Am. J. Hum. Genet. 61:1413–1423, 1997

Diversity and Age of the Four Major mtDNA Haplogroups, and Their Implications for the Peopling of the New World

Sandro L. Bonatto and Francisco M. Salzano

Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

Summary

Despite considerable investigation, two main questions on the origin of Native Americans remain the topic of intense debate—namely, the number and time of the migration(s) into the Americas. Using the 720 available Amerindian mtDNA control-region sequences, we reanalyzed the nucleotide diversity found within each of the four major mtDNA haplogroups (A–D) thought to have been present in the colonization of the New World. We first verified whether the within-haplogroup sequence diversity could be used as a measure of the haplogroup's age. The pattern of shared polymorphism, the mismatch distribution, the phylogenetic trees, the value of Tajima's *D*, and the computer simulations all suggested that the four haplogroups underwent a bottleneck followed by a large population expansion. The four haplogroup diversities were very similar to each other, offering a strong support for their single origin. They suggested that the beginning of the Native Americans' ancestral-population differentiation occurred ~30,000–40,000 years before the present (ybp), with a 95%-confidence-interval lower bound of ~25,000 ybp. These values are in good agreement with the New World–settlement model that we have presented elsewhere, extending the results initially found for haplogroup A to the three other major groups of mtDNA sequences found in the Americas. These results put the peopling of the Americas clearly in an early, pre-Clovis time frame.

Received March 4, 1997; accepted for publication October 1, 1997; electronically published November 26, 1997.

Address for correspondence and reprints: Dr. Sandro L. Bonatto, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Caixa Postal 15053, 91501-970 Porto Alegre, RS, Brazil. E-mail: sandro@if1.if.ufrgs.br

0002-9297/97/6106-0027$02.00

Introduction

The question of the origin of the indigenous peoples of the Americas has been the object of great debate. Some major problems have been slowly resolved since the last century, the main agreement achieved so far being that concerning these peoples' origin by migrations from Asia through the region of the Bering Strait $\geqslant$12,000 years before the present (ybp) (Cavalli-Sforza et al. 1994). However, the time and number of such migrations, as well as the size of the ancestral populations, are still important unsettled questions.

Previous studies based on high-resolution RFLPs and control region (CR) sequences have shown that the great majority of the Native American mtDNAs screened so far could be classified into four distinct clusters, called haplogroups "A"–"D" (for a review of the RFLP data, see Wallace 1995; for the CR sequence data, see Forster et al. 1996). The distribution of these four haplogroups in the populations that spoke the three main sets of languages found in the Americas (Amerind, Na-Dene, and Eskaleut), as well as the estimates of the internal diversity of each haplogroup, led to several hypotheses regarding the number and age of the migration(s) that colonized the New World. Although Amerind populations in general have all four haplogroups, Na-Dene and Eskimo groups have mainly sequences from haplogroup A (Merriwether et al. 1995; Wallace 1995). Moreover, using RFLP data, Torroni et al. (1992, 1993, 1994) found a much lower haplogroup A sequence diversity in the Na-Dene than in the Amerinds, whereas, within Amerinds, haplogroup B had sequence diversity lower than those of the other three haplogroups. These results led these authors to suggest that the Na-Dene entered the continent by means of an independent migration (Wallace 1995). The Amerinds, on the other hand, would have migrated to the Americas in two waves; the more ancient carried haplogroups A, C, and D, whereas the more recent carried haplogroup B sequences only. On the basis of the mean diversity found in the haplogroups, Wallace and co-workers dated the major Amerind migration into the Americas as having occurred ~26,000–34,000 ybp, the haplogroup B migration as having occurred ~12,000–15,000 ybp, and the Na-Dene

migration as having occurred ~7,000–9,000 ybp (Wallace 1995). In contrast, Horai et al. (1993), using CR sequence data, postulated that each major haplogroup would represent separate migrations that occurred ~14,000–21,000 ybp.

We have recently shown (Bonatto and Salzano 1997), on the other hand, that mtDNA sequence data strongly support a single and early (>20,000 ybp) origin for the Amerinds, Na-Dene, and Eskimo, in agreement with other molecular studies (Merriwether et al. 1995; Forster et al. 1996; Kolman et al. 1996). Our results were based mainly on the analysis of the CR sequences from haplogroup A, since it is the only haplogroup widely distributed among all Native Americans (Merriwether et al. 1995). It remained unknown whether the other three major Native American haplogroups would indicate the same picture. Also, although most studies on the problem of dating the colonization of the Americas have used sequence diversity as a measure of age, few (e.g., Bonatto and Salzano 1997) have investigated whether their samples met the very stringent assumptions required by this practice (Rogers and Jorde 1995).

The goals of the present study were to analyze the four major Native American mtDNA haplogroups' diversity and to evaluate the implications of these results for the estimation of the number and age of the migration(s) to the Americas. Specifically, we tested the hypothesis that one or more of the haplogroups may represent different migrations to the continent (e.g., Horai et al. 1993; Wallace 1995). It should be noted that here we will not test the hypothesis of different migrations that is based on linguistic groups (Amerinds, Na-Dene, and Eskimo), since this has been done elsewhere (Bonatto and Salzano 1997). The analyses involved the use of two data sets, one including 720 Native Americans, with their hypervariable segment I (HVS-I) sequences, and another composed of 217 individuals, with their HVS-I+HVS-II sequences. Several methods were applied, including computer simulations over a wide range of demographic scenarios, to determine whether the data were consistent with a bottleneck followed by a large population expansion. We finally calculated the within-haplogroup nucleotide-diversity values, and, using appropriate substitution rates and methods, estimated their mean ages and 95% confidence interval (CI) values.

Subjects and Methods

Population Samples

All available CR sequences from Native Americans were employed, with the exception of the two populations described by Horai et al. (1993), since they were not sequenced for the first 100 bases of HVS-I. Also, only some sequences from Easton et al.'s (1996) Yanomami were used, since many of them present several unusual features that preclude their utilization until they are further investigated (authors' unpublished data). The other sequences have the complete set of nucleotides for HVS-I (positions 16024–16383 [numbering is according to Anderson et al. 1981]) or HVS-I+HVS-II (positions 45–390 for HVS-II), or only a small fraction of them are missing. Two data sets were assembled, one with HVS-I sequences and the other with HVS-I+HVS-II sequences. The sequences were aligned by hand, and insertions in relation to the reference sequence (Anderson et al. 1981) were not considered. For the HVS-I, the Native American sample consists of 720 individuals from a total of 24 populations (with sample sizes $n \geqslant 5$) from North, Central, and South America, for each continent, as follows: for South America (n = 318)—Xavante (n = 25), Zoró (n = 30), and Gavião (n = 27) (Ward et al. 1996); Wai Wai (n = 26) and Suruí (n = 24) (authors' unpublished data); Mapuche (n = 39) (Ginther et al. 1993); Yanomama (n = 27), Wayampi (n = 21), Kayapo (n = 13), Arara (n = 9), Katuena (n = 9), Poturujara (n = 9), Awa-Guaja (n = 2), and Tiriyo (n = 2) (Santos et al. 1996); Yanomami (n = 50) (Easton et al. 1996); and Colombian mummies (n = 5) (Monsalve et al. 1996); for Central America (n = 136)—Huetar (n = 27) (Santos et al. 1994); Ngöbé (n = 46) (Kolman et al. 1995); and Kuna (n = 63) (Batista et al. 1995); and, for North America (n = 228)—Nuu-Chah-Nulth (n = 63) (Ward et al. 1991); Bella Coola (n = 40) and Haida (n = 41) (Ward et al. 1993); and Yakima (n = 42), Athapascan (n = 21), Inupiaq Eskimo (n = 5), and western Greenland Eskimo (n = 16) (Shields et al. 1993). The 38 individuals whose mtDNA Torroni et al. (1993) have sequenced from several populations all over the Americas were also included.

For HVS-I+HVS-II, sequences were available from a total of 217 individuals from the Huetar (Santos et al. 1994), Ngöbé (Kolman et al. 1995), Mapuche (Ginther et al. 1993), and Yanomami (Easton et al. 1996) and from 24 Suruí, 26 Wai Wai, 3 Xavante, 1 Gavião, and 1 Zoró (authors' unpublished data).

Phylogenetic Analysis

Several DNA distances were used in the tree constructions, from the simplest (proportion of differences) to the most complex (Tamura-Nei gamma [Tamura and Nei 1993]), but all gave essentially the same results; therefore, only those with the Kimura two-parameter (K2P [Kimura 1980]) distance were presented. Because of the large number of sequences used, trees were constructed with the neighbor-joining (NJ) method, by use of the Njboot program (N. Takezaki; available at Internet address http://iubio.bio.indiana.edu). The inte-

rior-branch-test confidence probability (CP) values for branches in the trees (Rzhetsky and Nei 1992) were estimated by the CheckSzDv program (from the TreePack package [I. Belyi; http://trantor.cse.psu.edu/~belyi]), by means of the pairwise option and the K2P distance. Minimum-spanning trees (Excoffier et al. 1992) were also constructed, by means of the Minspnet program (L. Excoffier; ftp://acasun1.unige.ch/pub/comp/win).

Diversity and Divergence Estimates

The nucleotide diversity within and between haplogroups was calculated by means of the Sendbs program (N. Takezaki; http://iubio.bio.indiana.edu). Several DNA distances were calculated, and the standard error (SE) values of these estimates were obtained with a bootstrap approach with 1,000 replications over sites. The 95% CI for the diversity and divergence values were calculated by use of ± 2 SE. The 95% CI for the time of origin (expansion) of the haplogroups, on the basis of the nucleotide diversity values, was estimated as described by Bonatto and Salzano (1997), by use of their formula 1 (modified from Redd et al. 1995) for the calculation of the minimum SE of the time (T_{MSE}) and by use of ± 2 T_{MSE} for the lower- and upper-bound values. We should note that our 95% CI considers both the nucleotide diversity and the mutation-rate errors.

For the time estimates, we need the substitution rates for the HVS-I and HVS-I+HVS-II regions, as well as their SEs. For HVS-I, we used the slow and fast rates given by Bonatto and Salzano (1997): 10.3% ($\pm 1.35\%$)/million years (Myr) and 15% ($\pm 1.97\%$)/Myr. For the HVS-I+HVS-II data sets, we used the following two rates: 8.85% ($\pm 0.9\%$)/Myr and 11.5% ($\pm 1.15\%$)/Myr. Both slow rates were taken from Horai et al. (1995), and the fast rates were taken from Ward et al. (1991) (in the case of HVS-I) and Stoneking et al. (1992) (in the case of HVS-I+HVS-II), whereas the SEs were those either given by Horai et al. (1995) or estimated by use of their approach.

The α parameter for our data sets was calculated by means of Yang and Kumar's (1996) method and the Pamp program (from the Paml package [Z. Yang; http://iubio.bio.indiana.edu]), by use of trees calculated by use of the K2P distance.

Mismatch Distributions

The evolutionary history of the four haplogroups was also examined, by use of the mismatch-distribution approach (Rogers and Harpending 1992; Rogers 1995; Rogers and Jorde 1995). The relevant parameters were calculated by means of the method of moments (Rogers 1995), by the Mmest program (from the Mismatch package [A. Rogers; ftp://anthro.utah.edu]). The 95% CI for the times of expansion of each haplogroup was estimated in a manner similar to that used in the nucleotide-diversity approach presented above, as described elsewhere (Bonatto and Salzano 1997).

Simulations

Rogers and Jorde (1995) showed that the only sense in which sequence diversity can be constructed as a measure of age is as an estimation of the time during which a population has expanded since a severe bottleneck. Although it is clear that the peopling of the Americas was probably characterized by a population reduction followed by expansion (Bonatto and Salzano 1997), there is considerable uncertainty about the sizes of the founding and the more recent pre-Columbian populations (Cavalli-Sforza et al. 1994). Besides, we are dealing here with groups of sequences (haplogroups), not with distinct populations. Thus, we want to test also the assumptions of a small founding population and a large expansion, for each haplogroup independently, so that we may apply dating methods that use sequence diversity as a measure of age.

Following the work by Eller and Harpending (1996), we designed simulations of various situations of stationary and expanding populations, to test (1) in which conditions the empirical estimates would be reproduced and (2) whether we could reject or accept the hypothesis that the haplogroups were stationary or had expanded and, if the latter was true, to what degree. Specifically, we tested in which demographic scenarios the simulations would give values at least as extreme as the ones estimated directly from the samples, for two statistics—Harpending's raggedness (r; Harpending 1994) and Tajima's D (Tajima 1989). Raggedness quantifies the smoothness of a distribution: the smaller the value, the smoother the (mismatch) distribution. Harpending et al. (1993) found that expanding populations showed very small r values, since their mismatch distributions have the shape of a smooth wave. However, Aris-Brosou and Excoffier (1996) showed that a high heterogeneity of substitutions among the sites (a lower α parameter for the gamma distribution) may cause a stationary population to exhibit very smooth distributions. Therefore, in such cases the results of the simulations using the raggedness of a distribution may not readily distinguish between stationary and expanding scenarios. Moreover, they also showed that, although (large) population expansions shift Tajima's D to (significant) negative values, substitution-rate heterogeneity has the opposite effect, moving Tajima's D to more-positive values for more-uneven substitution rates. Since the mtDNA CR in humans is known to have substitution-rate heterogeneity (Kocher and Wilson 1991; Wakeley 1993), Tajima's D may be a better statistic to distinguish between station-

ary and expanding populations than is the r used by Eller and Harpending (1996) in their simulations.

The simulations were performed by the Mmgen program (from the Mismatch package; see above), which uses the coalescent model to generate simulated histories by assuming some input parameters, such as size of the mismatch distribution, number of sites, sample size, and time since the expansion (for details of the coalescent algorithm, see Rogers et al. 1996). All the above input parameters were calculated from the empirical data for each haplogroup, so that the simulations mirrored as closely as possible the actual demographic parameters for each haplogroup. The other parameters of interest are the degree of expansion of the population and its final size, in units of θ, where $\theta = 2N_f u$, with N_f denoting the number of females and with u denoting the per-generation mutation rate for the nucleotide region (see Rogers and Harpending 1992).

We modified the program so that it generated empirical distributions of 10,000 D and r values for each combination of final θ and degree of expansion. Final θ ranged from 0.1 to 1,000, and degree of expansion ranged from 1 (for a stationary population) to 100,000,000. Most of the simulations were performed considering only one final random mating population, but, to test whether geographic population structure could influence the results, some simulations were also generated considering that, after expansion, the population would split into 3 or 20 groups. Besides the model of "infinite sites," we also did simulations by using a mutation model that takes into account the mutation-rate heterogeneity in the mtDNA CR (Rogers et al. 1996), using the "finite sites with gamma-distributed rates" model of substitution (Rogers et al. 1996). The α parameters of the gamma distribution used in the simulations were those calculated for our data sets, as describe above. As in the work of Eller and Harpending (1996), a specific scenario of final θ and degree of expansion was rejected if $\leqslant 500$ ($\leqslant 5\%$) simulations showed a D or r value more extreme than that calculated from the data.

Results

Of the 720 Native American individuals sequenced for HVS-I, 592 (82%, comprising 125 different sequences) have sequences with all the markers for one of the four haplogroups (the marker substitutions for four major haplogroups are those listed by Forster et al. 1996 as the founding sequences A2, B, C, and D1). For HVS-I+HVS-II, this value is 161/217 (74%, comprising 52 different sequences). If we exclude Easton et al.'s (1996) Yanomami sample, these values are 87% for HVS-I and 89% for HVS-I+HVS-II, respectively. To minimize the possibility of the occurrence of multiple, yet closely related founding sequences in each haplogroup, which would result in overestimation of the diversity values since colonization, we used for each haplogroup only the sequences that have all its marker substitutions. By doing this we tried to ensure that all sequences analyzed here were derived from just one founding sequence per haplogroup. Also, we used for haplogroup A sequences from Amerinds, Na-Dene, and Eskimo, since we have shown elsewhere (Bonatto and Salzano 1997) that they all have a common origin. However, it is important to note that the diversity values for haplogroup A (see below) did not change much if we remove the non-Amerind sequences.

A striking feature of the two data sets is that, for each haplogroup, the polymorphisms, especially at HVS-I, either exist in only one sequence or are shared by a small fraction of the sequences, with no substitution occurring in >20% of the haplogroup sequences (not shown). According to Slatkin and Hudson (1991), this pattern is exactly what we would expect in a situation of exponential growth from a single ancestral sequence (from whom the descendant sequences inherited the marker substitutions).

Haplogroups' Nucleotide Diversity

One important requirement in the coalescence theory (Donnelly and Tavaré 1995) and the mismatch-distribution methods (Rogers and Jorde 1995) is the use of a random sample of genes from the population under study. The population for the problem in which we are interested—the peopling of the Americas—is the entirety of Native Americans, not the local groups. However, two individuals from the same local population will have a much higher probability of being closely related than will two individuals from different populations, especially if we consider the generally small sizes of the Native American local populations (Salzano and Callegari-Jacques 1988). The use of the within-local-population frequency of the sequences, highly affected by each population's specific recent demographic history, will underestimate the nucleotide diversity of Native Americans as a whole. On the other hand, the occurrence of the same sequence in different populations is more likely to have been affected by more-ancient events. Therefore, since we are interested in the early evolutionary history of the continent, for the estimation of the within-haplogroup nucleotide diversity we used the between-populations frequency of the sequences (also see Bonatto and Salzano 1997). To estimate the between-populations frequency of each different sequence in the data set, we counted only the number of populations in which it occurred, disregarding its frequency within the populations.

Table 1 shows the nucleotide diversities (with their

Table 1

Nucleotide Diversity and Age Estimates for Native American's mtDNA Haplogroups

Haplogroup	No. of Individuals	No. of Sequences	Nucleotide Diversity[a] (95% CI) (%)	Mean Age (95% CI) (years) 10.3%/Myr	15%/Myr
HVS-I:					
A	71	45	.84 (.72–.97)	41,014 (34,887–47,142)	28,163 (23,949–32,377)
B	45	31	.80 (.71–.89)	39,017 (36,240–44,569)	26,791 (22,972–30,611)
C	36	25	.84 (.68–.99)	40,680 (34,150–47,210)	27,933 (23,443–32,423)
D	36	24	.96 (.77–1.16)	46,778 (38,979–54,576)	32,121 (26,759–37,482)
Average[b]	188	125	.86 (.79–.93)[c]	41,576 (35,869–47,283)	28,549 (24,623–32,475)
Divergence[d]			2.67 (2.42–2.92)[c]	129,469 (111,458–147,480)	88,902 (76,513–101,292)
				8.85%/Myr	11.5%/Myr
HVS-I+HVS-II:					
A	17	17	.83 (.76–.89)	46,635 (41,504–51,765)	35,889 (31,940–39,837)
B	16	15	.71 (.62–.80)	40,144 (35,360–44,929)	30,894 (27,212–34,576)
C	11	10	.64 (.45–.82)	36,034 (29,615–42,453)	27,730 (22,790–32,670)
D	10	10	.83 (.74–.92)	46,999 (41,610–52,388)	36,169 (32,022–40,316)
Average[b]	54	52	.75 (.70–.81)[c]	42,620 (38,029–47,211)	32,799 (29,266–36,332)
Divergence[d]			2.03 (1.92–2.13)[c]	114,639 (102,604–126,674)	88,222 (78,961–97,484)

[a] Tamura-Nei gamma distance, for α values given in the text.

[b] Weighted by the number of sequences in each haplogroup.

[c] SEs for the weighted averages were calculated as the square root of the sum of the squared weighted SEs from the individual comparisons.

[d] Weighted average of the pairwise haplogroup divergence.

95% CIs) for the four haplogroups, for both HVS-I and HVS-I+HVS-II. The remarkable feature is the high similarity of the values in each data set, especially for HVS-I (with higher sample sizes), which have a range of 0.80%–0.96% with a mean of 0.86%. Note that, on the contrary, the studies with RFLPs (Torroni et al. 1994) found that haplogroup B had a much lower diversity than the other three. For the RFLP data the mean diversity value of the other three haplogroups is 2.2 times higher than the haplogroup B diversity, whereas this ratio for the CR data is only ~1.1. This ratio for the CR data is maintained even when the within-population frequency of the sequences is used, the sample size being 563 individuals in this case (not shown), which is higher than the 335 individuals used in the RFLP studies. Therefore, this difference between CR sequences and RFLP data cannot be either explained by sample size or attributed to the different ways in which the haplotype frequencies were treated, more probably being due to the different populations, regions of the mtDNA studied, or haplogroup definitions. Diversity values for the HVS-I+HVS-II data set were more variable, possibly because of smaller sample sizes and, perhaps, the higher mutation-rate heterogeneity in HVS-II (see below).

The between-haplogroup divergence values are much higher than the within-haplogroups diversity, and the average among all pairwise comparisons is more than three times (for the HVS-I) and more than two times (for the HVS-I+HVS-II) higher than the within-haplogroup averages (table 1). This result supports the notion that the haplogroups' divergence (not diversification) began well before their entering the Americas and that any analysis that lumps together different haplogroups (e.g., the mismatch distributions of Horai et al. 1993 and Aris-Brosou and Excoffier 1996) will furnish results from events that occurred much earlier than the colonization of the New World.

In light of both the existence of rate heterogeneity in

the mtDNA CR (Wakeley 1993) and its effects on the estimation of the true DNA distance (Yang 1996), the diversity values were calculated by means of the Tamura-Nei gamma distance. The α parameter used in the calculations was estimated the Yang and Kumar (1996) new method. For the Native American HVS-I and HSV-I+HVS-II data sets, the α values were .5 and .16, respectively, in close approximation with those calculated for similar data sets (Wakeley 1993; Yang and Kumar 1996). The lower number for the HVS-I+HVS-II is due to the extremely low value for the HVS-II region, which alone has an estimated α of .07.

Mismatch Distributions

A one-wave–shaped distribution of the number of nucleotide differences between all pairs of individuals within a population, the mismatch distribution, is a signature of a population expansion in the past (Rogers and Harpending 1992), and extensive simulations have corroborated this finding (Slatkin and Hudson 1991; Rogers and Harpending 1992; Harpending et al. 1993). Figure 1 shows the mismatch distributions for the Native American haplogroups, for both the HVS-I and HVS-I+HVS-II data sets. The haplogroups' wave profiles are remarkably similar to each other, suggesting that these sequences were taken from the same ancestral population, which underwent a large expansion in the past. The waves are not so similar for the HVS-I+HVS-II data, probably because of smaller sample sizes and the much higher mutation-rate heterogeneity in the HVS-II region. The raggedness values for the HVS-I (table 2) data set are very low, as is generally found in expanding populations (Harpending et al. 1993).

Several studies have shown that the phylogeny of a sample of genes taken from a population that has experienced a large expansion after a bottleneck has the shape of a star tree (DiRienzo and Wilson 1991; Slatkin and Hudson 1991; Rogers and Jorde 1995). Figure 2 shows the NJ tree of the 125 different HVS-I sequences from the four haplogroups; only two D sequences did not cluster with the others of their respective haplogroups. The statistical support for the haplogroups is high, all CP values being >85%, with the exception of haplogroup D, which does not have unique markers (see Forster et al. 1996). The most remarkable feature of this tree is that each haplogroup presents a clear star-shaped subtree, similarly to what was found with the use of the minimum-spanning tree (not shown). This result supports again the hypothesis of a large population expansion for Native Americans.

Simulation Results

Tajima's D values for the haplogroups were significantly negative for the HVS-I data set (table 2). Aris-Brosou and Excoffier (1996) demonstrated that a large (>100-fold) population expansion moves Tajima's D to significantly negative values but that mutation-rate heterogeneity shifts it to more positive values. Therefore, the significantly negative D values obtained for the four haplogroups when the HVS-I data are used, despite the existence of a moderate mutation-rate heterogeneity ($\alpha = .5$) in this region, is a strong support for a large expansion that affected all four haplogroups. However, the much higher mutation-rate heterogeneity ($\alpha = .16$) in the HVS-I+HVS-II data shifted the D values to numbers inside the 95% CI, although they still are moderately negative, as was found by Aris-Brosou and Excoffier (1996) in their simulations. The much lower α (.07) for the HVS-II region alone shifted Tajima's D for haplogroups B and C to positive values (not shown).

Figure 3 shows the results of the simulations for each haplogroup when Tajima's D statistics and the finite-sites gamma-rates model are used for the simulation. The darker-shaded values denote specific combinations of degree of expansion and value for final θ (population size) that could not be rejected by the simulations; that is, these are scenarios in which >500 (>5%) of the simulations resulted in a value of D lower than that estimated from our HVS-I data set. The minimum degree of expansion not rejected by the simulations was 100-fold for haplogroups A, B, and C and 50-fold for D. Table 2 presents the maximum size, of the initial population, that was not rejected by the simulations (the size of the initial θ was calculated by dividing the final θ by the degree of expansion for each combination of values that were not rejected, and the maximum value for each haplogroup was taken). From these values of initial θ the effective number of females was calculated by $\theta_0 = 2N_f u$, as given above. These results suggest that each Native American haplogroup, as defined here, was founded by a small number of females. Since the values above are the number of females in the founding population that carried each haplogroup-founding sequence, to estimate the size

Table 2

Summary Statistics for the Four Native American mtDNA CR Haplogroups, Based on the HVS-I Data Set

Haplogroup	Tajima's D	r	θ_0[a]	N_{f0}[b] 10.3%/Myr	N_{f0}[b] 15%/Myr
A	−2.363*	.0496	.25	135	93
B	−2.359*	.0414	.3	162	111
C	−2.226*	.0498	1.0	539	370
D	−2.060*	.0294	2.0	1,078	740

NOTE.—Number of individuals and number of sequences are as in table 1.

[a] Maximum value of θ not rejected by the simulations.

[b] Effective number of females in the initial population, calculated as $\theta_0/2u$.

*$P < .05$.

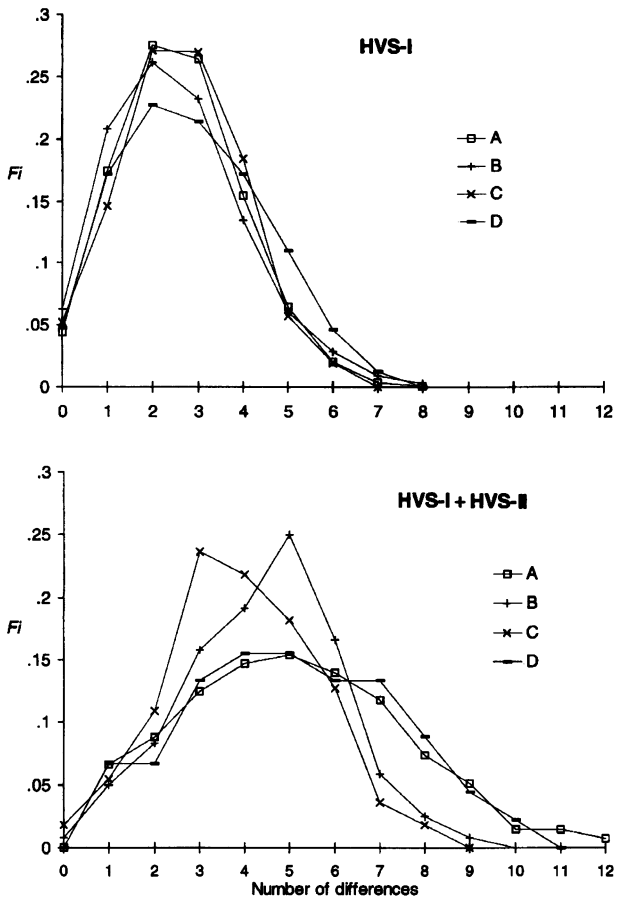

Figure 1 Mismatch distributions for the four haplogroups, with HVS-I and HVS-I+HVS-II data sets. *Fi* denotes the relative frequency of pairs of sequences that differ by *i* nucleotide sites.

of the whole founding population for the four haplogroups we should sum the individual values for each haplogroup, which results (when we use the 10.3% rate) in a maximum value of ~2,000 females and a founding population of <5,000 individuals. The possible existence of other, less successful founding haplogroups (e.g., see Forster et al. 1996; Merriwether and Ferrell 1996), which may account for ~10% of the mtDNA now found in the Americas, may increase these estimates to some degree. These figures, although approximate, suggest that during the colonization process the ancestral population was never much higher than ~10,000 individuals.

The use of the finite-sites model with gamma-distributed rates in the simulations—rather than the unrealistic, infinite-sites model (e.g., see Eller and Harpending 1996)—turned the tests more stringent in relation to the scenarios that could be rejected. Also, in distinguishing the stationary scenario from the expanding scenario, Tajima's *D* had a discriminating power much higher than that of the raggedness statistics, especially when a mutation-rate–heterogeneity model was used in the simulations (not shown). The use of final, geographically structured populations, instead of a randomly mating one, in the simulations had no qualitative effect on the tests.

Estimating the Age of the Four Native American mtDNA Haplogroups

All the results that we have presented so far strongly argue in favor of the hypothesis that, in the process of the colonization of the Americas, there was, for each haplogroup, a bottleneck, followed by a large (>100-fold) expansion. Therefore, we now have justification to use the sequence diversity found in each Native American haplogroup as a measure of the latter's expansion age. Table 1 shows the mean ages and their 95% CIs for each haplogroup, for both data sets and all mutation rates. The diversification times are very similar both to each other and between the HVS-I and HVS-I+HVS-II data sets. The average ages for the HVS-I values were ~42,000 and ~29,000 ybp, and those for the HVS-I+HVS-II values were ~43,000 and ~33,000 ybp, for the slower and faster substitution rates, respectively. The

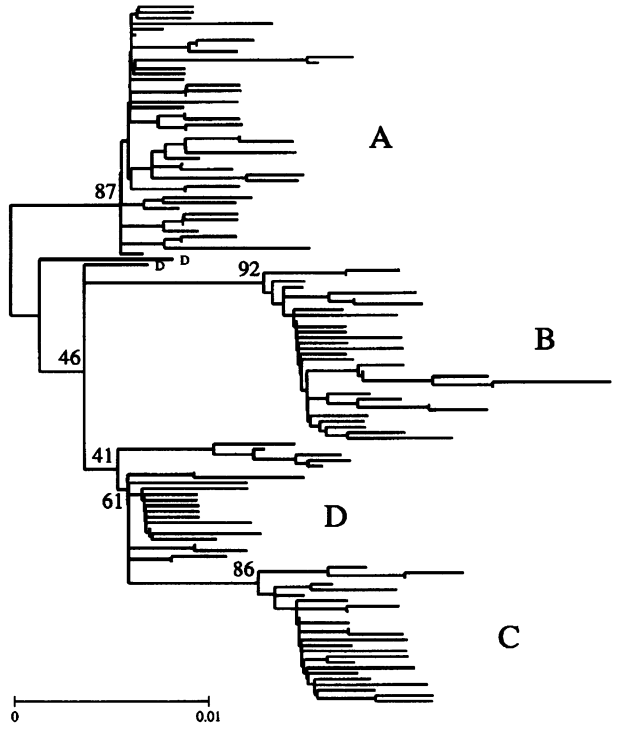

Figure 2 NJ tree of 125 different Native American HVS-I sequences from the four major haplogroups. All sequences clustered according to the haplogroup (A–D) to which they belong, except for two haplogroup D sequences. The interior-branch-test CP values for the main clusters are shown above the branches.

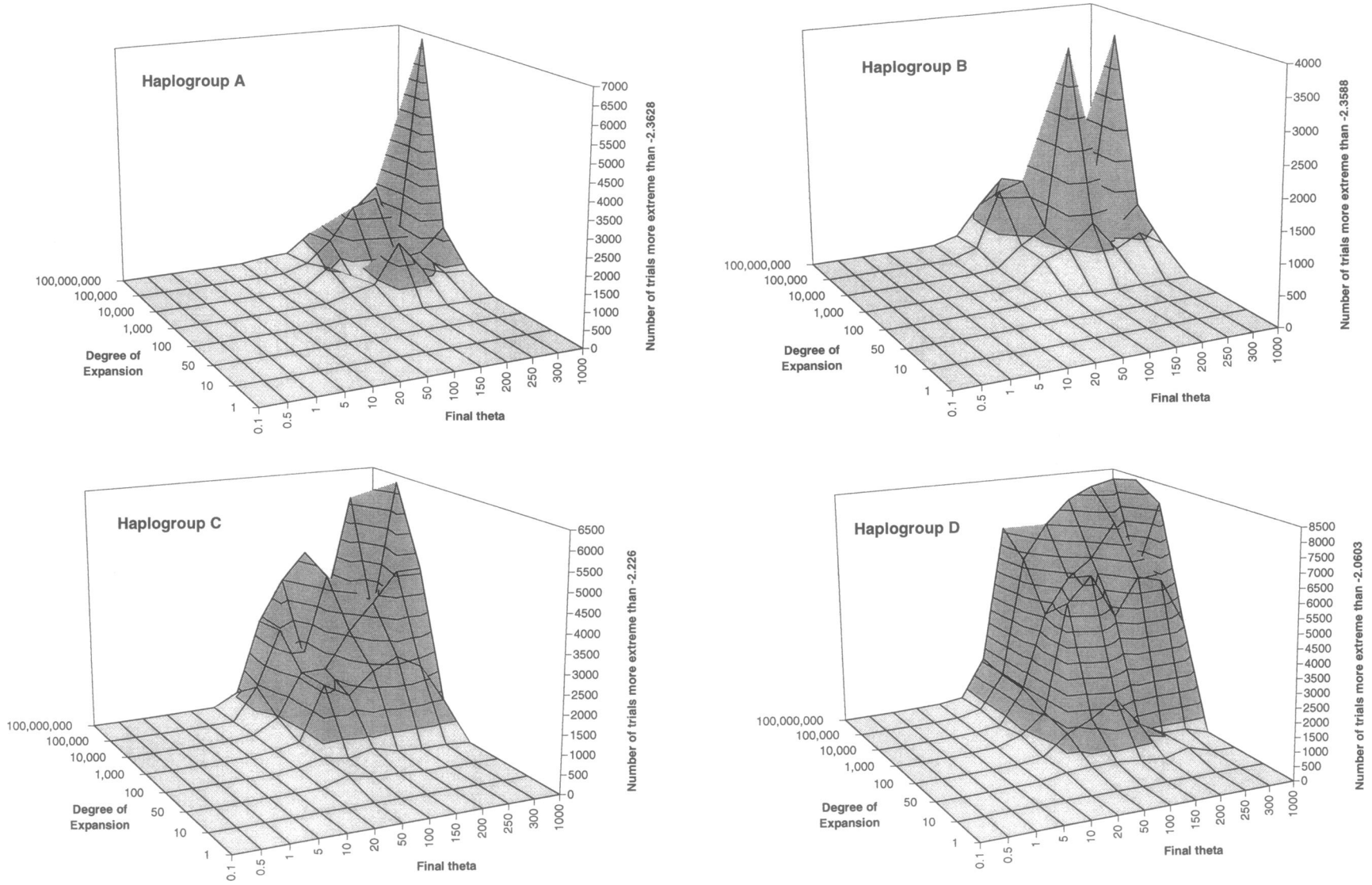

Figure 3 Simulation surfaces for the four haplogroups (A–D), by use of Tajima's *D*. The number of simulations for which Tajima's *D* were more extreme than the values calculated for each haplogroup are plotted for each value of final θ and degree of expansion. Darker shading denotes those models that could not be rejected at the 5% level, and lighter shading denotes those models that could be rejected at the 5% level.

lower-bound estimates for the averages were ~25,000 (HVS-I) and ~29,000 (HVS-I+HVS-II) ybp; and the minimum value for all estimates was ~23,000 ybp, for haplogroup C with use of the HVS-I+HVS-II data. The upper-bound age was ~50,000 ybp. The average divergence time between the haplogroups was >110,000 ybp, being, in general, three times higher than the haplogroup ages. The ages estimated by use of the mismatch-distribution approach, by means of the method of moments, were identical or very similar to those calculated by use of nucleotide diversity (not shown), further indicating a strong initial bottleneck (Bonatto and Salzano 1997), although the CIs of the mismatch-distribution method were larger.

Discussion

When we consider mainly the HVS-I data, the pattern of shared polymorphisms, the mismatch distribution (fig. 1), the phylogenetic tree (fig. 2), the values of Tajima's *D* and raggedness (table 2), and the simulation results (fig. 3), all suggest that the four major haplogroups underwent a bottleneck followed by a large population expansion. These results give strong support for our further use of the within-haplogroup sequence diversity as an estimate of the time since that bottleneck. The very similar diversity values found for the four haplogroups, both with the HVS-I data set and with the HVS-I+HVS-II data, strongly suggest that they all expanded at approximately the same time and, therefore, that they most likely came from the same population, a result that is in agreement with a single-migration model suggested by several recent studies (Merriwether et al. 1995; Forster et al. 1996; Kolman et al. 1996; Bonatto and Salzano 1997).

In our previous study (Bonatto and Salzano 1997), using mainly haplogroup A sequences, we concluded that those mtDNA data strongly indicate that all Native Americans originated from a single colonization event that occurred in Beringia >22,000 ybp ago, possibly ~30,000–40,000 ybp. We suggested a scenario, based on Szathmary's works (e.g., see Szathmary 1993), in which the Native American ancestral population settled in the Beringian landmass during sometime before expanding. Eventually they crossed the Alberta ice-free corridor and colonized the rest of the American continent. The collapse of that corridor, ~25,000–14,000 (Hoffecker et al. 1993) or ~30,000–11,000 (Lemmen et al. 1994) ybp, isolated the people still living in Beringia, from whom originated the Na-Dene and Eskimos (with their reduced overall mtDNA diversity); those south of the ice sheets gave rise to the Amerind-speaking peoples. The present results for the four major haplogroups' diversification ages agree very well with these estimates. When only the mean values are considered, these estimates suggest a very early date (~30,000–40,000 ybp) for the beginning of the diversification of the Native American ancestral population, with a lower bound of ~25,000 ybp.

At least two types of evidence support the idea that haplogroups' sequence differentiation probably began during Beringia's settlement and not in Asia before the colonization process: (1) our estimates of $\geqslant$100-fold ancient population expansion suggest that the diversification began during an intensive colonization process; and (2) if the expansion had occurred somewhere else in Asia, then one should find there sequences, with all markers for each haplogroup, at a high number and frequency, similar to the ~90% frequency found in Native Americans; however, only the founding sequences for each haplogroup have been found in Asia so far—and they have been found at a very low frequency (see Forster et al. 1996; Kolman et al. 1996; Bonatto and Salzano 1997). The few additional founding sequences for haplogroup A that have been suggested—in the Na-Dene and Eskimo (see Forster et al. 1996)—are probably derived ones and will be discussed elsewhere (authors' unpublished data).

We agree that some additional founding haplogroups (such as group X from Forster et al. 1996; also see Bailliet et al. 1994; Merriwether and Ferrell 1996) might exist, besides the four major ones studied here. However, they constitute only ~10% of the sequences now found in the Americas and, because of their very small sample size, could not be analyzed in the study. Since we analyzed each haplogroup separately, and since the number of haplogroups was not a relevant parameter, including these putatively additional founding haplogroups should not significantly change the results presented here.

Some recent studies also tried to estimate the time of entry into the Americas by means of haplogroup-diversity values, on the basis of both RFLP data (Torroni et al. 1992, 1994) and CR sequence variation (Forster et al. 1996). We emphasize that our inferred CIs took into account both the mutation-rate heterogeneity and the nucleotide-diversity variance, whereas the estimates of other recent studies considered, at most, only one source of error; the CI for their estimates would be much broader than the range that they have provided. As for Torroni et al.'s (1992, 1994) hypothesis, our previous results do not support the idea of an independent Na-Dene migration (Bonatto and Salzano 1997), and our present analyses also do not support their suggestion of a more recent haplogroup B migration. Similarly, neither Horai et al.'s (1993) proposal of different migrations, ~14,000–21,000 ybp, for each haplogroup nor the hypothesis of a Polynesian contribution for haplogroup B sequences found in America (see Bonatto et al. 1996) was supported. In any case, Torroni et al.'s (1994) estimated average arrival date, ~26,000–34,000 ybp, for

the other three haplogroups is very close to our estimates (table 1).

In general, Forster et al.'s (1996) scenario for the peopling of the Americas is similar to that which we proposed (see above and Bonatto and Salzano 1997). They postulated a single and early entry (>20,000 ybp) and suggested that, although the Amerinds colonized all the continent and maintained their original diversity, Beringians (Eskimo + Na-Dene) reduced their diversity, because of the climate's deterioration until ~11,000 ybp, at which time they reexpanded to their present size. Forster et al. also have presented coalescence ages for Native American haplogroups, using a data set very similar to our HVS-I—but very different methods—to estimate the haplogroups' age. Although they did not calculate any CI for their age estimates, they suggested ~20,000–25,000 ybp as the arrival time for the Amerinds, which is near our lower-bound estimates. Their haplogroup coalescence ages, however, are probably underestimates of the diversification times since these populations' entrance in the Americas, since they estimated the diversity values on the basis of each haplogroup within each tribe separately. Their results would receive a strong influence from the recent demographic history of each tribe, which could significantly change the ancient parameters that we are interested to estimate. A good example of this can be seen in their estimated age for the Central American Amerinds, which showed a coalescence age lower than that of the South Americans. Far from suggesting that Central American Amerinds originated more recently than South American Amerinds, this result only reflects the reduced mtDNA diversity found in the Chibcha groups, from which all Central American mtDNA sequences came. The Chibcha's reduced mtDNA diversity is thought to have occurred because of recent events (Kolman et al. 1995).

Acknowledgments

We thank to Mark Stoneking for helpful comments on an earlier version of this manuscript. This work was funded by Financiadora de Estudos e Projetos, Conselho Nacional de Desenvolvimento Científico e Tecnológico, and Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior.

References

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

Aris-Brosou D, Excoffier L (1996) The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol Biol Evol 13:494–504

Bailliet G, Rothhammer F, Carnese FR, Bravi CM, Bianchi NO (1994) Founder mitochondrial haplotypes in Amerindian populations. Am J Hum Genet 55:27–33

Batista O, Kolman CJ, Bermingham E (1995) Mitochondrial DNA diversity in the Kuna Amerinds of Panamá. Hum Mol Genet 4:921–929

Bonatto SL, Redd AJ, Salzano FM, Stoneking M (1996) Lack of ancient Polynesian-Amerindian contact. Am J Hum Genet 59:253–256

Bonatto SL, Salzano FM (1997) A single and early origin for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci USA 94:1866–1871

Cavalli-Sforza LL, Piazza A, Menozzi P (1994) History and geography of human genes. Princeton University Press, Princeton

DiRienzo A, Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci USA 88:1597–1601

Donnelly P, Tavaré S (1995) Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421

Easton RD, Merriwether DA, Crews DE, Ferrell RE (1996) mtDNA variation in the Yanomami: evidence for additional New World founding lineages. Am J Hum Genet 59: 213–225

Eller E, Harpending H (1996) Simulations show that neither population expansion nor populations stationarity in a west African population can be rejected. Mol Biol Evol 13: 1155–1157

Excoffier L, Smouse P, Quattro J (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA data. Genetics 131:479–491

Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935–945

Ginther C, Corach D, Penacino GA, Rey JA, Carnese FR, Hutz MH, Anderson A, et al (1993) Genetic variation among the Mapuche Indians from the Patagonian region of Argentina: mitochondrial DNA sequence variation and allele frequencies of several nuclear genes. In: Penna SDJ, Chakraborty R, Epplen JT, Jeffreys AJ (eds) DNA fingerprinting: state of the science. Birkhäuser, Basel, pp 211–219

Harpending HC (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol 66:591–600

Harpending HC, Sherry ST, Rogers A, Stoneking M (1993) The genetic structure of ancient human populations. Curr Anthropol 34:483–496

Hoffecker JF, Powers WR, Goebel T (1993) The colonization of Beringia and the peopling of the New World. Science 259: 46–53

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

Horai S, Kondo R, Nakagawa-Hattori Y, Hayashi S, Sonoda S, Tajima K (1993) Peopling of the Americas, founded by four major lineages of mitochondrial DNA. Mol Biol Evol 10:23–47

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region

and protein-coding region. In: Osawa S, Honjo T (eds) Evolution of life: fossils, molecules and culture. Springer, Tokyo, pp 391–413

Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, Guionneau-Sinclair F (1995) Reduced mtDNA diversity in the Ngöbé Amerinds of Panamá. Genetics 140:275–283

Kolman CJ, Sambuughin N, Bermingham E (1996) Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. Genetics 142:1321–1334

Lemmen DS, Duk-Rodkin A, Bednarski JM (1994) Late glacial drainage systems along the northwestern margin of the Laurentide ice sheet. Q Sci Rev 13:805–828

Merriwether DA, Ferrell RE (1996) The four founding lineage hypothesis for the New World: a critical reevaluation. Mol Phylogenet Evol 5:241–246

Merriwether DA, Rothhammer F, Ferrell RE (1995) Distribution of the four founding lineage haplotypes in Native Americans suggests a single wave of migration for the New World. Am J Phys Anthropol 98:411–430

Monsalve MV, Cardenas F, Guhl F, Delaney AD, Devine DV (1996) Phylogenetic analysis of mtDNA lineages in South American mummies. Ann Hum Genet 60:293–303

Redd AJ, Takezaki N, Sherry ST, McGarvery ST, Sofro ASM, Stoneking M (1995) Evolutionary history of the COII/ $tRNA^{Lys}$ intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol Biol Evol 12:604–615

Rogers A (1995) Genetic evidence for a Pleistocene population explosion. Evolution 49:608–615

Rogers A, Fraley AE, Bamshad MJ, Watkins WS, Jorde LB (1996) Mitochondrial mismatch analysis is insensitive to the mutation process. Mol Biol Evol 13:895–902

Rogers A, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569

Rogers A, Jorde L (1995) Genetic evidence on modern human origins. Hum Biol 67:1–36

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol 9: 945–967

Salzano FM, Callegari-Jacques SM (1988) South American Indians: a case study in evolution. Clarendon Press, Oxford

Santos M, Ward RH, Barrantes R (1994) mtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. Hum Biol 66:963–977

Santos SEB, Ribeiro-dos-Santos AKCR, Meyer D, Zago MA (1996) Multiple founder haplotypes of mitochondrial DNA in Amerindians revealed by RFLP and sequences. Ann Hum Genet 60:305–319

Shields GF, Schmiechen AM, Frazier BL, Redd A, Voevoda MI, Reed JK, Ward RH (1993) mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern North American populations. Am J Hum Genet 53:549–562

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequence in stable and exponentially growing populations. Genetics 129:555–562

Stoneking M, Sherry ST, Redd AJ, Vigilant L (1992) New approaches to dating suggest a recent age for the human mtDNA ancestor. Philos Trans R Soc Lond B 337:167–175

Szathmary EJE (1993) Genetics of aboriginal North Americans. Evol Anthropol 1:202–220

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595

Tamura K, 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

Torroni A, Neel JV, Barrantes R, Schurr TG, Wallace DC (1994) Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. Proc Natl Acad Sci USA 91:1158–1162

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, et al (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53:563–590

Torroni A, Schurr TG, Yang C-C, Szathmary EJE, Williams RC, Schanfield MS, Troup GA, et al (1992) Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. Genetics 130:153–162

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

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

Ward RH, Frazier BL, Dew-Jager K, Pääbo S (1991) Extensive mitochondrial diversity within a single Amerindian tribe. Proc Natl Acad Sci USA 88:8720–8724

Ward RH, Redd A, Valencia D, Frazier B, Pääbo S (1993) Genetic and linguistic differentiation in the Americas. Proc Natl Acad Sci USA 90:10663–10667

Ward RH, Salzano FM, Bonatto SL, Hutz MH, Coimbra CEA Jr, Santos RV (1996) Mitochondrial DNA polymorphism in three Brazilian Indian tribes. Am J Hum Biol 8:317–323

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. TREE 11:367–372

Yang Z, Kumar S (1996) Approximate methods for estimating the patterns of nucleotide substitution and the variation of substitution rates among sites. Mol Biol Evol 13:650–659