Am. J. Hum. Genet. 59:935–945, 1996

Origin and Evolution of Native American mtDNA Variation: A Reappraisal

Peter Forster,[1] Rosalind Harding,[3] Antonio Torroni,[4] and Hans-Jürgen Bandelt[2]

[1]Heinrich-Pette-Institut für Experimentelle Virologie und Immunologie an der Universität Hamburg and [2]Mathematisches Seminar, Universität Hamburg, Hamburg; [3]Institute of Molecular Medicine, John Radcliffe Hospital, Oxford; and [4]Dipartimento di Genetica e Biologia Molecolare, Università di Roma "La Sapienza," Rome

Summary

The timing and number of prehistoric migrations involved in the settlement of the American continent is subject to intense debate. Here, we reanalyze Native American control region mtDNA data and demonstrate that only an appropriate phylogenetic analysis accompanied by an appreciation of demographic factors allows us to discern different migrations and to estimate their ages. Reappraising 574 mtDNA control region sequences from aboriginal Siberians and Native Americans, we confirm in agreement with linguistic, archaeological and climatic evidence that (i) the major wave of migration brought one population, ancestral to the Amerinds, from northeastern Siberia to America 20,000–25,000 years ago and (ii) a rapid expansion of a Beringian source population took place at the end of the Younger Dryas glacial phase ~11,300 years ago, ancestral to present Eskimo and Na-Dene populations.

Introduction

The number of prehistoric migrations into the American continent and their timing has not yet been conclusively resolved. Archaeological estimates for the first human entry to America vary from 15,000 years (Morell 1990) to >33,000 years ago (Dillehay and Collins 1988), and a controversial linguistic classification has been taken to postulate that three different migrations from Asia across Beringia account for the presence of (*a*) Amerind speakers throughout North, Central, and South America, (*b*) Na-Dene speakers in the Pacific northwest, and (*c*) Eskimo-Aleut speakers in the circumpolar region (Greenberg et al. 1986).

These questions have been addressed by studying mtDNA variation by using sequencing as well as restriction (RFLP) analysis, but a consensus has not yet emerged. In particular, no one has combined the extensive data of different mtDNA control region sequencing studies. The first sequencing studies (Ward et al. 1991, 1993) yielded an unexpectedly high level of diversity in Amerinds, complicating the relationship between ethnic groups and mitochondrial types. Subsequent work (Horai et al. 1993; Torroni et al. 1993*a*, 1993*b*) explained this diversity by showing that a small but diverse set of founding Asian sequences entered America. Shields et al. (1993) demonstrated the common and relatively recent ancestry of Siberian Beringians and American Eskimo and Na-Dene mtDNA sequences. Ivanova (1993) and Voevoda et al. (1994) further found that only coastal Siberian mtDNA is closely related to American Eskimo and Na-Dene mtDNA. The lowest mtDNA diversity among Amerinds was observed in Central Americans (Santos et al. 1994; Batista et al. 1995; Kolman et al. 1995), and these researchers proposed ancient demographic events rather than decimation after European conquest as an explanation. Recently, Merriwether et al. (1995) have enlivened the debate on American origins by arguing that the ubiquity of some haplogroups in America implies a single migration to America.

Here we argue that any time estimate of a migration must be preceded by a precise phylogenetic identification of founding sequences. Therefore, we reevaluate 574 Native American and Siberian mtDNA sequences from the first hypervariable region by using an intraspecific network method. We identify the major founding Native American control region sequences, attaining a higher resolution than has been possible with RFLP analysis. The resulting phylogenies reflect a common founding mitochondrial pool for Native American tribes as well as distinct demographic histories.

In order to determine the arrival time of each founding control region sequence in America, we then search for a time estimator that is appropriate for the demographic situations displayed by the American phylogenies. We argue and demonstrate by computer simulations that the average distance ρ to the founding sequence is a simple but reliable time estimator. We calibrate ρ by use

Received April 30, 1996; accepted for publication July 17, 1996.

Address for correspondence and reprints: Dr. Hans-Jürgen Bandelt, Mathematisches Seminar, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany.

0002-9297/96/5904-0026$02.00

of phylogenetic, archaeological, and paleoclimatological information on Eskimo and Na-Dene evolution and suggest an entry date of 20,000–25,000 years ago for the first Americans and a reexpansion of Beringians ~11,300 years ago.

Subjects and Methods

Population Samples

The phylogenetically analyzed samples comprise unambiguous mtDNA control region sequences from nucleotide positions (np) 16090–16365 for all sequences except those of Horai and Hayasaka (1990) and Horai et al. (1993), which begin at np 16129 (nucleotide positions according to Anderson et al. 1981). Native American and Siberian sequences were taken from the following publications: (1) Ward et al. (1991), (2) Ward et al. (1993), (3) Shields et al. (1993), (4) Ivanova (1993) and Voevoda et al. (1994), (5) Torroni et al. (1993*a*, 1993*b*), (6) Ginther et al. (1993), (7) Horai et al. (1993), (8) Santos et al. (1994), (9) Batista et al. (1995), and (10) Kolman et al. (1995). Geographic locations, sample sizes *n*, and references for the populations are as follows: *Siberians:* Chukchi (northeast Siberia, $n = 7 + 35$, references 3 and 4), Siberian Eskimo (northeast Siberia, $n = 6 + 22$, references 3 and 4), other Siberians ($n = 6 + 7 + 16$, references 3, 4, and 5); *American Eskaleut speakers:* Inupiaq (Alaska, $n = 5$, reference 3), Inuit (West Greenland, $n = 17$, reference 3); *Na-Dene speakers:* Athapascan (Alaska, $n = 21$, reference 3), Haida (British Columbia, $n = 41$, reference 2), other Na-Dene speakers ($n = 7$, reference 5); and *Amerind speakers:* Bella Coola (British Columbia, $n = 40$, reference 2), Nuu-Chah-Nulth (Vancouver Island, $n = 63$, reference 1), Yakima (Washington State, $n = 42$, reference 3), Huetar (Costa Rica, $n = 27$, reference 8), Ngöbé (Western Panama, $n = 46$, reference 10), Kuna (Eastern Panama, $n = 63$, reference 9), native Columbians ($n = 20$, reference 7), native Chileans ($n = 45$, reference 7), and Mapuche (Patagonia, $n = 38$, reference 6).

Three of these 574 sequences were discarded for the analyses because they represent obvious cases of admixture: one Chilean, already identified as an outlier by Horai et al. (1993), differs by only one unique mutation from a common European/Middle Eastern sequence, and a Haida, and a West Greenland Inuit (35 and 83 in Shields et al. 1993) are not found in any other American or Siberian sample but reveal two exact matches with two of the most-common European sequences. It is assumed that sequence 66 in the data of Shields et al. (1993) is baseshifted at np 16107/16111 (cf. the guidelines of Bandelt et al. 1995). Furthermore, 210 non-Siberian Asian sequences were used for comparisons. These were compiled from the data of Vigilant (1990), Horai and Hayasaka (1990), Shields et al. (1993), Torroni et al. (1993*a*), and Mountain et al. (1995).

Phylogenetic Analysis

Intraspecies data, in contrast to interspecies data, are typically characterized by low divergence, resulting in phylogenies with multifurcations and unresolvable parallelisms. Reduced median networks (Bandelt et al. 1995), which we use here, reproduce multifurcations and resolve parallelisms and reversals (i.e., discard improbable trees) with a compatibility argument when justified by character-state frequency criteria, whereas unresolvable parallelisms and reversals are highlighted as reticulations, thus displaying multiple solutions in a single diagram. A reduced median network will typically include several or all most-parsimonious trees.

Demographic Analysis

Time estimates from intraspecific genetic diversity call for a population-genetic approach to allow for the effects of genetic drift on sequence divergence. The extent of genetic drift is a consequence of the demographic history of that population (Donnelly and Tavaré 1995). Demographic factors that may have been important during the evolution of a population are the rate and pattern of population growth and extent of population structure (nonrandom mating) within and among tribes. In population-genetic studies, a widely used statistic for estimating time is π, referred to as the nucleotide diversity when expressed per nucleotide, or alternatively referred to as the average pairwise sequence difference when expressed per sequence stretch (Tajima 1989*a*, 1989*b*). The statistic π is, however, not appropriate for populations where constant size and random mating cannot be assumed.

Qualitatively, the star-shaped topology of the Native American mtDNA tree suggests a population expansion for most tribes. In an attempt to confirm or reject the assumption of constant size from the structure of the gene tree for Native American mtDNA, we refer to simulation studies of coalescent models that show how gene tree topology varies with demographic history (Griffiths and Tavaré 1994*a*, 1994*b*; Donnelly and Tavaré 1995; Harding, in press). Tajima (1989*a*) proposed that the violation of the assumption of constant size can be investigated by comparing the sample diversity statistics π and S. Under the assumptions of constant size, random mating and neutrality, π and S, calculated from mtDNA variation of a tribal unit, estimate the parameter $\theta = 2N_f\mu$ where N_f is the effective number of females and μ is the mutation rate. The mean pairwise sequence diversity π averages over the number of nucleotide differences in all pairwise comparisons of sequences in a sample drawn randomly from a population. S is the number of segregating sites (i.e., positions variable in

the sample sequences) adjusted to the sample size n of sequences by division with the harmonic number for n − 1:

$$1 + 1/2 + \ldots + 1/(n-1) \approx \ln(n-1) + 0.5772 + 1/(2(n-1)) - 1/(12(n-1)^2) .$$

We compute these statistics π and S, but not from the raw sample diversity. In contrast to general usage, we correct the mismatch differences for multiple mutation hits by scoring from the estimated tree. In the case of equally probable trees, π is averaged across all potential solution trees. S, likewise, is corrected for multiple hits. Values of S significantly greater than π are a signature of population expansion.

We are further concerned that the phylogenies of Native American mtDNA reveal not a randomly mating population but an almost completely subdivided population of different tribes sharing only the set of ancestral founding haplotypes. In the following, we examine the effect of this scenario on π, S, and on a third statistic, ρ, which is the average number of sites differing between a set of sequences and a specified common ancestor (which need not be among the sampled sequences). Computing ρ is thus based on the topology of a rooted tree. The statistics ρ and π can be calculated for the pooled population by summing over the values π_i and ρ_i (ρ_i referring to the last common ancestor of the pooled population) for each subpopulation with sample sizes n_i ($n = n_1 + \ldots + n_m$ is the sample size for the total population):

$$\rho = \sum_{i=1}^{m} (n_i/n)\rho_i$$

and

$$\pi = 1/[n(n-1)] \sum_{i=1}^{m} [n_i(n_i - 1)\pi_i + 2n_i(n - n_i)\rho_i] .$$

For $m \to \infty$ one has $\pi \to 2\rho$, while all n_i are bounded by some constant. If n is sufficiently large compared to m, and $n_i \approx n/m$ with $m \geq 2$, then the total number of segregating sites in the combined sample is approximately $S \ln n \approx \theta m(\ln n - \ln m)$, assuming constant population size, and therefore $S \approx m\theta$ while π and ρ can be given approximately as $\pi \approx (2 - 1/m)\theta$ and $\rho \approx \theta$.

In summary, we expect in the case of complete subdivision, $S > \pi \geq 3/2\rho$ (constant size, but complete subdivision) provided that ρ is small. In this situation, S is almost independent of the total sample size n. With high exponential growth within each tribal subpopulation, the loss of new mutations due to genetic drift is reduced, the number of segregating sites grows faster than the logarithm of the sample size n. There is an expectation of $S > \pi \geq 3/2\rho$ (considerable growth), as above, but now S grows with sample size n.

We examine the relationship between S, π, and ρ by simulating a coalescent model with an exponential decline backwards in time of the population size at rate 30 from the current size, N_0, such that N_t at times in the past is given by $N_0 e^{-t30}$ and N_0 is determined by a choice of $\theta = 10$ (Griffiths and Tavaré 1994*b*; Harding, in press). With $\theta = 10$, and $\mu = 0.001$ per generation, N_0 is 5,000 females. Simulations were run for samples of 20 and 50 sequences in each of 20 subpopulations. We found, on average for $n = 20$, S = 1.59, $\pi = 1.05$, and $\rho = 0.65$, and for $n = 50$, S = 2.34, $\pi = 1.05$, and $\rho = 0.63$. Not unexpectedly, these statistics varied widely among subpopulations. However, these simulations demonstrate that the relationship $S > \pi \geq 3/2\rho$ holds for a model of complete subdivision with independent expansions founding from a shared ancestral population.

Estimates of Time

Simple population genetic estimates of coalescence time assuming constant population size are not appropriate in application to the Native American mtDNA data. Because population expansion reduces the effect of drift, an assumption of sequence divergence accumulating as a function of time only may suffice. In phylogenetic studies, the time-depth of a lineage in a given tree is typically estimated by ρ, defined above. For the simulation examples, the average times back to the common ancestral sequence are 13,249 years for $n = 20$ and 12,897 years for $n = 50$. Estimates of these times given by ρ are 13,000 years for $n = 20$ and 12,600 years for $n = 50$, demonstrating that ρ is a suitable estimator when such a model applies.

Averaged Diversity

In the presence of more than one founder sequence and data from different tribes, we propose the following procedure to obtain meaningful diversity values. In order to avoid an elevation of S compared to π, or of π compared to ρ, incurred by deliberate pooling of tribes and haplogroups, we calculate S and π for each subhaplogroup (descended from a single founder) within each tribe separately and then take the averages (weighted according to subsample sizes). Since the subsample sizes here are usually <20, we would expect no larger values than $S \leq 1.5 \leq 2.4\rho$ (generalizing from the above simulations) even under considerable growth. At the extreme, subsample sizes smaller than four would necessarily contribute equal values for S and π and are

therefore excluded here from the calculation of averages. In contrast, the parameter ρ is not influenced by pooling.

Results

Criteria for Founding Sequences among Native Americans

In order to identify the major founding control region sequences among Native American mtDNA, we employed the three criteria with which Torroni et al. (1993*a*) justified their RFLP founding haplotypes: (*a*) American founding sequences should still be found in Asians, (*b*) founding sequences should occupy central branching nodes in a phylogeny, and (*c*) founding sequences should be widespread within Native Americans, whereas derived sequences should have a limited distribution. It should be noted, however, that these criteria could miss minor founding haplotypes and would furthermore only apply to major founding events that are recent relative to the mutation rate and to source and descendent populations that have expanded sufficiently to retain ancestral sequences. The two latter assumptions appear to be reasonable in the case of the settlement of America, which represented an expansion into a large, uninhabited area probably <30,000 years ago. Another caveat concerns extremely variable nucleotide positions, which can easily arise in the Asian and American populations independently from the same root sequence, thus creating an additional sequence match that could be mistaken for a founder sequence. This case seems to occur because of frequent transitions at np 16519 (cf. the data of Horai et al. 1993), which causes restriction site variation at 16517 *Hae*III (fortunately outside the stretch of control region considered here): The extreme variability of this site is well documented in American and African mtDNA (cf. Torroni et al. 1993*a*, 1993*b*; Chen et al. 1995); curiously, this site appears to be "frozen" in haplogroup B, for some unknown reason. The decision taken by Bailliet et al. (1994) and Bianchi et al. (1995) to identify founding haplotypes according to 16517 *Hae*III status is therefore phylogenetically questionable, and we do not recommend the use of their haplogroups or their nomenclature.

Matches with Asian Sequences

To implement criterion (*a*), we compared the 472 Native American sequences with 309 Asian sequences, revealing that several Native American sequences are identical to Asian sequences. For this search the stretch from np 16129 to np 16362 was compared so as to include the data of Horai and Hayasaka (1990) and Horai et al. (1993), yielding 10 Asian-American matches. Subsequent inspection of the region np 16090 to np 16129 of these 10 sequences revealed that one match could be further subdivided on the basis of a variant nucleotide at np 16111. The 11 candidate founder sequences are marked in figure 1 as large circles. Most Eskimo/Na-Dene sequences cluster around only one match, as will be shown, in contrast to Amerinds who harbor several clusters. We therefore discuss these populations separately.

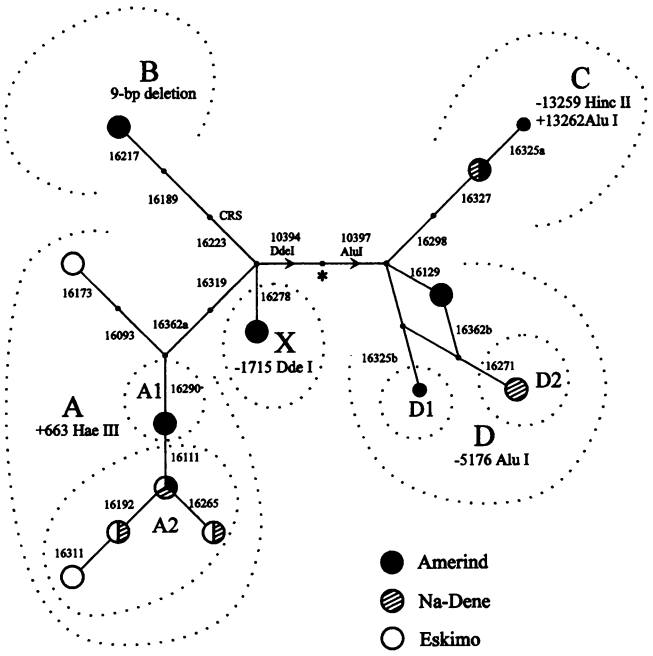

Figure 1 Reduced median network of mtDNA control region matches between Native Americans and Asians, indicated by large circles carrying the distinction between Amerind, Na-Dene, and Eskimo. Two sequences represented by medium-sized black circles are added, since they clearly meet criteria (*b*) and (*c*) for founding sequences. RFLP haplogroups A, B, C, and D, along with their characteristic sites, are denoted as by Torroni et al. (1993*a*, 1993*b*), whereas control region haplogroup X was classified as "Others" in those RFLP studies. Control region sequencing permits further subdivision of A and D. The nomenclature used here is not related to that of Bailliet et al. (1994). The Cambridge reference sequence (Anderson et al. 1981) is indicated by CRS, and the putative root of the network (cf. Torroni et al. 1993*a*) is marked by an asterisk (*). Links are labeled by mutations in the control region or restriction site changes; parallel lines in a reticulation represent the same nucleotide position; the order of mutations on a path not interrupted by any branchings or distinguished nodes is arbitrary; suffixes *a* and *b* indicate recurrent mutations; arrows point to restriction site gains. Putative Amerind control region founding sequences (listed as transitions relative to the Anderson sequence less 16,000) are A1 (223 290 319 362), A2 (111 223 290 319 362), B (189 217), C (223 298 325 327), D1 (223 325 362), and X (223 278).

Founding Sequences among Amerinds

To investigate whether any of these matches represent central branching nodes in Amerind mtDNA (criterion *b*), each Amerind sequence was assigned to 1 of the

11 nodes according to its sequence motif. The resulting reduced median networks of the four largest groups (containing all most parsimonious trees) are displayed in figure 2, while a fifth group (which we shall call "X") has previously been displayed by Bandelt et al. (1995) in their figure 7. The most obvious feature of the phylogenies in figure 2 is their perfectly starlike structure, but they also show that the branching centers of C and D1 are not found in this Asian sample, indicating that they have acquired mutations before or during the transition from Asia to America. These additional mutations are entered as links to two new potential founding nodes indicated by medium-sized black circles in figure 1. One of these additional potential American founders incidentally matches a partially sequenced Japanese (Oota et al. 1995). We then completed figure 1 by adding the RFLP status and RFLP haplogroup labels A, B, C, D, and X to the nodes according to the nomenclature of Torroni et al. (1993*a*). The restriction sites 10394 *Dde*I and 10397 *Alu*I separated the transition at np 16362 into two parallel events. Two of the sequence matches in figure 1 (Asian-Eskimo and Asian-Na-Dene) have not been analyzed for their RFLP status and have been preliminarily assigned to haplogroups A and D, pending restriction analysis. Further sampling may decide whether these two matches are also possible founding sequences. As can be seen in figure 1, there is a one-to-one relation with the traditional RFLP haplogroups, but the higher resolution of control region sequences resolves RFLP group A into A1 and A2, D into D1 and D2, while X (formerly assigned to "Others") is seen to constitute an Amerind haplogroup in its own right.

According to criterion (*c*), founding sequences and their descendant haplogroups should be expected to be widespread within America. The geographic frequencies of all haplogroups observed in the Amerinds are summarized in table 1, which shows this criterion is indeed met, although there are considerable frequency variations in different geographic areas. Haplogroup X is rather infrequent in the populations presented here, and further samples are needed to clarify its precise phylogeny and diversity. However, the antiquity and geographic range of this haplogroup is underlined by its identification in a 7,000-year-old brain from Florida (Hauswirth et al. 1994). The third criterion further demands that derived sequences should have a limited geographic distribution. The phylogenies presented in figure 2 indeed strongly reflect the ethnogeographic affinities of the tribes, since the deeper a node is, the more widespread it is.

In summary, we have evidence for at least six Amerind founding sequences which gave rise to the groups A1, A2, B, C, D1, and X. Remarkably, only 4 of the 384 Amerind sequences do not fit into these haplogroups in a trivial manner (one Chilean sequence is European [see Methods]; one Nuu-Chah-Nulth and one Huetar lack informative variants for unambiguous group assignment, and one Amerind sequence match, preliminarily assigned to haplogroup D, requires additional data as discussed), suggesting that no major Amerind founding sequence has been overlooked. Internal nodes are generally shared between tribes, evidently because these nodes predate tribal differentiation. Peripheral nodes are rarely shared between tribes so that intertribal female gene flow can be roughly estimated by counting the number of individuals sharing peripheral nodes with other tribes, which amounts to only 9% (10/115) of individuals in peripheral nodes. The ubiquity of the haplogroups within Amerind tribes (table 1) cannot therefore be explained by intertribal migration and suggests that the Amerind founding sequences arrived as one wave from Asia.

Founding Sequences among Eskimos/Na-Dene

American Eskimo and Na-Dene mtDNA in our sample is predominantly A2 (90%), a feature that distinguishes them from Amerinds. Figure 3 represents a reduced median network of the combined Siberian/Eskimo/Na-Dene sequences with the A2 motif. The network offers many most parsimonious trees, which are generated mainly by reversions at np 16362 (a feature not found in the other American data sets), but consideration of the sequence frequencies leaves only few plausible evolutionary pathways; and if, in addition, np 16362 is regarded as unstable then essentially only one phylogenetic tree is plausible (indicated in the figure by unbroken lines). This phylogeny is strongly supported by the linguistic and geographic hierarchies evident in the network: the three deepest American nodes of the A2 haplogroup are shared with Siberians, whereas the derivatives of these nodes are virtually specific to either Siberia or America. These three nodes therefore appear ancestral to populations on both continents (criteria *b* and *c*). A fourth, peripheral, node, characterized by a transition at np 16311, is also shared between an Asian and an American, but since this site is hypervariable (see next section), it may well represent a parallelism. In summary, Eskimo and Na-Dene mtDNA was founded by at least three sequences, all within group A2. In addition, an Apache harbors a D2 root sequence also found in Siberia. Disregarding two European sequences (see Subjects and Methods), only 8/89 American Eskimo/Na-Dene sequences belong to other haplogroups (three Haida and one Athapascan are C, two Haida are D1, two Eskimos belong to the node preliminarily assigned to A in fig. 1), suggesting no major Eskimo/Na-Dene founders have been overlooked in our sample. The starlike structure and low diversity of haplogroup A2 in figure 3 indicates a recent reexpansion, supporting the

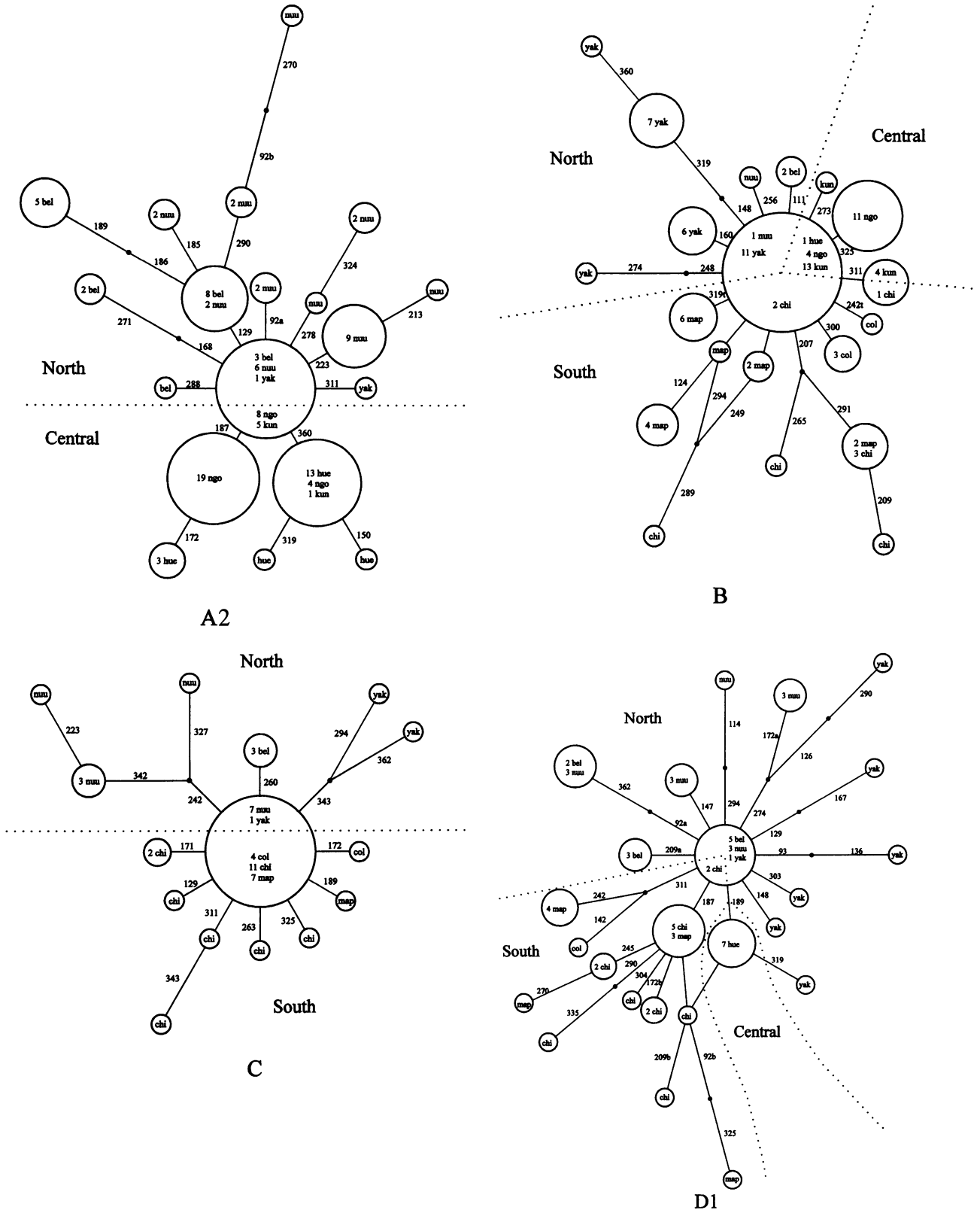

Figure 2 Reduced median networks of the four largest Amerind haplogroups A2, B, C, and D1. The Chileans and Colombians are not sequenced at np 16111, precluding their assignment to haplogroup A1 or A2. Nucleotide positions are as in Anderson et al. (1981) less 16,000; all mutations are transitions except those with a suffix *t*. Circle areas are proportional to sequence frequencies, and the exact number and tribal identity of individuals is inscribed inside the circles. Population labels are as follows: bel = Bella Coola; nuu = Nuu-Chah-Nulth; yak = Yakima; hue = Huetar; ngo = Ngobe; kun = Kuna; col = Colombian; chi = Chilean; and map = Mapuche. For total number of sequences from North/Central/South America, consult Subjects and Methods.

Table 1

Number of Individuals in the Major Control Region Founding Haplogroups in Amerind Tribes

Region and Amerind Tribe	Mitochondrial Haplogroup[a]					
	A1	A2	B	C	D1	X
North America:						
Bella Coola	0	19	2	3	10	0
Nuu-Chah-Nulth	0	28	2	12	13	7
Yakima	0	2	26	3	7	2
Central America:						
Huetar	0	18	1	0	7	0
Ngöbé	0	31	15	0	0	0
Kuna	39	6	18	0	0	0
South America:						
Colombians		10[b]	4	5	1	0
Chileans		2[b]	9	18	15	0
Mapuche	6	0	15	8	9	0

[a] Eleven Amerind sequences are not listed here: six Bella Coola and one Yakima appear to be Na-Dene incursion (see fig. 3); one Chilean has a European sequence; and one Nuu-Chah-Nulth, one Yakima, and one Huetar require further information for cluster assignment (see text).

[b] The Chileans and Colombians are not sequenced at np 16111, precluding their assignment to haplogroup A1 or A2.

interpretation of Shields et al. (1993) and Torroni et al. (1993*b*) that Eskimos and Na-Dene underwent an expansion distinct from and more recent than the expansion of Amerinds.

In order to address the question of whether the ancestral Eskimo and Na-Dene mtDNA pool was in interior Siberia, Beringia, or North America, we can compare the frequency of the deepest A2 sequence in each population as well as the diversity accumulated by the three founders on each side of the Bering straits. The combined data sets of Shields et al. (1993), Ward et al. (1993), and Ivanova (1993) and Voevoda et al. (1994) indicate that the deepest A2 founder sequence is found in 30/68 Na-Dene speakers but is absent in both American and Siberian Eskimos (0/49) and rare in Siberian Chukchi (2/44). This result agrees with the lower diversity of the derivatives in Siberia (0.42 mutations average distance from the A2 founders, as opposed to 0.63 for American Eskimo/Na-Dene) and with the coastal distribution of A2 in Siberia (Ivanova 1993; Voevoda et al. 1994). Phylogeny, diversity, and geographic distribution therefore indicate that A2 reexpanded in Beringia or in the former unglaciated regions of northern Northwest America, giving rise to both Na-Dene and Eskimo mtDNA.

Time Depth of Circumpolar Mitochondrial Diversity

In this section, we argue that the diversity of haplogroup A2 in Eskimos and Na-Dene has accumulated since the end of the Younger Dryas glacial relapse. Circumpolar populations from both Siberia and North America are known to exhibit a markedly reduced mtDNA diversity compared to their respective neighbors further south, as is exemplified by the lack of haplogroup B in northern populations on both continents, by the dominance of haplogroup A2 in Eskimos and Na-Dene, and by the reduced diversity within their A2 haplogroup compared to that of neighboring Amerind populations (Shields et al. 1993; Torroni et al. 1993*b*). This situation is mirrored in Europe, where the major North German–specific haplogroup, which is also starlike, has the same, low diversity as Eskimo and Na-Dene haplogroup A2, while the general European diversity is three times greater (Richards et al. 1995). One explanation for finding similarly lowered levels of mtDNA diversity on the northern extremes of three different continents would be a phase of climatic deterioration drastic enough to extinguish northern populations. The most recent such event is the abrupt, worldwide return of full glacial conditions during the Younger Dryas 13,000–11,300 years ago, which was immediately followed by the present, exceptionally stable warm period. This scenario is explicitly borne out by archaeological findings that demonstrate a population hiatus in Alaska coincident with the Younger Dryas event (Kunz and Reanier 1994). We therefore assign the archaeologically and paleoclimatologically determined age of 11,300 ± 500 years to the mtDNA variation accumulated in the founding Eskimo and Na-Dene sequences. A later large-scale migration to Beringia would appear less plausible, be-

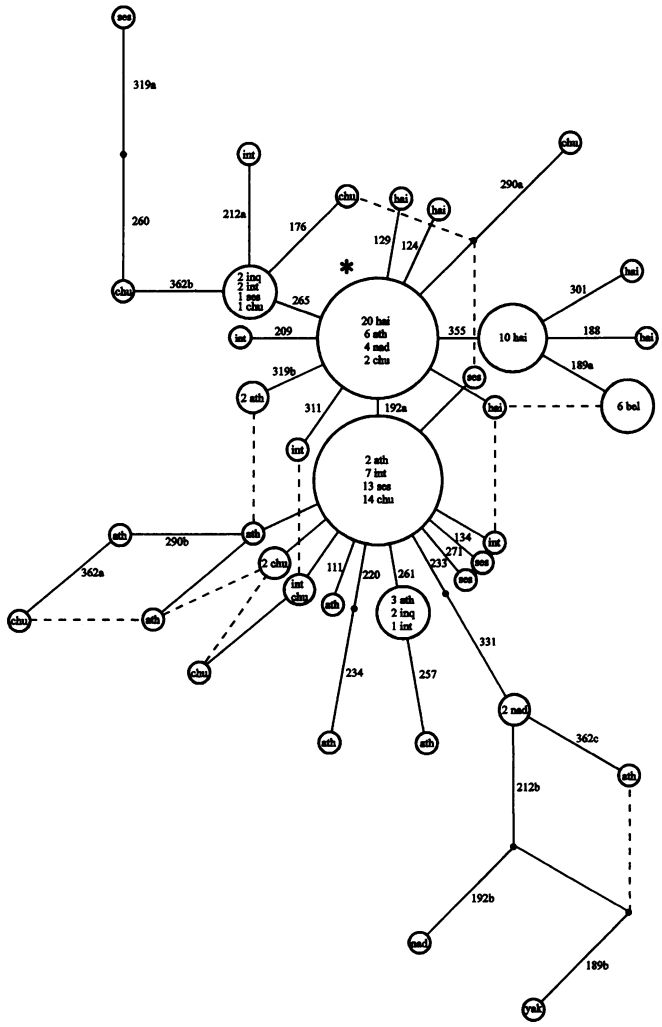

Figure 3 Reduced median network of control region mtDNA sequences belonging to haplogroup A2 in Siberians, Eskimos, and Na-Dene. Nucleotide positions are as in Anderson et al. (1981) less 16,000. Circle areas are proportional to sequence frequencies, and the exact numbers and tribal identities of individuals are inscribed inside the circles. Population labels are as follows: chu = Chukchi; ses = Siberian Eskimo; int = Inuit; inq = Inupiaq; ath = Athapascan; hai = Haida; and nad = other Na-Dene. Seven Amerind sequences (six Bella Coola and one Yakima) are also entered, since they presumably represent admixture. The ancestral sequence of A2 is labeled by an asterisk (*).

cause of not only the existence of a North European cluster of similar diversity and the inundation of the Bering straits after the ice age (Peltier 1994) but also the very low level of mitochondrial gene flow between even adjacent Native American tribes (9%, see Founding Sequences among Amerinds).

To take into account that not one but three major sequences founded Eskimo and Na-Dene haplogroup A2, we disregard the mutations at the two links separating these founders and calculate S, π, and ρ for each of the three subgroups in Eskimo and in Na-Dene separately and then take the weighted averages. For S, π, and ρ we obtain 2.33, 0.93, and 0.42, respectively, for Eskimo A2, and 2.17, 1.16, and 0.69, respectively, for Na-Dene A2. The diversity thus contracted strongly suggests considerable population growth for Eskimos and Na-Dene. Therefore, we use ρ as an appropriate measure for time depth (table 2), yielding a rate of one transition per 20,180 $\pm$ 1,000 years in the mtDNA control region (np 16090–16365). In order to check whether hypervariable sites distort the time estimate by mutating much faster than others, we can inspect the pooled transition events of the four displayed haplogroups (fig. 2). Only two sites, np 16189 and 16311, attain the approximate saturation level of four transitions (in agreement with Hasegawa et al. 1993), and their omission hardly affects the time estimates.

Previous calibrations of the mitochondrial clock have yielded similar values. Our mutation rate agrees roughly with those of Ward et al. (1991) and Hasegawa and Horai (1991), although their calibrations are based on uncertainties concerning the time depth of human and chimpanzee divergences, the similarity of mutation rates in humans and chimps, and the transition-transversion ratio. Our haplogroup time depths also correspond to those of Torroni et al. (1994), whose RFLP-based calibration draws on a postulated genetic split and isolation of the Chibchan tribes of Central America, although their age estimate for haplogroup B is too low, mainly because of the large proportion of Central Amerind sequences in their data. The fast mutation rate proposed by Lundstrom et al. (1992), who applied coalescence theory to the Nuu-Chah-Nulth data, is not valid because (*a*) the Asian coalescent of the Nuu-Chah-Nulth was inadvertantly used to estimate divergence time within the Americas, (*b*) the constant size model was chosen, and (*c*) an evolutionary average value for effective population size was estimated from a modern population census. In order to test our mutation rate (which is based on the assumption of population continuity in Beringia since the Younger Dryas), we in the following compare it with published radiocarbon dates.

Time Depth of Amerind Mitochondrial Diversity

The star-shape of mtDNA phylogenies and the relationship between *S*, π, and ρ (averaged over the tribal haplogroups to yield the contracted diversity) for North and South Amerind variation suggest that tribal populations have greatly expanded from an initially small number of founders. Table 2 demonstrates that ρ (the average distance to the coalescent of interest) is a more reliable time measure than π (the mean pairwise distance), in agreement with computer simulations of expanding populations (see Methods). In North and South

Table 2

Diversities and Coalescence Age Estimates for Native American Haplogroups

Population	*n*	*S*	π	ρ	ρ/μ (coalescence time, years)
Siberians/Eskimos/Na-Dene					
A2 averaged[a]	124	2.25	1.00	.56	11,300[b]
North American Amerinds:					
A2	49	2.22	1.81	1.24	
B	30	1.57	1.35	.93	
C	18	1.32	1.23	1.00	
D1	30	2.38	2.09	1.17	
North averaged	127	2.02	1.72	1.12	23,000
Central American Amerinds:					
A1	39	.24	.05	.97	
A2	55	.81	.71	.85	
B	34	.46	.45	.47	
D1	7	.00	.00	1.00	
Central averaged	135	.52	.42	.80	16,000
South American Amerinds:[c]					
A	18	.99	1.09	.84	
B	28	1.86	1.74	1.26	
C	31	1.31	.69	.36	
D1	25	2.83	2.11	1.85	
South averaged	102	1.78	1.40	1.05	21,000

NOTE.—Diversities averaged over tribes after correction for multiple hits and elimination of transversions (cf. Methods).

[a] Within Siberian/Eskimo/Na-Dene A2, three founders were assumed for the diversity calculations (cf. Results).

[b] Age assignment according to archaeological and paleoclimatological data (cf. Results).

[c] The shorter sequences of Horai et al. (1990, 1993) were taken into account by incrementing diversity values by 12%.

Amerind populations, S is always greater than π, and π is $\sim 3/2\rho$, which is a probable outcome of a recently expanded population. In North and South Amerind populations, ρ averages 1.12 and 1.05, yielding 23,000 and 21,000 years, respectively, for the migration of the Amerind ancestors from Siberia to America. Central Amerind populations show a different picture, with ρ even exceeding π, which may indicate a long phase of constant population size and/or population collapse. This interpretation is compatible with the absence of several ancestral sequences and of haplogroup C in Central Amerinds (fig. 2 and table 1).

The North and South Amerind dates coincide well with the earliest radiocarbon dates of 24,000 and 23,600 years for human artifacts in Alaska (reviewed by Szathmary 1993). The frequent occurrence of founding sequences even in South America implies a rapid advance south, which can be dated thanks to a South American–specific subcluster of D1, characterized by a transition at np 16187. Larger samples are needed, but its diversity (*n*, S, π, and ρ are 18, 2.5, 1.7, and 0.9, respectively) suggests a founding event in South America $\sim$18,000 years ago, in qualitative agreement with the earliest secure South American radiocarbon date of 13,000 years (Szathmary 1993).

Discussion

In conclusion, we tentatively propose the following scenario for the settlement of the Americas: As recent radiocarbon dating has shown (reviewed by Morell 1995), Siberia was inhabited by 40,000 years ago. The ancient Siberian mitochondrial gene pool would have included the ancestors of haplogroups A1 (itself ancestral to A2), B, C, D1, and X. These founding sequences appear to have entered America $\sim$20,000–25,000 years ago in one wave, as is indicated by the similar genetic dates for the different Amerind haplogroups, as well as by their widespread occurrence within Amerind tribes, despite low intertribal female migration. They reached South America by at least 13,000 years ago, according to radiocarbon dating (Dillehay and Collins 1988), which we confirm by genetic dating of the South American–specific subcluster of D1. Conditions during the Ice Age may then have reduced northern populations, whereas the ancestral mitochondrial variation is re-

tained at more southerly latitudes in both Asia (Kolman et al. 1996) and America (Shields et al. 1993). After the collapse of glacial conditions at the end of the Younger Dryas 11,300 years ago, Beringians reexpanded demographically and geographically and evolved into present-day Eskimo and Na-Dene peoples, while further south the climatic improvement may have stimulated the spread of the Clovis culture.

What can be learned from the preceding case of identifying and dating the migrational waves to America? Any mtDNA study that attempts to date the first appearance of some population in a certain geographical area should be based on extensive sampling, not only of the population under consideration but also of potential source populations and neighboring populations. A thorough phylogenetic analysis can then identify the putative founder sequences. The mutations distinguishing these founders must be disregarded in order to set the evolutionary clock to zero at the arrival of the population in the new area. Then, a careful choice of realistic demographic models or ad hoc approaches is necessary to describe the process adequately. A comparison of demographic parameters such as S, π, and ρ, averaged over the groups descended from the founders, and the shape of the phylogenetic tree (or network in the case of unresolvable ambiguities) may assist this search. Population genetic models assuming constant population size and random mating will hardly be realistic in most situations for the following reasons: (1) arrival in a new area is likely to trigger subsequent population expansion for many generations; (2) for a population on the move, the demographic pressure is possibly relaxed in that population sizes can fluctuate rather freely, that is, without statistically conforming to an expected (time-dependent) size; (3) over a period of tens of thousands of years, environmental conditions may have changed drastically, leaving an imprint on the effective population sizes; and (4) subsequent gene flow from neighboring populations will inevitably distort the estimation of θ and thus coalescence time. We have therefore resorted to ρ (the average distance to the root sequence of interest) to estimate time, but clearly there is a need for more sophisticated population genetic methods that allow variable population size and nonrandom mating to be taken into account.

Acknowledgments

We thank Martin Richards (Oxford) for critical advice; Laurent Excoffier (Geneva) for pointing out a problem in an earlier version of the manuscript; Boris Malyarchuk (Magadan) for access to Siberian data; and Holger Koschek (Hamburg), Arne Röhl (Hamburg), and Vincent Macauley (Oxford) for valuable computational assistance. This work was performed by P.F. in partial fulfillment of the requirements for the doctoral degree.

References

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

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

Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

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

Bianchi NO, Bailliet G, Bravi CM (1995) Peopling of the Americas as inferred through the analysis of mitochondrial DNA. Brazil J Genet 18:661–668

Chen Y-S, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (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

Dillehay TD, Collins MB (1988) Early cultural evidence from Monte Verde in Chile. Nature 332:150–152

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

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: Pena SDJ, Chakraborty R, Epplen JT, Jeffreys AJ (eds) DNA fingerprinting: state of the science, Birkhäuser, Basel, pp 211–219

Greenberg JH, Turner CG, Zegura LZ (1986) The settlement of the Americas: a comparison of the linguistic, dental and genetic evidence. Curr Anthropol 27:477–497

Griffiths RC, Tavaré S (1994*a*) Ancestral inference in population genetics. Stat Sci 9:307–319

——— (1994*b*) Sampling theory for neutral alleles in a varying environment. Phil Trans R Soc Lond B 344:403–410

Harding RM. Using the coalescent to interpret gene trees. In: Boyce AJ, Mascie-Taylor CGN (eds) Molecular biology and human diversity, Cambridge University Press, Cambridge, pp 63–80 (in press)

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

Hasegawa M, Horai S (1991) Time of the deepest root for polymorphism in human mitochondrial DNA. J Mol Evol 32:37–42

Hauswirth WW, Dickel CD, Rowold, DJ, Hauswirth, MA (1994) Inter- and intrapopulation studies of ancient humans. Experientia 50:585–591

Horai S, Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am J Hum Genet 46:828–842

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

Ivanova AV (1993) Mitochondrial DNA polymorphism in native inhabitants of Chukotka. PhD thesis, University of Novosibirsk

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

Kunz ML, Reanier RE (1994) Paleoindians in Beringia: evidence from Arctic Alaska. Science 263:660–662

Lundstrom R, Tavaré S, Ward RH (1992) Estimating substitution rates from molecular data using the coalescent. Proc Natl Acad Sci USA 89:5961–5965

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

Morell V (1990) Confusion in earliest America. Science 248: 439–441

——— (1995) Siberia: surprising home for early modern humans. Science 268:1279

Mountain JL, Hebert JM, Bhattacharyya PA, Underhill PA, Ottolenghi C, Gadgil M, Cavalli-Sforza LL (1995) Demographic history of India and mtDNA-sequence diversity. Am J Hum Genet 56:979–992

Oota H, Saitou N, Matsushita T, Ueda S (1995) A genetic study of 2,000-year-old human remains from Japan using mitochondrial DNA sequences. Am J Phys Anthropol 98: 133–145

Peltier WR (1994) Ice age palaeotopography. Science 265: 195–201

Richards MB, Forster P, Tetzner S, Hedges R, Sykes BC (1995) Mitochondrial DNA and the Frisians. In: Faltings V, Walker A, Wilts O (eds) North-Western European language evolution suppl vol 12, University of Odense, Odense, pp 141–163

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

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

Szathmary EJE (1993) Genetics of aboriginal north Americans. Evol Anthropol 6:202–220

Tajima F (1989*a*) The effect of change in population size on DNA polymorphism. Genetics 123:597–601

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

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*a*) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53:563–590

Torroni A, Sukernik RI, Schurr TG, Starikovskaya YB, Cabell MF, Crawford MH, Comuzzie AG, et al (1993*b*) mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet 53:591–608

Vigilant L (1990) Control region sequences from African populations and the evolution of human mitochondrial DNA. PhD thesis, University of California, Berkeley

Voevoda MI, Avksentyuk AV, Ivanova AV, Astakhova TI, Babenko VN, Kurilovich SA, Duffy LK, et al (1994) Molecular genetic study on native inhabitants of Chukotka: mitochondrial DNA and ALDH, ADH polymorphisms. Sibirskii Ekologicheskii Zhurnal 2:149–162

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