	7.01	3.99	4.61	10.02	3.75	2.54	3.79	2.11	2.24	3.44	15.36
\pm S.D.	± 2.45	± 1.49	± 1.96	± 1.31	± 2.08	± 1.30	± 1.69	± 2.28	± 1.25	± 1.00	± 1.34	± 0.79	± 0.78	± 1.03	± 3.14
Tajima's D	-2.139 ^d	-0.552	-2.014 ^d	-1.455	-1.806 ^d	0.265	-1.385	-2.681 ^g	-2.117 ^d	-1.121	-2.012 ^d	-0.933	-2.249 ^e	-2.198 ^e	-2.528 ^g
Fu&Li's D	-4.203 ^f	-0.668	-4.186 ^f	-2.524 ^d	-2.744 ^d	0.963	-1.015	-8.572 ^f	-3.300 ^f	-1.407	-2.847 ^f	-1.272	-3.138 ^f	-4.440 ^f	-7.738 ^f
Fu&Li's F	-3.883 ^f	-0.735	-3.956 ^f	-2.560 ^d	-2.871 ^d	0.870	-1.291	-6.894 ^f	-3.437 ^f	-1.519	-3.024 ^f	-1.357	-3.349 ^f	-4.339 ^f	-5.872 ^f
To Thr&Val	78	10	36	13	20	7	12	74	10	6	14	9	3	14	158
From Thr&Val	47	2	34	10	22	2	3	52	9	2	13	3	5	11	99

^aRimaibe (N=10), Foulbe (10), Mossi (10), all from Burkina Faso, plus 2 Mandenka and 1 Ghanan (subset of West Africa)

^bIndividuals of African descent from the Dominican Republic (subset of West Africa)

^cSignificant ($P < 0.0001$) difference between African and non-African distribution of $d_N/(d_S+constant)$ values.

^d $P < 0.05$

^e $P < 0.01$

^f $P < 0.02$

^g $P < 0.001$

Table 4. The rate of non-synonymous/synonymous changes in relation to continental geography and haplogroup diversity.

Clade	Geography	Diversity (ρ) ^a	M_N	M_S	M_N/M_S	S.D.
L1c	West and Central Africa	12.7	25	65	0.38	
L0d	South Africa	9.1	15	42	0.36	
M7	Southeast Asia	7.3	8	21	0.38	
M8(CZ)	Northeast Asia	6.7	12	31	0.39	
U	Europe	6.6	40	111	0.36	
"Old" haplogroups		>6			0.37	0.01
K	Europe	2.6	10	23	0.43	
C	Northeast Asia	2.6	8	13	0.62	
D1	Native American	2.5	5	16	0.31	
L1b1	West Africa	2.2	10	11	0.91	
L2a1	West Africa	1.8	14	16	0.88	
H1	Europe	1.2	11	19	0.58	
"Young" haplogroups		<3			0.62	0.24

^a average number of synonymous substitutions to the ancestral sequence of the haplogroup. One or two of the most frequent haplogroups for a) Africa, b) Europe, c) East Asia and Native Americans are displayed for the upper and lower range of sequence diversity. Using the average mutation rate of synonymous transitions the threshold of 6 synonymous transitions for the "older" haplogroups means coalescent times of the haplogroups >40 thousand years, while the threshold of <3 synonymous transitions for the "younger" haplogroups means <20 thousand years of divergence.

Table 5. Significant differences ($p < 0.05$) in mutational direction in synonymous sites

Amino Acid	codon change		number of observed changes ¹	p	p'
	(count in CRS)		to/from		
Ala	GCG (8)	to GCA (80)	9/5	3×10^{-7}	2×10^{-4}
	GCU (43)	to GCC (123)	19/8	2×10^{-6}	0.002
Asn	AAU (32)	to AAC (132)	13/13	7×10^{-4}	0.042
Asp	GAU (15)	to GAC (51)	12/5	2×10^{-4}	0.004
Gln	CAG (8)	to CAA (82)	7/9	0.001	0.018
Glu	GAG (24)	to GAA (64)	18/10	4×10^{-4}	0.046
Gly	GGG (34)	to GGA (67)	44/37	0.005	
	GGU (24)	to GGC (87)	11/12	0.014	
Ile	AUU (124)	to AUC (196)	38/26	0.002	
Leu	CUG (45)	to CUA (276)	39/59	1×10^{-7}	0.006
	UUG (19)	to CUG (45)	4/3	0.099	
	UUA (73)	to CUA (276)	28/18	4×10^{-8}	2×10^{-4}
	CUU (65)	to CUC (167)	25/18	1×10^{-4}	0.032
	UUG (19)	to UUA (73)	14/11	5×10^{-4}	0.048
Lys	AAG (10)	to AAA (85)	6/10	0.004	
Met	AUG (40)	to AUA (167)	31/29	5×10^{-6}	0.008
Phe	UUU (77)	to UUC (139)	24/17	0.006	
Pro	CCG (7)	to CCA (52)	14/10	1×10^{-5}	3×10^{-4}
	CCU (41)	to CCC (119)	22/14	3×10^{-4}	0.008
Ser	AGU (14)	to AGC (39)	7/3	0.02	
	UCG (7)	to UCA (83)	5/10	0.02	
	UCU (32)	to UCC (99)	14/7	9×10^{-5}	0.008
Thr	ACU (52)	to ACC (155)	27/16	6×10^{-7}	0.002
	ACG (10)	to ACA (134)	24/11	9×10^{-16}	9×10^{-13}
Try	UGG (11)	to UGA (93)	18/36	5×10^{-4}	0.031
Tyr	UAU (46)	to UAC (89)	22/17	0.01	
Val	GUU (31)	to GUC (48)	15/6	0.008	
Total	NNG (223)	to NNA (1256)	229/237	4.6×10^{-52}	5.3E-19
	NNU (596)	to NNC (1444)	249/162	2.8×10^{-34}	1.8E-27
				inverse	
ND6	NNG (62)	to NNA (35)	14/19	0.03	
	NNU (72)	to NNC (6)	10/6	1.1×10^{-3}	

p is a Chi-square probability assuming equal rates of codon exchange and estimates the difference from the expected number of changes given the codon frequencies in the reference mtDNA sequence (ANDREWS *et al.* 1999).

p' is binomial probability taking into account additional transitional biases observed over the whole mitochondrial genome favouring transitions G to A over A to G and T to C over C to T by factors of 2.33 and 1.93, per respective nucleotides.

¹ - number of changes corresponds to mutations (including multiple hits per site) inferred in phylogenetic analysis (Figure 1)

Table 6. Number and location of recurrent mutations in the mtDNA coding region (577-16023).

Number of recurrences	nucleotide positions
14	709~r
10	13708(A to T)
8	1888~r, 8251
7	11914, 10398(A <-> T)
5	1438~r, 5460(A to T), 13105(I to V)
4	1598~r, 1719~r, 3010~r, 13928C(S to T), 13966(T to A), 15930~t, 5147, 13368, 15217
3	3394(Y to H), 5821~t, 12172~t, 14110(F to L), 15110(A to T), 15924~t, 5231, 6182, 6221, 7055, 8790, 9545, 9554, 9950, 12007, 12501, 13359, 15514
2	930~r, 1503~r, 2768~r, 3434(Y to C), 4025(T to M), 4048(D to N), 5046(V to I), 5773~t, 8027(A to T), 10084(I to T), 12236~t, 12950(N to S), 13759(A to T), 14687~t, 15758(I to V), 15927~t, 3666, 3915, 4562, 4580, 4688, 4703, 5417, 5471, 5585-nc, 6260, 6446, 6680, 6752, 7076, 7388, 8020, 8152, 8155, 8392, 8964, 8994, 9254, 9266, 9509, 9755, 9824, 9932, 10685, 10790, 11260, 11944, 12354, 12477, 12810, 14007, 14034, 14148, 14182, 14905, 15115, 15301, 15784, 15884-nc

Amino acid changes are indicated in parentheses, ~r – change in rRNA, ~t – change in tRNA sequences, nc – change in non-coding position. The following positions showed a single recurrence, having mutated twice:

593~t, 597~t, 719~r, 813~r, 827~r, 1018~r, 1193~r, 1243~r, 1694~r, 1811~r, 1822~r, 2245~r, 2332~r, 2352~r, 2416~r, 2706~r, 2757~r, 2772~r, 2789~r, 2885~r, 3203~r, 3206~r, 3505(T to A), 4500(S to P), 4596(V to I), 4824(T to A), 4917(N to D), 5442(F to L), 5910(A to T), 6253(M to T), 6261(A to T), 6480(V to I), 7389 (Y to H), 7444(Ter to K), 7569~t, 7673(I to V), 7805(V to I), 7853(V to I), 8329G~t, 8387(V to M), 8393(P to S), 8566(I to V), 8584(A to T), 9095(L to P), 9139(A to T), 9438(G to S), 9477(V to I), 9861(F to L), 9966(V to I), 10031~t, 10143(G to S), 10321(V to A), 10463~t, 11016(S to N), 11025(L to P), 12142~t, 12248~t, 12346(H to Y), 12358(T to A), 12397(T to A), 12940(A to T), 13135(A to T), 13145(S to N), 13651(T to A), 13879(S to P), 13889(C to Y), 14129(T to I), 14180(Y to C), 14315(S to N), 14798(F to L), 15287(F to L), 15314(A to T), 15317(A to T), 15323(A to T), 15326(T to A), 15479(F to L), 15907~t, 15928~t, 15939~t, 15951~t, 3483, 3591, 3693, 3777, 3834, 3852, 4038, 4117, 4200, 4248, 4655, 4715, 4823, 4883, 4907, 4916, 4937, 5054C, 5162, 5237, 5393, 5580-nc, 5581-nc, 5656-nc, 6026, 6179, 6392, 6431, 6455, 6827, 7184, 7337, 7424, 7861, 8050, 8104, 8227, 8269, 8277-nc, 8383, 8485, 8697, 8856, 9150, 9180, 9299, 9305, 9365, 9377, 9449, 9716, 9758, 9899, 10238, 10389, 10586, 10589, 10688, 11002, 11257, 11299, 11332, 11350, 11353, 11383, 11404, 11437, 11452, 11812, 11854, 12372, 12432, 12540, 12609, 12630, 12720, 12771, 13020, 13104, 13116, 13215, 13263, 13470, 13590, 13680, 13827, 13980, 14094, 14212, 14233, 14323, 14364, 14470, 14560, 14581, 14620, 14668, 15043, 15061, 15106, 15148, 15172, 15289, 15313, 15346, 15394, 15454, 15466, 15550, 15607, 15670, 15697, 15883, 15886-nc.

Table 7. Codon sites found to be under positive selection in the mitochondrion encoded protein genes in primates.

	gene	codon No	Nucleotide position	Nh	Post. Prob.
1	ND2	218	5121-3	0	0.9732
2	ND2	265	5262-4	2	0.9603
3	ATP6	10	8554-6	0	0.9521
4	ATP6	188	9088-0	0	0.9809
5	ND3	9	10083-5	3	0.9628
6	ND3	44	10188-0	0	0.9823
7	ND3	107	10377-9	0	0.9666
8	ND3	114	10398-0	8	0.9542
9	ND4L	9	10494-6	0	0.9641
10	ND4	55	10922-4	0	0.9897
11	ND4	424	12029-1	0	0.9738
12	ND5	109	12661-3	0	0.9514
13	ND5	202	12940-2	2	0.983
14	ND5	459	13711-3	1	0.9525
15	ND5	492	13810-2	0	0.9617
16	ND6	11	14641-3	0	0.954

The test of positive selection ($dN/dS \gg 1$) was applied to 13 protein coding genes of mtDNA by the phylogenetic tree involving 7 primate species: 1. Homo sapiens, 2. Pan troglodytes, 3. Gorilla gorilla, 4. Papio hamadryas, 5. Hylobates lar, 6. Pongo pygmaeus, and 7. Macaca sylvanus) using PAML (YANG 2002). A model of neutral selection on codon sites was rejected in favor of a model allowing for 3 ratios across sites with K=3 site classes ($p \approx 0$, $df = 5$). The 3-ratio model identified the listed 16 sites to be under significant (posterior probabilities > 0.95) positive selection ($dN/dS = 2.02$). Nucleotide positions and gene names are given as in human reference sequence (GenBank accession # NC_001807.3). Nh - Number of non-synonymous mutations per respective codon observed in the tree of 277 human mtDNA sequences (Figure 1).

Table 8. Predominant amino acid replacement types in mitochondrial genes by intra and interspecies comparisons.

	No ³	Ala<>Thr	Ile<>Val	Ile<>Thr	Phe<>Leu	Asn<>Ser	Met<>Thr	Ser<>Thr	Ile
Human-human ¹	414	0.273	0.145	0.08	0.07	0.058	0.029	0.01	0
Human-chimpanzee ²	167	0.18	0.132	0.078	0.042	0.066	0.072	0.042*	0
Chimpanzee-orangutan	452	0.119*	0.071*	0.082	0.066	0.031	0.053	0.04*	0
Cat-dog	96	0.055*	0.107	0.047	0.025*	0.02*	0.03	0.07*	0
Pig-cow	479	0.054*	0.09	0.035*	0.033	0.027	0.05	0.044*	0

¹ proportions of specified amino acid changes as reported in the tree of 277 human sequences (Figure 1).

² proportions of changes between rCRC (NC_001807.3) and Pan troglodytes (D38113). Orangutan (D38115); pig (AJ002189); dog (AY729880); cat (NC_001700); cow (V00654) sequences.

³ Total number of observed amino acid replacements

* Difference from the human-human pattern; χ^2 probability $p < 0.01$

Only the most frequent (>5% in at least one comparison) amino acid replacement types are reported.

Figure 2

