Analysis of the Complete Human mtDNA Genome: Methodology and Inferences for Human Evolution

M. Ingman and U. Gyllensten

The analysis of mitochondrial DNA (mtDNA) sequences has been a potent tool in our understanding of human evolution. However, almost all studies of human evolution based on mtDNA sequencing have focused on the control region, which constitutes less than 7% of the mitochondrial genome. The rapid development of technology for automated DNA sequencing has made it possible to study the complete mtDNA genomes in large numbers of individuals, opening the field of mitochondrial population genomics. Here we describe a suitable methodology for determining the complete human mitochondrial sequence and the global mtDNA diversity in humans. Also, we discuss the implications of the results with respect to the different hypotheses for the evolution of modern humans.

Mitochondrial Genetics

Nucleotide sequences from the mitochondrial genome differ from those of the cell nucleus in a number of characteristics. It is these attributes, which include high copy number per cell, lack of recombination (Olivio et al. 1983), high substitution rate (Brown et al. 1979), and uniparental inheritance (Giles et al. 1980; Gyllensten et al. 1985), that have made mitochondrial DNA (mtDNA) an attractive source of information for phylogenetic studies. Mitochondria constitute the powerhouse of the cell and have a key function in the production of adenosine triphosphate (ATP). The synthesis of ATP requires the transfer of energy across five protein complexes (I–V), each composed of a series of polypeptides encoded either by the nuclear or mitochondrial genome. The circular mitochondrial genome encodes 13 polypeptides, and in addition, 22 transfer RNA (tRNA) genes and 2 ribosomal RNA (rRNA) genes necessary for transcription and translation of the mitochondrial genome. The mitochondrial compartments within each cell harbor between 0 and 15 copies of the mtDNA genome, resulting in the entire cell containing 100–1000 copies (Cavelier et al. 2000). This multicopy feature facilitates the analysis of mtDNA sequences from a wide range of tissue sources, as well as from the partially degraded material found in forensic evidence materials and archaeological remains (Allen et al. 1998; Krings et al. 1997). Despite the large number of mtDNA copies per cell and the presence of mtDNA copies differing at one or more nucleotide sites within the same individual (heteroplasmy), recombination among different mtDNA molecules has not been convincingly demonstrated. The egg cytoplasm contains a large number of mitochondrial particles, while the midpiece of the sperm only carries 5–10 mitochondria. During intraspecific fertilization, paternal mitochondria are specifically labeled and directed toward selective degradation via the ubiquitin pathway, resulting in an exclusively maternal inheritance of mtDNA (Sutovsky et al. 1999). The mechanism for identification and degradation of paternal mitochondria can be inactivated in interspecific crosses, where mtDNA is inherited in a biparental fashion, although the paternal mtDNA represents only a minor portion of the mtDNA in the zygote (Gyllensten et al. 1991; Kaneda et al. 1995). The lack of recombination and the uniparental mode of inheritance in intraspecific crosses results in mtDNA molecules being inherited in a clonal fashion and that the evolutionary history of maternal lineages be reconstructed.

The substitution rate in the mtDNA genome amounts to 5–10 times that of nuclear DNA (Brown et al. 1982). The high substitution rate has been attributed to the lack of mitochondrial histones and a high concentration of oxidative radicals. This higher substitution rate provides increased resolution of more recent evolutionary events as compared to the infor-

From the Department of Genetics and Pathology, Section of Medical Genetics, Rudbeck Laboratory, Dag Hammarskjöldsväg 20, University of Uppsala, S-751 85 Uppsala, Sweden. This work was supported by grants from the Swedish Natural Sciences Research Council and Beijer Foundation. We are grateful for the comments of Mark Stoneking on an earlier version of the manuscript. Address correspondence to Ulf Gyllensten at the address above or e-mail: ulf.gyllensten@genpat.uu.se. This paper was delivered at a symposium entitled "Primate Evolutionary Genetics" sponsored by the American Genetic Association at Town and Country Resort and Convention Center, San Diego, CA, USA, May 19–20, 2001.

 92:454–461

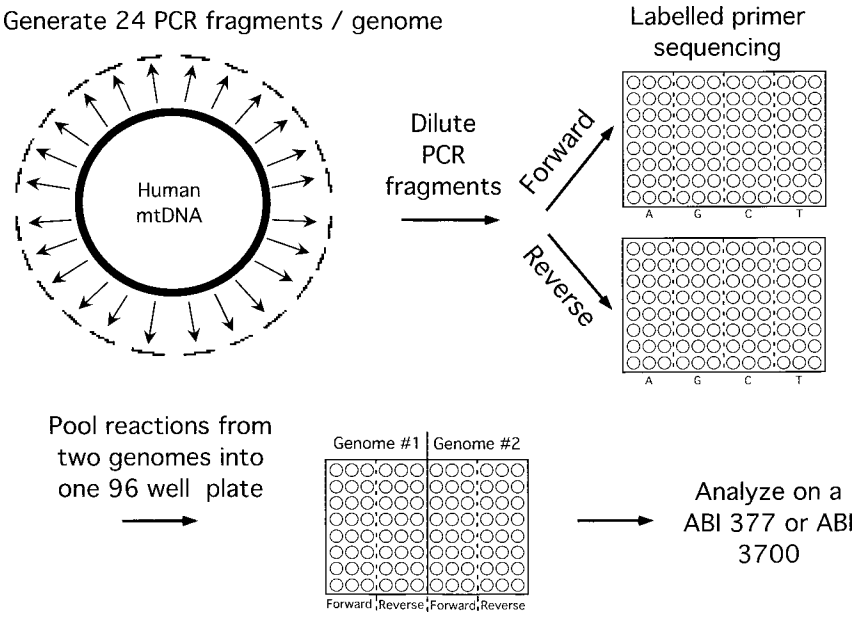

Figure 1. Overview of the methodology used for expedient sequencing of the complete human mtDNA genome.

mation provided by nuclear gene studies. For example, in our recent study of 53 individuals, a total of 657 segregating sites were identified in the 16,500 bp mitochondrial genome (Ingman et al. 2000), as compared to 33 sites in 10,200 bp of sequence from Xq13.3 of 69 globally distributed individuals (Kaessmann et al. 1999).

Methodology for Rapid Sequencing of the Complete mtDNA Genome

To facilitate the expedient analysis of complete mtDNA genomes, we optimized the sample flow as outlined in Figure 1. We used a set of 24 pairs of primers (Rieder et al. 1998) to polymerase chain reaction (PCR) amplify the entire human mitochondrial genome into fragments ranging from 765 to 1162 bases long and overlapping by about 200 bases. PCR products were examined on a 1% agarose gel and, depending on the amount of product obtained, were diluted between 1:2 and 1:6 for direct sequencing. It has not been necessary to purify the PCR products prior to sequencing. The forward and reverse PCR primers were synthesized with the universal M13 (–21) forward primer or M13 reverse primer, respectively, allowing for BigDye primer (Applied Biosystems, Foster City, CA) cycle sequencing. Cycle sequencing protocols were as recommended by the manufacturer for 0.5× reactions. Extension products were pooled, precipitated using ethanol, and air dried before rehydrating in loading buffer. Electrophoresis was performed on an ABI377 96-lane machine, loading only every second lane. Thus a complete genome (24 forward products and 24 reverse) is analyzed on one 4.5% 29:1 polyacrylamide gel. More recently the reactions have been analyzed on an ABI3700 capillary sequencer, enabling an even higher throughput of samples. Fragment assembly was performed by the program Sequencher (GeneCodes Corporation, Ann Arbor, MI), using a complete mtDNA genome sequence as a template. The template sequence was then removed and the contig checked and edited. Finally, the template sequence was again imported to verify all positions found to be variable.

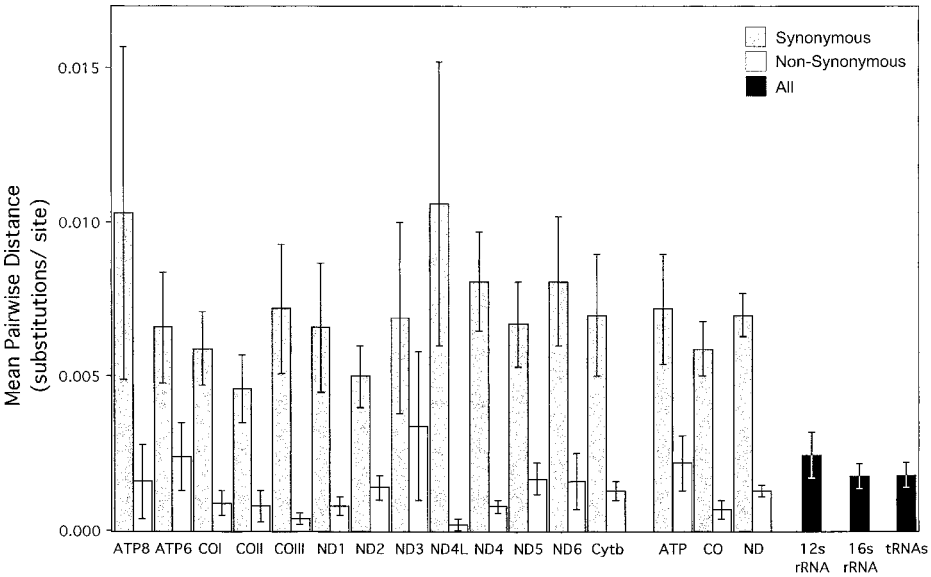

Figure 2. Mean pairwise synonymous and nonsynonymous distances with standard error bars for each of the 13 protein coding genes and 3 gene complexes based on comparisons among the 53 human mtDNA genomes. Mean pairwise distance with standard error bars for the two rRNA genes and the combined (22) tRNA genes.

Sampling Strategy and mtDNA Sequences

In our initial study we analyzed 53 individuals representing 14 of the major linguistic phyla in an attempt to assess the global genetic diversity in humans while limiting the number of samples (Ingman et al. 2000). This sampling strategy attempts to avoid the bias inherent in selecting individuals based on current world demographics, such as current population size or geographic location (Maddison et al. 1992). From nearly 900 kb sequenced, five heteroplasmic sites were confidently identified. A total of 657 segregating sites (141 in the D-loop; 516 outside) were identified among these 53 individuals, of which 283 (80 in the D-loop; 203 outside) showed the same polymorphism in at least two individuals (Ingman et al. 2000). The pairwise sequence distances between mtDNAs, corrected for multiple substitutions (Tamura and Nei 1993), vary from 6.0×10^{-5} substitutions per site between two South American Indians (Warao) to 6.8×10^{-3} substitutions per site between two Africans (Mbenzele pygmy and San). The average distance between mtDNA genomes is 3.8×10^{-3} substitutions per site.

Pattern of Selection

The pattern of selection acting on individual mtDNA coding regions was examined by a comparison of the rate of synony-

mous (dS) and nonsynonymous (dN) changes per such site. The substitution rates in mitochondrial protein coding genes, calculated using the chimpanzee as an outgroup and assuming a human-chimpanzee split of 5 million years, vary from 0.81 (for NADH dehydrogenase 4L) to 1.79 $\times 10^{-8}$ (for NADH dehydrogenase 2) substitutions per site per year (Table 1). All individual coding regions showed an excess of synonymous substitutions, consistent with the operation of conserving (purifying) selection (Figure 2). The significance of the difference between dS and dN for each protein coding gene was analyzed with a one-tailed z test. Although it is inappropriate to place too much emphasis on these test results due to the low number of observed changes, all comparisons showed a trend toward significance, with z scores in genes ranging between 1.542 ($P = .063$) and 4.421 ($P < .001$) (Table 1). Since the proteins encoded by mtDNA constitute parts of four different complexes, it is of interest to examine whether these complexes differ with respect to the type and degree of selection. Between all complexes, the same pattern of an excess of synonymous relative to nonsynonymous change was evident. This analysis of the selection pressure acting on human mtDNA peptides based on a set of complete human mtDNA genomes shows that purifying selection is maintaining the structure of mitochondrial proteins in humans.

The substitution rate in protein coding genes was generally found to be higher than in the rRNA and tRNA genes (Table 1). The mean rate for the protein coding regions (1.46×10^{-8}) is 3-fold, 1.7-fold, and 3-fold higher than for the 12s, 16s, and tRNA genes, respectively. The mean pairwise distance for the rRNA and tRNA genes is similar to that for the nonsynonymous changes in the coding regions (Figure 2), implying that purifying selection is preserving the structure of the rRNA and tRNA molecules. At present, the available dataset is insufficient to address the issue of whether the rRNA and tRNA genes are more highly conserved than the protein coding genes.

Footprints of Recombination?

Recently, based on an analysis of the correlation between linkage disequilibrium (LD) and distance between variable sites, it was claimed that mtDNA sequences show signs of recombination (Awadalla et al. 1999; Eyre-Walker et al. 1999). These analyses were criticized for the methodology employed (Kumar et al. 2000), particularly for the use of an LD measure (Hill and Robertson measure, r2) that doesn't take allele frequency into account. Also, the datasets initially used consisted of either D-loop sequences or of a limited selection of restriction fragment length polymorphism (RFLP) sites; both types of data may be unsuitable for these purposes. We examined LD among all informative sites in our set of complete mtDNA genomes (including the D-loop) using a standard estimate (D'), which allows for all variable positions to be examined independent of their allele frequency (Lewontin 1964). In this analysis, African and non-African sequences were studied separately. There is no correlation between D' and nucleotide distance between sites in the 53 sequences (African $R^2 = 1 \times 10^{-3}$; non-African $R^2 = 5 \times 10^{-3}$) (Ingman et al. 2000). Also, no evidence of a correlation was seen between r2 and distance in this dataset ($R^2 = 2.23 \times 10^{-6}$ and 1×10^{-3}, respectively). Based on these results (Ingman et al. 2000), and those of others (Elson et al. 2001), we maintain that there is no evidence that recombination contributes to the evolution of human mtDNA.

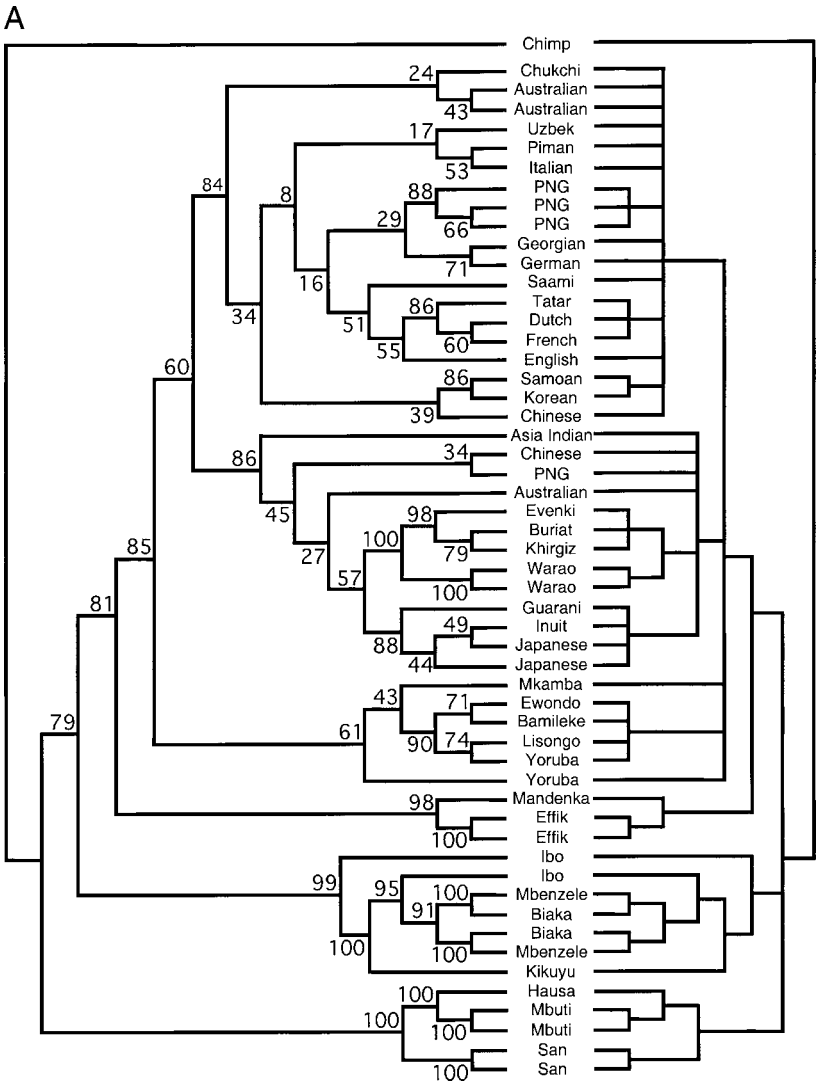

Figure 3. Neighbor-joining cladograms reconstructed from **(a)** 53 mitochondrial genomes minus the D-loop, and **(b)** 53 D-loop sequences. The bifurcating cladograms on the left of each figure show bootstrap values (1000 replicates) on the nodes. The trees on the right are identical except that all branches with less than 80% bootstrap support have been collapsed.

Homoplasy in the Human Mitochondrial D-Loop

Among the 53 sequences examined, nearly 30% of the polymorphic sites are confined to the D-loop, which represents a mere 7% of the genome. However, previous analyses have indicated the existence of hotspots for substitutions in the D-loop (Mad-

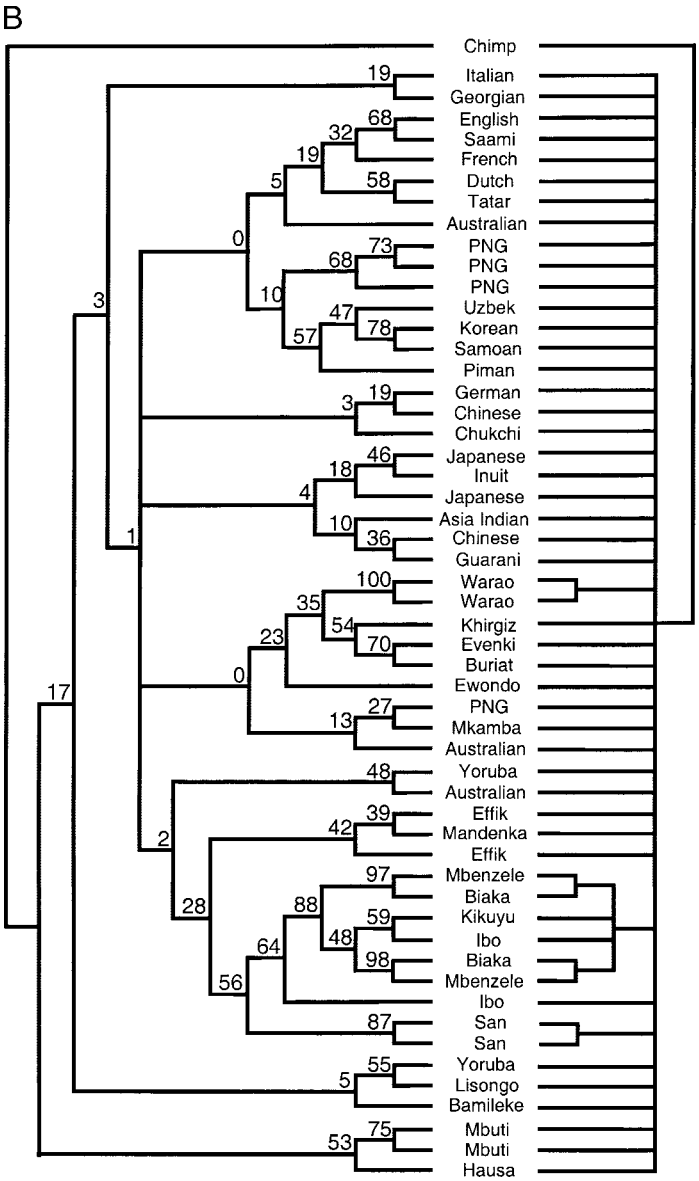

Figure 3. Continued

dison et al. 1992; Tamura and Nei 1993; Wakeley 1993), rendering this particular sequence less suitable for evolutionary inferences.

Since the molecular clock hypothesis assumes that DNA sequence evolution is approximately constant over time for all evolutionary lineages, we performed tests to assess the suitability of D-loop data to the assumption of clocklike rates. A test comparing the log likelihood of trees reconstructed with and without the molecular clock assumption was used to examine the supposition that the mtDNA lineages evolve at "clocklike" rates within humans (Strimmer and von Haesseler 1996). The human mtDNA sequences, excluding the D-loop, have evolved at approximately constant rates ($\delta = 64.7$, df = 51, $P = .095$), while the hypothesis of constant rates is rejected for the D-loop ($\delta = 107.98$, df = 51, $P < .001$). A relative rate test (Sarich and Wilson 1973), using a gorilla sequence as an outgroup, further shows that there is no significant difference between the evolutionary rate of mtDNA on the lineages leading to human and chimpanzee, excluding the D-loop ($P = .123$). In contrast, the D-loop has not evolved at a constant rate across human lineages ($P < .001$), and is consequently less suitable for dating evolutionary events.

Given that the D-loop did not conform to clocklike behavior, we examined the substitution pattern in this part of the genome, using the unique opportunity provided by having access to an evolutionary tree constructed from linked, but independent, DNA sequences in the coding regions of the mtDNA genome. We first reconstructed neighbor-joining (Saitou and Nei 1987) trees for the mitochondrial genome sequences without the D-loop (Figure 3a; left tree) and for the D-loop sequences only (Figure 3b; left tree). The tree reconstructed from the D-loop sequences alone has considerably less statistically supported branching than the tree reconstructed from the rest of the genomes. This difference is particularly evident when branches with less than 80% bootstrap support are collapsed in both trees (Figure 3a,b; right trees). In order to study the pattern of substitution at individual nucleotide sites in the mitochondrial D-loop, we inferred the number of changes that must have occurred for the D-loop data to fit a phylogenetic tree constructed from the coding region of these genomes. This approach depends on complete linkage between the control region sequences and the rest of the mtDNA genome. The nucleotide sites in the D-loop range between those that perfectly fit the inferred phylogeny (no homoplasy) and those that appear to have changed independently in up to 10 parallel lineages (Figure 4). Of the 80 parsimony informative sites in the control region, 57 are present independently on more than one branch of the tree, representing parallel substitutions or back mutations. The sites in the D-loop with parallel changes appear to cluster in groups. Regions that have been previously identified as "conserved sequence blocks" (CSBs) and those associated with certain functions, such as termination associated sequences (TAS) (Doda et al. 1981) and possible control element (CE) (Ohno et al. 1991), contain no sites with parallel changes (Figure 4).

Given the high number of parallel substitutions at many nucleotide sites in the D-loop, we determined the relative influence of each of the parsimony informative sites on tree reconstruction by calculating a ratio of the minimum number of changes required for them to fit the inferred phylogeny and the frequency of lineages carrying the particular variant at that site (Figure 4). Sites with a higher ratio are considered to be in greater conflict with the data from the rest of the mitochondrial genome than sites with a low ratio. Sites having more than two nucleotide variants were first removed from the D-loop dataset, and conflicting sites were then sequentially removed, beginning with the sites showing the highest ratio. With the

Table 1. Statistics for the different parts of the human mitochondrial genome

Component	Length[a]	S[b]	Inform.[c]	Sub. rate[d]	dS-dN[e]	SE[f]	z[g]	P[h]
ATP8	207	8	6	1.33×10^{-8}	0.0087	0.0058	1.542	.063
ATP6	681	30	14	1.52×10^{-8}	0.0042	0.0021	1.977	.025
COI	1542	49	21	1.05×10^{-8}	0.0050	0.0013	3.755	<.001
COII	684	23	8	1.33×10^{-8}	0.0038	0.0013	3.015	.002
COIII	783	29	10	1.25×10^{-8}	0.0069	0.0021	3.332	<.001
ND1	957	27	13	1.33×10^{-8}	0.0058	0.0022	2.769	.003
ND2	1043	42	15	1.79×10^{-8}	0.0037	0.0011	3.220	<.001
ND3	345	13	6	1.55×10^{-8}	0.0033	0.0018	1.745	.042
ND4L	297	10	5	0.81×10^{-8}	0.0102	0.0045	2.274	.012
ND4	1377	54	25	1.78×10^{-8}	0.0073	0.0016	4.421	<.001
ND5	1811	74	28	1.55×10^{-8}	0.0051	0.0015	3.435	<.001
ND6	524	25	11	1.20×10^{-8}	0.0065	0.0023	2.711	.004
cyt *b*	1140	42	16	1.76×10^{-8}	0.0060	0.0020	2.840	.003
12s rRNA	954	23	7	0.62×10^{-8}	—	—	—	—
16s rRNA	1560	32	13	0.87×10^{-8}	—	—	—	—
tRNAs	1512	37	12	0.50×10^{-8}	—	—	—	—
ATP	888	38	20	1.47×10^{-8}	0.0050	0.0020	2.460	.008
CO	3009	101	39	1.16×10^{-8}	0.0052	0.0010	5.451	<.001
ND	6357	245	103	1.54×10^{-8}	0.0060	0.0010	7.633	<.001

[a] Length in base pairs.
[b] Number of segregating sites.
[c] Number of phylogenetically informative sites.
[d] Substitution rate (substitutions per site per year).
[e] Difference between mean pairwise synonymous and nonsynonymous distance.
[f] Standard error of note e.
[g] Result of one-tailed *z* test.
[h] *P* value of note g.

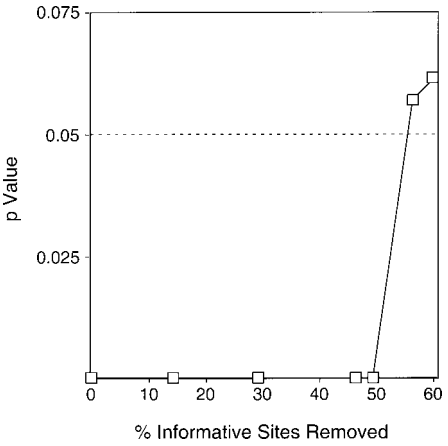

Figure 5. Graph indicating how the D-loop becomes "clocklike" as sites are removed. The critical significance level is shown on the *y*-axis and the percentage of informative D-loop sites that were removed on the *x*-axis. In reference to Figure 4, the data points on the *x*-axis are 14% (>0.6), 29% (>0.4), 46% (>0.3), 49% (>0.25), 55% (>0.24), and 60% (>0.2). Clocklike behavior is reached with the removal of 55% (44 of 80) of the informative sites (72, 93, 95, 143, 146, 150, 151, 152, 182, 185, 189, 194, 198, 200, 204, 456, 16111, 16126, 16145, 16148, 16172, 16184, 16187, 16188, 16189, 16209, 16214, 16234, 16249, 16265, 16270, 16271, 16291, 16293, 16304, 16319, 16325, 16357, 16362, 16390, 16399, 16438, 16519, 16527).

removal of 44 of the 80 phylogenetically informative sites in the D-loop, "clocklike" behavior was achieved, as determined by a test (Strimmer and von Haesseler 1996) that compares the log likelihoods of trees constructed with and without the molecular clock assumption (Figure 5). A neighbor-joining tree (Saitou and Nei 1987) was then reconstructed for the D-loop sequences with these 44 sites removed. The D-loop tree based on these sites (tree not shown) not surprisingly showed very little bootstrap-supported resolution in the branching pattern. For the purpose of comparison, a second neighbor-joining tree was reconstructed from the entire mitochondrial genomes, but with these 44 sites removed. When branches of less than 80% bootstrap support are collapsed (Figure 6; left tree), this tree gives some improvement in resolution when compared to a tree reconstructed with the complete genomes (Figure 6; right tree), even in one deep African branch (group of Ibo, Mbenzele, Biaka, Mbenzele, and Kikuyu). The tree reconstructed from the mitochondrial sequences with 44 D-loop sites removed also has some improvement in terminal branching resolution when compared to the tree constructed from the sequences with the D-loop completely removed. For example, more resolution is evident for the group of Saami, Tatar, Dutch, French, and English. Thus with the removal of sites showing the highest number of parallel changes, the D-loop contributes a small but useful amount of information for the construction of a robust mtDNA tree.

In summary, there is at least a 10-fold difference in the number of inferred changes between sites, indicating a pronounced variation in substitution rate among sites in the D-loop. More than 70% of the informative sites in the D-loop have undergone parallel substitution. The D-loop can be forced to conform to a state of constant evolutionary rate across branches, but this requires the removal of more than 50% of the parsimony informative sites. Exclusion of these sites from a dataset of complete mitochondrial genomes results in a small improvement in the stability and resolution of the mtDNA tree.

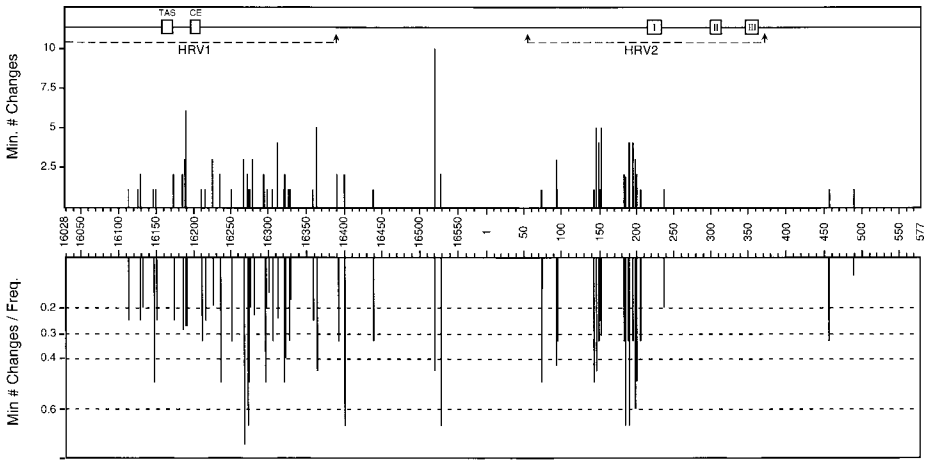

Figure 4. Graphical representation (top) of the minimum number of changes per site required for the D-loop data to fit to the collapsed tree on the right of Figure 3a. The regions listed are HVR1 and HVR2 (positions 16024–16383 and 57–372, respectively), TAS (Doda et al. 1981) (sites 16157–16172), CE (Ohno et al. 1991) (sites 16194–16208), and the Conserved Sequence Blocks I to III (marked in the figure as I, II, and III) (sites 216–235, 299–315, and 346–363, respectively). Below is a graph of the minimum number of changes divided by the frequency of the variant at each site. The dashed lines represent threshold values for site removal. These sites are >0.6 (and sites with more than two variants) (185, 189, 16111, 16184, 16188, 16265, 16270, 16291, 16399, 16527); 0.4–0.6 (72, 93, 143, 146, 198, 200, 16145, 16234, 16271, 16293, 16319, 16362, 16519); 0.3–0.4 (95, 150, 152, 182, 194, 204, 456, 16209, 16249, 16304, 16325, 16390, 16438); 0.25–0.3 (16187, 16189); 0.24–0.25 (151, 16126, 16148, 16172, 16214, 16357); and 0.2–0.24 (195, 16278, 16311).

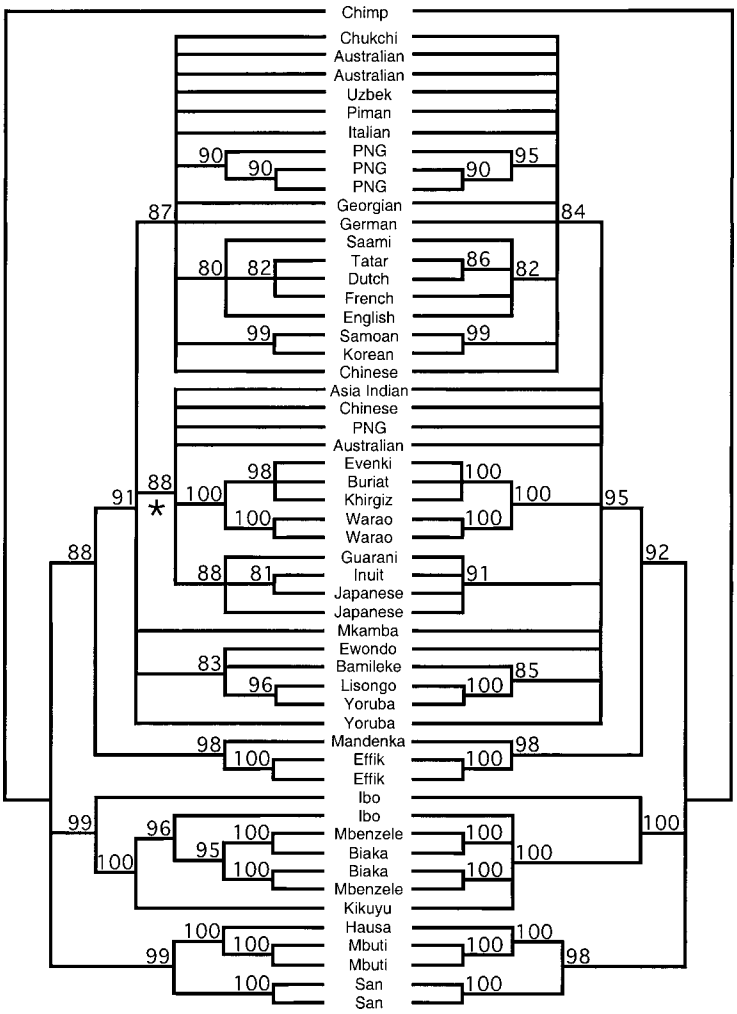

Figure 6. Neighbor-joining trees reconstructed from 53 mitochondrial genomes. All branches with less than 80% bootstrap support (1000 replicates) have been collapsed. The cladogram on the left represents the complete mitochondrial sequences with removal of the 42 D-loop sites showing the highest degree of homoplasy; and the right tree is reconstructed from complete genomes. The branch marked with an asterisk on the left tree highlights a supported branch that is unsupported on the right tree.

The mtDNA Genome Diversity and Origin of Modern Humans

Much of the discussion regarding the origin of modern humans has focused on two main competing hypotheses on the origin of modern humans. The "multiregional" hypothesis proposes that the transformation to anatomically modern human forms occurred in different parts of the world in parallel from a widely distributed progenitor species, such as *Homo erectus* (Wolpoff 1989; Wolpoff et al. 2001). Support for this hypothesis has come mainly from fossil evidence, where reports have claimed evidence of cultural and morphological continuity between archaic forms and modern humans outside of Africa (Wolpoff 1989). The alternative "recent African origin" hypothesis states that anatomically modern humans originated in Africa 100,000–200,000 years ago and subsequently spread to other parts of the world, eventually replacing the archaic human forms with little or no genetic mixing (Cann et al. 1987; Vigilant et al. 1989). A number of mtDNA studies have provided support for a recent African origin of modern humans (Horai et al. 1995; Ruvolo et al. 1993; Vigilant et al. 1991; Zietkiewicz et al. 1998). However, these results have been criticized for their lack of statistical support for tree topology, especially the deep African branches (Nei 1992; Templeton 1992). Evidently, lacking sufficiently strong empirical data, it is impossible to confidently place the root of modern human mtDNA lineages in sub-Saharan Africa. The neighbor-joining (Saitou and Nei 1987) tree reconstructed from our complete mtDNA sequences, excluding the D-loop, has a strongly supported basal branching pattern, with the three deepest branches leading exclusively to sub-Saharan mtDNAs and the fourth branch containing both Africans and non-Africans (Figure 7). The deepest statistically supported branch (bootstrap = 100) provides compelling evidence of a human mtDNA origin in Africa.

In order to date the most recent common ancestor (MRCA) for human mtDNA, the substitution rate was estimated from the mean genetic distance between all humans and the one chimpanzee sequence (0.17 substitutions per site) and the assumption, based on paleontological (Andrews 1992) and genetic (Kumar and Hedges 1998) evidence, of a divergence time between humans and chimpanzees of 5 million years, to be 1.70×10^{-8} substitutions per site per year. The MRCA based on the maximum distance between two humans (5.82×10^{-3} substitutions per site between the Africans: Mkamba and San) is then estimated to be 171,000 ± 50,000 years before present (YBP) (95% CI). This estimate of genetic distance was calculated with the Tamura and Nei (1993) model for nucleotide substitution and gamma rates. Due to the relatively close relationship of these sequences, the use of gamma rates may provide an overestimate of genetic distance and return an inflated standard error. If the same replacement model is used, but without gamma rates, the deepest split becomes 5.67×10^{-3} substitutions per site, yielding an estimate of 167,000 ± 18,000 YBP. We can also estimate the age of the MRCA for the youngest clade that contains both African and non-African sequences (Figure 7; node marked "‡") from the mean distance of all members of that clade to their common node (8.85×10^{-4} substitutions per site), as 50,000 ± 27,500 YBP (95% CI). Without gamma rates this becomes 49,000 ± 4500 YBP. Since genetic divergence is expected to precede the separation of populations, this date can be considered as the lower bound for an exodus from Africa. These dates evidently depend on the assumed divergence time of chimpanzee and human.

RFLP Haplogroups and Genome Sequence Variation

Mitochondrial DNA phylogenies have long been based on fragments of the mtDNA genome: either D-loop sequences or data from small sets of RFLPs distributed across the genome (Torroni et al. 1998, 2000; Wallace et al. 1999). The drawbacks of relying solely on the D-loop have been

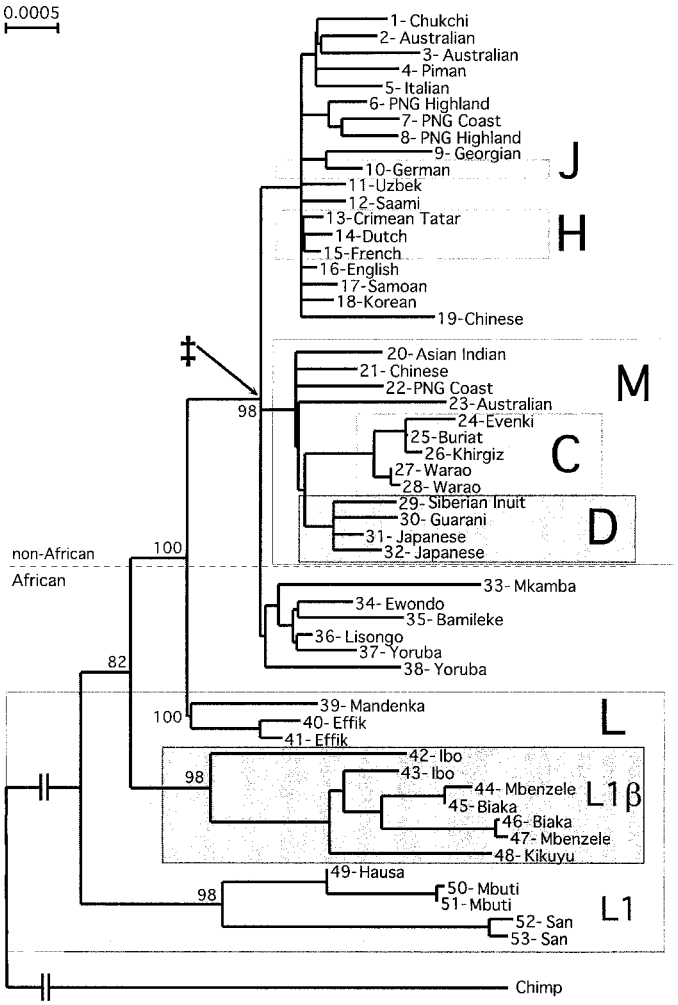

Figure 7. Neighbor-joining phylogram (Saitou and Nei 1987) based on 53 complete human mtDNA genome sequences (but excluding the D-loop), constructed using PAUP*4.0 Beta (Sinauer Associates, Sunderland, MA) and bootstrapped with 1000 replicates (bootstrap values shown on nodes). The population origin of the individual is given at the twigs. Individuals of African descent are found exclusively below the dashed line and non-Africans above. The node marked "‡" refers to the MRCA of the youngest clade containing both African and non-African individuals. The boxed areas define previously described haplogroups (Torroni et al. 1998, 2000; Wallace et al. 1999) that can be found within our dataset. The sites that define these haplogroups are M, 10394+ *Dde*I and 10397+ *Alu*I; L, 3592+ *Hpa*I; L1, 10806+ *Hin*fI, L1(_); 10319+ *Alu*I; C, 13262+ *Alu*I; D, 5176– *Alu*I; H, 7025– *Alu*I; and J, 13704– *Bst*NI. Several lineages are defined with D-loop sites, but these are uninformative in our dataset with the exception of an *Hin*fI site gain at 16389 defining lineage L2, which groups the Mandenka, 2 Effik, and 2 Mbuti sequences.

discussed previously. RFLP analysis on the other hand is limited by the recognition sequences of available enzymes. Since only a fraction of the genome is analyzed, many informative polymorphisms may not be detected, potentially resulting in ambiguities in tree topology and difficulty in resolving differences between closely related sequences. The extent to which data obtained from D-loop sequencing and RFLP analysis have affected phylogeny reconstruction has not been known, but we are now in a position to assess these influences using our dataset of 53 complete human mtDNA sequences from diverse genetic backgrounds. We compared the information obtained from the dataset of complete mtDNA genomes with that from the RFLP sites used to define major mtDNA lineages. A number of continent- and population-specific mtDNA variants have been described on the basis of published RFLP data (Wallace et al. 1999) and are summarized on the website "MITOMAP A human mitochondrial genome database" (http://infinity.gen.emory.edu/mitomap.html). From the position of restriction sites and the enzyme used, we inferred the position of these nucleotide changes and superimposed these onto our data outside the D-loop. Due to differences in sampling strategy, we were unable to assimilate some of these previously identified haplogroups with our data. However, we would have expected our Mbuti pygmies to have contained the sites anticipated for haplogroup L2(γ), which are described as population-specific for the Mbuti. These sites were not present in the Mbuti we examined. In general the RFLP sites identified as lineage specific represent only a portion of the informative sites present in the complete sequence data. The RFLP sites define some of the major clades, such as those with the deepest branches. The complete genome data provide stronger support for the haplogroups and, more importantly, distinguish a number of entirely new haplogroups. For example, the clade including the six individuals previously discussed (Figure 7; sequences 33–38) was not identified by the RFLP sites. Phylogenetic trees reconstructed from RFLP data typically lack statistical support for major branches, resulting in ambiguous internal and external structure. Also, estimates of substitution rates obtained from RFLP data are based only on the sites examined, and therefore are less reliable. In contrast, the topology of the tree based on the entire mtDNA genome (excluding the D-loop) is very robust and the substitution pattern more suited to evolutionary analysis, leading to more firm conclusions (Ingman et al. 2000). Technological advances in recent years have facilitated large-scale DNA sequencing, and further possibilities exist for automation using a DNA chip (Chee et al. 1996). In light of this, reconstruction of human evolutionary history based on fractions of the mtDNA genome may no longer be necessary and, in fact, give an incorrect view of human evolution.

In summary, our results shows that whole mtDNA genome analysis will provide important information for studies of human evolution, signaling the advent of the field of mitochondrial population genomics. All these complete genome sequences are available on our website, which is intended as a resource for phylogenetics and evaluation of disease associated polymorphisms (http://www.genpat.uu.se/mtDB). Nevertheless, it should noted that mtDNA is only one locus and the information is therefore limited in that it only reflects the genetic history of females. Comparative studies of different genetic systems are needed for a balanced view. For nuclear DNA, the regions studied will have to span at least 100 kb in order to include a similar number of informative sites as in the 16.5 kb mtDNA genome. While determining 100 kb for 100–200 chromosomes may seem a daunting project, the technology for DNA analysis continues to develop and such stud-

ies are likely to be performed shortly, providing us with an ever more detailed understanding of our genetic history.

References

Allen M, Engstrom AS, Meyers S, Handt O, Saldeen T, von Haesseler A, Paabo S, and Gyllensten U, 1998. Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. J Forensic Sci 43:453–464.

Andrews P, 1992. Evolution and environment in the Hominoidea. Nature 360:641–646.

Awadalla P, Eyre-Walker A, and Smith JM, 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524–2525.

Brown WM, George M Jr, and Wilson AC, 1979. Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci USA 76:1967–1971.

Brown WM, Prager EM, Wang A, and Wilson AC, 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239.

Cann RL, Stoneking M, and Wilson AC, 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Cavelier L, Jazin E, Jalonen P, and Gyllensten U, 2000. MtDNA substitution rate and segregation of heteroplasmy in coding and noncoding regions. Hum Genet 107: 45–50.

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, and Fodor SP, 1996. Accessing genetic information with high-density DNA arrays. Science 274:610–614.

Doda JN, Wright CT, and Clayton DA, 1981. Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. Proc Natl Acad Sci USA 78:6116–6120.

Elson JL, Andrews RM, Chinnery PF, Lightowlers RN, Turnbull DM, and Howell N, 2001. Analysis of European mtDNAs for recombination. Am J Hum Genet 68:145–153.

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

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

Gyllensten U, Wharton D, Josefsson A, and Wilson AC, 1991. Parental inheritance of mitochondria in mice. Nature 352:255–257.

Gyllensten U, Wharton D, and Wilson AC, 1985. Maternal inheritance of mitochondrial DNA during backcrossing of two species of mice. J Hered 76:321–324.

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

Ingman M, Kaessmann H, Pääbo S, and Gyllensten U, 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713.

Kaessmann H, Heissig F, von Haesseler A, and Paabo S, 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet 22:78–81.

Kaneda H, Hayashi J, Takahama S, Taya C, Lindahl KF, and Yonekawa H, 1995. Elimination of parental mitochondrial DNA in intraspecific crosses during early mouse embyrogenesis. Proc Natl Acad Sci USA 92: 4542–4546.

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, and Paabo S, 1997. Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30.

Kumar S and Hedges SB, 1998. A molecular timescale for vertebrate evolution. Nature 392:917–920.

Kumar S, Hedrick P, Dowling T, and Stoneking M, 2000. Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931.

Lewontin RC, 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67.

Maddison D, Ruvolo M, and Swofford D, 1992. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. Syst Biol 41:111–124.

Nei M, 1992. Age of the common ancestor of human mitochondrial DNA. Mol Biol Evol 9:1176–1178.

Ohno K, Tanaka M, Suzuki H, Ohbayashi T, Ikebe S, Ino H, Kumar S, Takahashi A, and Ozawa T, 1991. Identification of a possible control element, Mt5, in the major noncoding region of mitochondrial DNA by intraspecific nucleotide conservation. Biochem Int 24:263–272.

Olivio PD, Van de Walle MJ, Laipis PJ, and Hauswirth WW, 1983. Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. Nature 306:400–402.

Rieder MJ, Taylor SL, Tobe VO, and Nickerson DA, 1998. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. Nucleic Acids Res 26:967–973.

Ruvolo M, Zehr S, von Dornum M, Pan D, Chang B, and Lin J, 1993. Mitochondrial COII sequences and modern human origins. Mol Biol Evol 10:1115–1135.

Saitou N and Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425.

Sarich VM and Wilson AC, 1973. Generation time and genomic evolution in primates. Science 179:1144–1147.

Strimmer K and von Haesseler A, 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969.

Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, and Schatten G, 1999. Ubiquitin tag for sperm mitochondria. Nature 402:371–372.

Tamura K and Nei M, 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.

Templeton AR, 1992. Human origins and analysis of mitochondrial DNA sequences. Science 255:737.

Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B, and Scozzari R, 1998. mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62:1137–1152.

Torroni A, Richards M, Macaulay V, Forster P, Villems R, Norby S, Savontaus ML, Huoponen K, Scozzari R, and Bandelt HJ, 2000. mtDNA haplogroups and frequency patterns in Europe. Am J Hum Genet 66:1173–1177.

Vigilant L, Pennington R, Harpending H, Kocher TD, and Wilson AC, 1989. Mitochondrial DNA sequences in single hairs from a southern African population. Proc Natl Acad Sci USA 86:9350–9354.

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

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

Wallace DC, Brown MD, and Lott MT, 1999. Mitochondrial DNA variation in human evolution and disease. Gene 238:211–230.

Wolpoff MH, 1989. Multiregional evolution: the fossil alternative to Eden. In: The human revolution: behavioural and biological perspectives on the origins of modern humans (Mellars P and Stringer C, eds). Princeton, NJ: Princeton University Press; 62–108.

Wolpoff MH, Hawks J, Frayer DW, and Hunley K, 2001. Modern human ancestry at the peripheries: a test of the replacement theory. Science 291:293–297.

Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd K, Modiano D, Scozzari R, Stoneking M, Tishkoff S, Batzer M, and Labuda D, 1998. Genetic structure of the ancestral population of modern humans. J Mol Evol 47:146–155.

Received October 29, 2001

Corresponding Editor: Oliver A. Ryder