Anthropological Science 110(3), 293–312, 2002

A Large-scale Analysis of Human Mitochondrial DNA Sequences with Special Reference to the Population History of East Eurasian

H. Oota[1], N. Saitou[2], and S. Ueda[1]

[1] Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo
[2] Division of Population Genetics, National Institute of Genetics, Mishima

(Received June 4, 2001; accepted May 20, 2002)

Abstract Ancient DNA technique is a very powerful tool for the studies on past human populations. However, in most cases ancient DNA is extremely degraded into short fragments, and the information is limited because of the damaged state. A large-scale data analysis for human mitochondrial DNA (mtDNA) was carried out to assess validity of the short nucleotide sequence for closely related human populations. We retrieved mtDNA data from the DDBJ/EMBL/GenBank nucleotide sequence database and constructed a data set containing 414 distinct mtDNA types derived from 19 populations of East Eurasia and the surrounding area. A series of new procedures were applied and an mtDNA phylogenetic tree was constructed. Six major star-like clusters were observed in this tree, and the corresponding six radiation groups (I-VI) were characterized. Frequency distributions of each radiation group showed remarkable difference in each geographical area, suggesting that the short mtDNA nucleotide sequences were valuable in analyzing ancient human populations. The efficient procedure for data analysis will enhance the usefulness of ancient DNA data. Additionally, we discuss a possibility of two human migration routes from Africa to East Eurasia based on the mtDNA tree topology and the coalescence times in each radiation group.

Keywords: ancient DNA, mtDNA, East Eurasian populations, star-like clusters, a large scale of data analysis

Introduction

Mitochondrial DNA (mtDNA) has been most often used to ancient DNA analysis because of the high copy number per cell which gives higher chances for successful amplification on DNAs from archaeological human remains such as bones and teeth (e.g., Pääbo et al., 1988; Horai et al., 1989; Kurosaki et al., 1993; Handt et al., 1994; Oota et al., 1995; Krings et al., 1997). We recently analyzed the human remains

Corresponding author: Hiroki Oota
Department of Biological Sciences, Graduate School of Science, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Current address:
Department of Genetics, Yale University School of Medicine
333 Cedar Street, P.O. Box 208005, New Haven, CT 06510-8005, U.S.A.
E-mail: hiroki.oota@yale.edu

excavated from a 2,000-year old site in the Shandong Peninsula of China and determined nucleotide sequences of their mitochondrial hypervariable regions I and II (HV1 and 2) (Oota et al., 1999a). The 185bp fragment sequence data were used for the phylogenetic analyses, and showed that the 2,000-year old population was closely related to modern Han Chinese in Taiwan. Our ancient DNA analyses thus have offered important information regarding phylogenetic relationships among East Asians (roughly corresponds to the so-called Mongoloid, Saitou 1995).

However, in general, ancient DNA is extremely damaged and degraded into short fragments in most cases (Oota et al., 1999b), which occasionally do not indicate any statistical significance in phylogenetic analyses using these nucleotide sequences. This difficult situation calls for an efficient phylogenetic analysis to enhance the usefulness of ancient DNA data. Recently a large number of nucleotide sequences, especially for the hypervariable region I of the control region, have been deposited in the DDBJ/EMBL/GenBank International Nucleotide Sequence Database (e.g., Vigilant et al., 1991; Di Rienzo and Wilson 1991; Ward et al., 1991, 1993; Shields et al., 1993; Batista et al., 1995; Redd et al., 1995; Sykes et al., 1995; Betty et al., 1995; Kolman et al., 1996; Comas et al., 1996; Horai et al., 1996). In this study, we retrieved these human mtDNA nucleotide sequences from the DDBJ/EMBL/GenBank database, in order to examine precision of the phylogenetic analyses using the mtDNA short sequence. Using these available data, we traced the genealogy of mtDNA of human populations in East Eurasia and its surrounding area, showing remarkable characters in each geographical area. Here we propose a series of efficient procedures that have not been detailed in our previous study (Oota et al. 1999a) in order to interpret ancient DNA data exactly.

Materials and Methods

Sequence Retrieval

Nucleotide sequences of mtDNA were retrieved from the DDBJ/EMBL/GenBank International Nucleotide Sequence Database by using BLAST 1.49 (Altschul et al., 1990) implemented at the supercomputer system of the National Institute of Genetics, Japan. The 185-bp fragment corresponding to nucleotide positions 16,194 to 16,378 in the Cambridge Reference Sequence (CRS) (Anderson et al., 1981) was used as a query. (Following this, we subtract 16,000 from the position number in the CRS, and this simplifies the nucleotide position [np] designation. For example, CRS position 16,194 is designated as np194.) There are three reasons for choosing *this* sequence region for this analysis and for ancient DNA analyses we have done (Oota et al., 1995; 1999a). First, it includes many informative sites shared among East Asians (Horai and Hayasaka 1990; Betty et al., 1995; Kolman et al., 1996; Horai et al., 1996). Second, it does not contain the cytosine-repeat between np184 and np193 that Bendall and Sykes (1995) mentioned the possibility of heteroplasmy, which is

confusing for contamination. Third, most of the mtDNA sequences deposited in the database included this 185-bp region. Therefore, this region is suitable for analyzing phylogenetic relationships between closely related human populations, especially in East Asia.

Sequences with insertions and/or deletions are not retrieved under the BLAST version we used. In this study we did not include such sequences because the frequency of insertions and deletions was much lower than that of substitutions (Saitou and Ueda 1994) and less than 0.1% among all sequences deposited in the database.

We retrieved 2,215 sequences from the DDBJ database. Subsequently, we eliminated those sequences containing various lengths and/or having undetermined nucleotides designated by "N," and also excluded non-Circum-Pacific populations (European and African). In the International Nucleotide Sequence Database, sequences were deposited mainly in two forms. Some research groups (e.g., Ward et al., 1991, 1993) combine identical sequences while others deposit all sequences tested (e.g., Vigilant et al., 1991; Redd et al., 1995). In the former case, we examined the number of individuals by referring to the original papers. There was no information about the name of ethnic groups in the database entries for Sykes et al. (1995), though several populations were specified. Therefore, the sequences, the names of ethnic groups, and the number of individuals were added from their Table 2 after eliminating all of their sequences from 2,215 sequences of our original data set. As a result, mtDNA sequences for a total of 1,298 individuals were compared.

The naming system of human populations varied from paper to paper. Some populations were referred by the names of ethnic groups and others by geographical regions. To facilitate comparison between populations in the continental region of the East Eurasia, we used the names of ethnic groups defined by the authors when describing East Asians. For example, we distinguished Taiwan Han Chinese from Cantonese, and Siberian Altai from Mongolian. "Asian" described in Vigilant et al. (1991) contained the following 11 groups according to Vigilant (1990) and Vigilant (personal communication): Chinese, Southern Chinese, American Chinese, Hmog, Japanese, Tongan, Indonesian, Taiwanese, Philippine, American Philippine and a hybrid of Amerindian mother and African American father. We divided those "Asian" into eight groups; Chinese (containing American Chinese), Southern Chinese (containing Hmog), Japanese, Indonesian, Taiwanese (as Aboriginal Taiwanese), Philippine, Tongan (as Polynesian) and Amerindian (or Native American). Yet distinction between Chinese and Southern Chinese was vague, so we refer to them as "PRC Chinese," meaning Chinese from the People's Republic of China. For the New World, various ethnic groups were combined as a single population called "Native American."

The original data set was thus classified into 19 ethnically defined populations, and the 19 populations were combined into seven major geographical populations

Table 1. Citations and acession numbers in the DNA data base

Geographic region [total]	Population ID	Population Name	No. of individuals	Accession numbers	Reference
West Asia [45]	1	Turk	45	U59009-U59053	Comas et al., 1996
Far East Asia [227]	2	Ainu	51	D84965-D85015	Horai et al., 1996
	3	Mainland Japanese	62	D84723-D84784	Horai et al., 1996
				(M76257...M76368)[a]	Vigilant et al., 1991
	4	Southern Korean	64	D84785-D84848	Horai etal., 1996
	5	Ryukyu Japanese	50	D84849-D84898	Horai etal., 1996
Continental East Asia [221]	6	Taiwan Han Chinese	66	D84899-D84964	Horai et al., 1996
	7	Cantonese	20	U37734-U37753	Betty et al., 1995
	8	PRC Chinese	17	(M76257.....M76368)[a]	Vigilant et al., 1991
	9	Mongolia	101	U33336-U33418	Kolman et al., 1996
	10	Altai of Central Siberia	17	L20198-L20213	Shields et al., 1993
New World [353]	11	Native Americans	353	L20143-L20155	Ward et al., 1991; Shields et al., 1993
				L20157-L20197	Ward et al., 1993
				L39319-L39325	Batista et al., 1995
				L39327-L39355	Kolman et al., 1995
				M75991-M76018	Ward et al., 1991; Shields et al., 1993
				(M76257.....M76368)[a]	Vigilant et al., 1991
Southeast Asia [136]	12	Indonesian	65	(U25334-U25413)[b]	Redd et al., 1995
				(U47145-U47271)[b]	Sykes et al., 1995
				(M76257...M76368)[a]	Vigilant et al., 1991
	13	Aboriginal Taiwanese	35	(U47145-U47271)[b]	Sykes et al., 1995
				(M76257.....M76368)[a]	Vigilant et al., 1991
	14	Philippine	36	(U47145-U47271)[b]	Sykes et al., 1995
				(M76257...M76368)[a]	Vigilant et al., 1991
Australia [5]	15	Australian Ahorigine	5	U37730-U37733	Betty et al., 1995
				M76283	Vigilant et al., 1991
Oceania [311]	16	Melanesian	104	(M76257...M76368)[a]	Vigilant et al., 1991
				(U25354-U25413)[b]	Redd et al., 1995
				(U47145-U47271)[b]	Sykes et al., 1995
	17	Micronesian	5	(U47145-U47271)[b]	Sykes et al., 1995
	18	Kapingamrangi	16	(U47145-U47271)[b]	Sykes et al., 1995
	19	Polynesian	186	(U47145-U47271)[b]	Sykes et al., 1995
				(U25354-U25413)[b]	Redd et al., 1995
				(M76257...M76368)[a]	Vigilant et al., 1991
Total [1298]			1298		

a. not series and containing other populations; b. including all populations

(Table 1). The data sets are available at the following web site: http://smiler.lab.nig.ac.jp/mtDNA/.

Phylogenetic Analysis

Conventional phylogenetic tree-making methods are not designed for short and closely related sequences such as ancient human mtDNA sequences. To enhance usefulness of such data, we therefore devised a series of procedures. The BLAST outputs are preprocessed by using a series of programs developed by Saitou (2000), and identical sequences as well as invariant sites were eliminated. We then count the number of populations and individuals for each mtDNA type thus obtained. This is because the frequency information for each population is used to build a phylogenetic tree in this new procedure. Using mtDNA types shared with more than one population, we first constructed a skeleton network. MtDNA types were joined stepwise, and types shared with more populations were given priority. When two types were shared with the same number of populations, the type containing more individuals was joined first. To represent incompatible parallel mutations, we permitted reticulations (Bandelt 1994; Bandelt et al., 1995) in this step. However, the relationship of non-recombining mtDNA sequences (Kumar et al., 2000) was expected to have a tree structure without reticulation. Therefore, all the reticulations were resolved according to the three criteria: (1) a connection between the types shared with more populations is favored over a connection between the types shared with fewer populations, (2) when two types are shared with the same number of populations, a connection between the types shared with more individuals is favored over a connection between the types shared with fewer individuals (Excoffier and Langaney 1989; Excoffier and Smouse 1994; Bandelt 1994; Bandelt et al., 1995), and (3) parallel transition is favored over parallel transversion (Bandelt et al., 1995).

Fig. 1 illustrates this tree buildup procedure from the hypothetical data of Table 2. Hypothetical mtDNA types (A—I) of Table 2 differ in one substitution from one or more other types. We first construct a skeleton network using the types (A–F) shared with more than one population. The types shared with more populations are given priority, so we start by connecting types A and B, and the skeleton network is constructed (see Fig. 1). Reticulations appear when incompatible partitions exist, such as between np1 and np3. We then resolve those reticulations in order to make a tree structure according to our criteria. Type B is shared with 45 individuals from nine populations, whereas type E is shared with 30 individuals from five populations. The connection of type B with type D has priority over that of type E with type D according to criterion 1. Therefore, the connection between types D and E is eliminated. Type C as well as type E is shared with the same number of populations. In this case, the connection between types E and F is given priority over the connection of types C and F, because type E is shared with more individuals than type C (criterion 2).

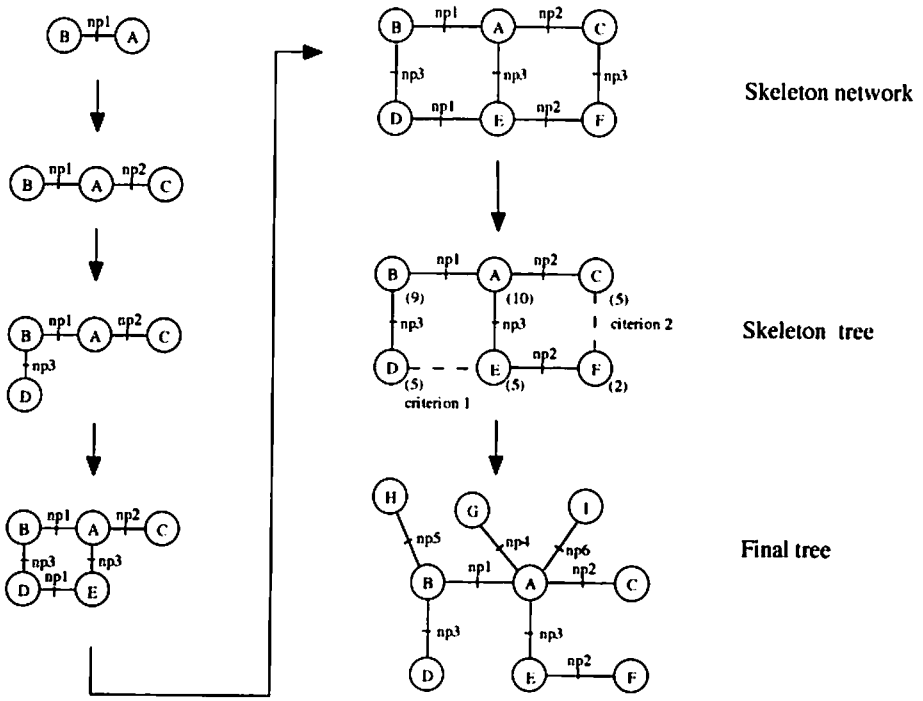

Figure 1. Tree construction based on the hypothetical data of Table 2.

Table 2. Hypothetical data

type	No. of shared populations	No. of individuals	Sequence data np1	np2	np3	np4	np5	np6
A	10	40	0	0	0	0	0	0
B	9	45	1	0	0	0	0	0
C	5	5	0	1	0	0	0	0
D	5	40	1	0	1	0	0	0
E	5	30	0	0	1	0	0	0
F	2	4	0	1	1	0	0	0
G	1	10	0	0	0	1	0	0
H	1	2	1	0	0	0	1	0
I	1	1	0	0	0	0	0	1

"np" represents nucleotide position. "0" means a consensus sequence, "1" means a nucleotide difference from that.

These procedures produce the skeleton tree. After that, other types (G, H, and I) shared with a single population are joined. This final tree has identical topology with that constructed by using the site-by-site method of Bandelt (1994).

Because the actual data was complicated, mtDNA types found in a single population were joined by applying the following two additional criteria; (4) a connection between types shared with the same population is given priority, and (5) when all criteria described above are not fitted, reticulations are split arbitrarily. The fifth criterion, however, did not affect the entire result. To confirm the topology of the phylogenetic tree, we also constructed neighbor-joining trees (Saitou and Nei 1987) using CLUSTAL W (Thompson et al., 1994) and maximum likelihood trees (Felsenstein 1981) using DNAML of PHYLIP 3.5 (Felsenstein 1993).

Results

Distribution of Variant Sites

Among 1,298 individual nucleotide sequences of mtDNA, 414 distinct types were identified and differences among them reside in 109 variant sites. Fig. 2 shows the frequency distribution of the variant sites. The frequency of transversional difference was less than 4%. Seven nucleotide positions with more than 10% frequencies are shown. Np223 and np362 reached 47% and 29%, respectively among those seven positions.

Frequency of the Types Sharing Multiple Populations

Table 3 shows frequency distribution of mtDNA types. Some mtDNA types contained only one individual while others were shared among many individuals. Three hundred and fifty two types were found only in one population, whereas 62 mtDNA types were shared with multiple populations. Thirteen out of those 62 types (1, 3, 9, 13, 15, 25, 32, 33, 58, 61, 113, 138, and 173) were shared among five populations or

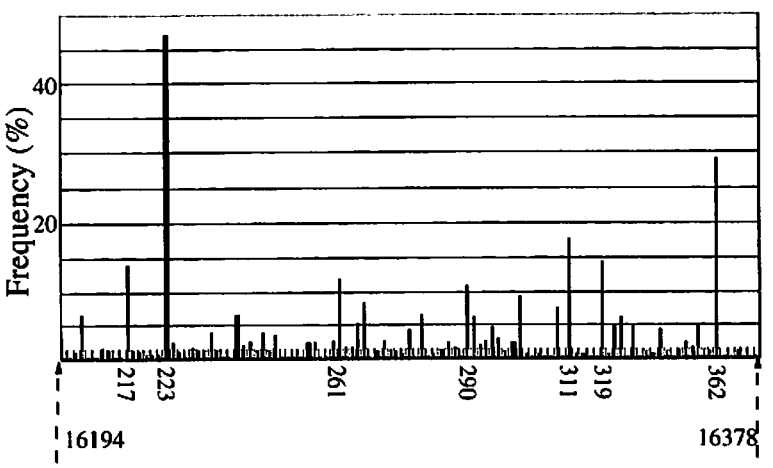

Figure 2. The frequency distribution of the 109 variant sites in 414 mtDNA types. Nucleotide position numbers showing 10% or more variant types are given.

Table 3. mtDNA types shared multiple populations

No. of populations	No. of mtDNA types	Type	No. of individuals	No. of descendants
14	1	**13**	30	32
11	1	**25**	49	30
9	1	**138**	45	13
8	1	**3**	18	15
7	2	**9**	53	17
		1	23	33
6	3	61	14	
		15	7	
		58	11	
5	4	**173**	135	14
		113	107	15
		32	10	
		33	5	
4	10		72	
3	12	**7** & others	70	12
2	27		101	
1	352		548	
Total	414		1298	

Types with more than 10 descendants types or shared with more than 5 populations are shown. Bolds present backbone types.

more. Type 13 was shared with the largest number of populations while type 1 corresponds to the CRS (Anderson et al., 1981). Those 62 types were used to construct the skeleton tree.

Multiple Star-like Clusters in the Phylogenetic Tree

We first constructed an unrooted skeleton tree for the 62 mtDNA types shared with more than one population following our five criteria, and then joined the remaining 352 types found in a single population to the skeleton tree. In the unrooted tree using these 414 types, though it is just one of many possible trees, we observed many star-like clusters, most of which originated from the mtDNA skeleton types (tree not shown). Nine mtDNA types (1, 3, 7, 9, 13, 25, 113, 138 and 173) among those skeleton types had more than 10 descendant types, which are considered to be the centers of radiations. We called those nine types RC (radiation center) types. Eight of the nine types were shared with five or more populations and with more than 10 individuals in each (Table 3).

Examination of the Network of RC Types

In our data set, we observed parallel mutations at np223, np311, and np362 that were shared with 15% mtDNA types or more (Fig. 2). We thus constructed a phylogenetic network of those nine RC mtDNA types (Fig. 3). Pairs of incompatible sites are expressed as reticulations in the network. Thirteen types (8, 18, 33, 41, 44, 61, 73, 74, 96, 164, 175, 268 and 356) shown in open circles were added to the phylogenetic network of nine RC types in Fig. 3 to clarify the reticulation patterns. Numbers in parentheses represent the number of the populations sharing that type. A connection between the types shared with more populations is favored over a connection between the types shared with fewer populations (criterion 1). For example, it was more probable for type 44 in the connection with type 9 than with type 18 because type 9 was shared with seven populations while type 18 was shared with only one population. All the reticulations were resolved according to this rule, and parallel mutations at np223, np311, np319 and np362 were assumed to produce a tree from the network. The dotted lines represent the unconnected edges.

To test reproducibility of the tree topology, two neighbor-joining trees were constructed using nine RC types based on the two kinds of nucleotide lengths: 185-bp and 341-bp nucleotide sequences corresponding to 16,194–16,378 and 16,038–16,378

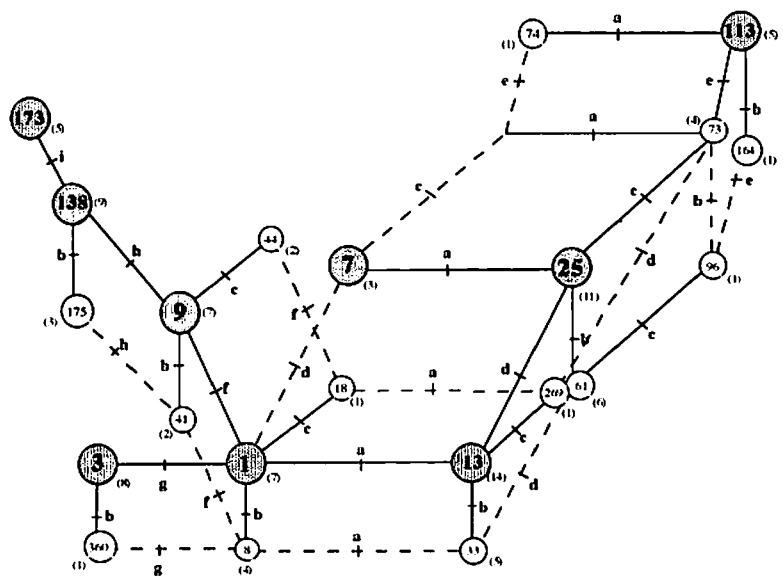

Figure 3. A phylogenetic network of nine backbone types (shaded) and related 14 types (non-shaded). Small letters correspond to nucleotide positions (a: np223, b: np311, c: np319, d: np362, e: np290, f: np217, g: np304, h: np261, i: np247). The numerals in parentheses represent the number of populations sharing the mtDNA types. Dotted lines represent disconnected edges.

in the CRS, respectively (Fig. 4). We also included sixteen sequences of Pygmies (Vigilant et al., 1991) with distinct mtDNA lineages. Furthermore, the mtDNA sequence of a Neanderthal (Krings et al., 1997) was used as the outgroup. The tree of Fig. 4a was constructed based on 185-bp nucleotide sequences used to construct the tree of Fig. 3. Types 25 and 113 form one cluster, while types 9, 138 and 173 form another cluster, and this is further clustered with types 1 and 3. The topology of the NJ tree of Fig. 4a was in accord with that of the tree of Fig. 3, except for the location of type 7 that differed from types 1 and 25 in only one nucleotide (Fig. 3). Based on the RFLP analysis of mtDNA of Native Americans, Forster et al. (1996) pointed out that transitions at np362 were separated into two parallel events (Figure 1 of Forster et al., 1996). This ambiguity in the location of type 7 probably caused the incompatible parallel mutations at np362. However, we connected type 7 with type 25 in the tree of Fig. 3 according to our criteria described in Materials and Methods. The tree of Fig. 4b is based on 341-bp nucleotide sequences including the 185-bp ones, showing the same topology as Fig. 4a. Furthermore, we constructed another unrooted tree from 62 skeleton types based on 341-bp nucleotide sequences (tree not shown); this tree had the same topology as Fig. 3.

In both trees of Fig. 4a and 4b type 13 is the closest sequence to the root of the tree, which was confirmed by the maximum likelihood analysis based on 185-bp and

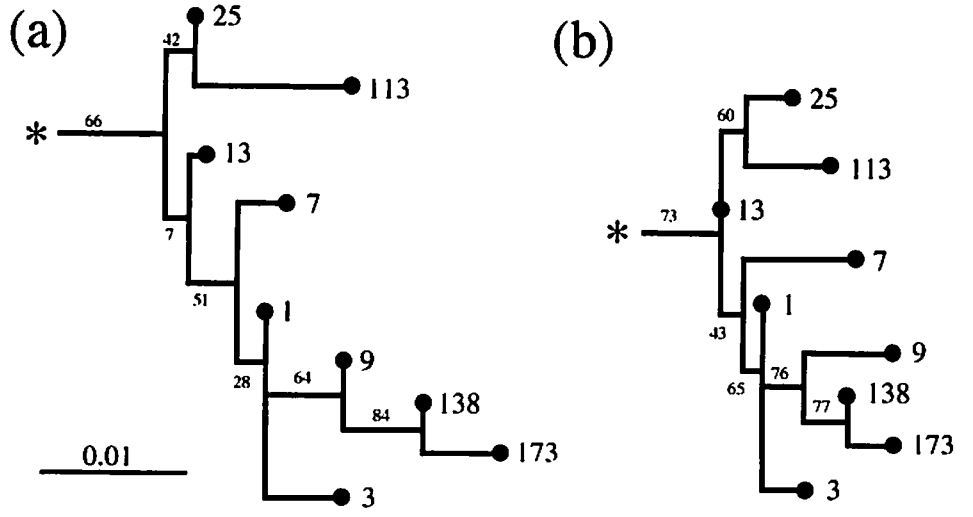

Figure 4. Neighbor-joining trees for the nine backbone types were constructed based on a) 185-bp mtDNA sequence, and b) 341-bp mtDNA sequence. Both trees were rooted by including 16 mtDNA types of Pygmies (Vigilant et al., 1991) and a Neanderthal sequence (Krings et al., 1997). Positions of the root are denoted by asterisks. Bootstrap resampling was employed 1,000 times, and the resulting bootstrap probabilities (%) are shown on interior branches.

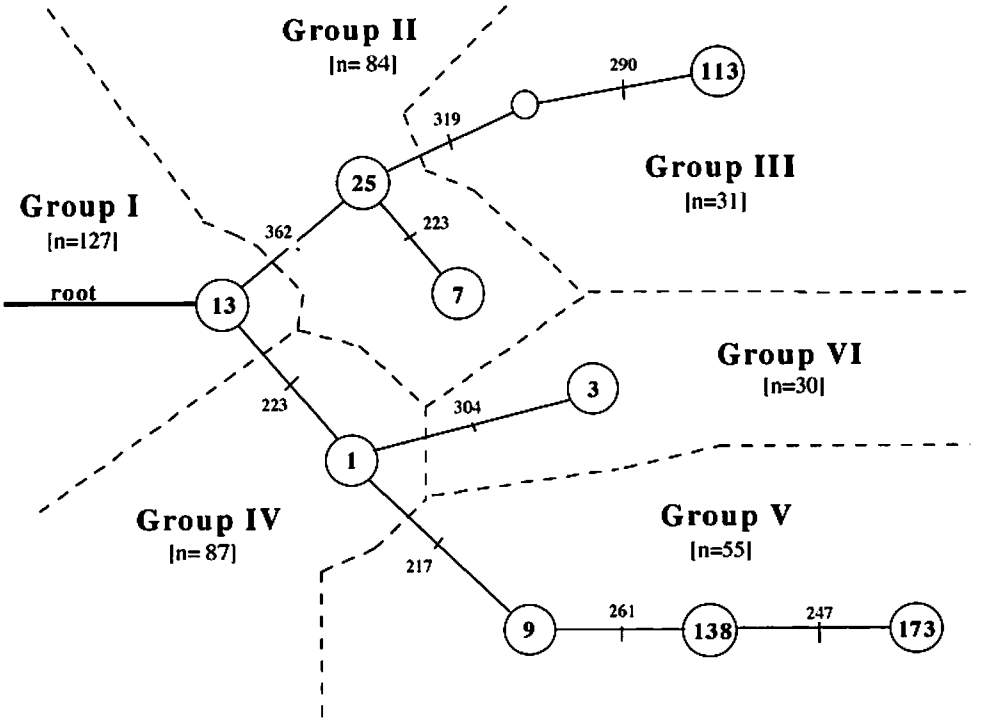

Figure 5. Characterizations of six radiation groups (I–VI). The nine radiation center (RC) mtDNA types correspond to the centers of stars. Numbers on the branches represent nucleotide positions. The numerals in brackets represent the number of mtDNA types included in the radiation groups.

341-bp nucleotide sequences (trees not shown). Among 414 types, type 13 was shared with 14 populations (Table 3) that were derived from all of East Eurasia and the Circum-Pacific region. We thus concluded that type 13 was the root of the tree.

Characterization of Six Radiation Groups

To reveal the general picture of the population structure, we characterized six radiation groups. The rooted tree of Fig. 5 shows the genealogy of those radiation groups (I–VI). Group I, which radiated from type 13, includes 127 mtDNA types. We inferred that Group I was the root of six radiation groups at least in East Eurasia and the Circum-Pacific area. Group I was characterized by both T at np223 and T at np362.

Based on the D-loop region sequencing and the restriction fragments length polymorphism (RFLP) analysis, Torroni et al. (1993a, 1993b) divided mtDNA types found in Native Americans into four major groups (haplogroups A, B, C, and D). In this study haplogroup C belongs to group I, whereas haplogroup D belongs to group II.

Group II (containing 84 mtDNA types) corresponds to the radiation group centered at type 25 with type 7 as a sub-center. Group II was characterized by C at np362 and G at np319. Group III (31 types), containing descendants of type 113, was characterized by T at np223, C at np362, and A at np319. Group V (55 types) was characterized by C at np217, C at np223, and T at np362. Groups III and V corresponded to haplogroups A and B in Native Americans, respectively. Thus, haplogroups A and B that contained the majority of Native Americans were located on the periphery of the whole tree.

Group VI (30 types), which radiated from type 3, was characterized by C at np304, T at np217, C at np223 and T at np362. Because group IV (87 types) was the CRS and its descendants, it was characterized by subtracting groups I, II, III, V and VI. There were eight exceptional types in the characterization of radiation groups (see the web site).

Pattern of Frequency Distribution

The number of individuals in each radiation group was counted and the pattern of frequency distribution showed geographical characteristics (Table 4). A high frequency (67%) in Group IV was found in West Asia. In Oceania, 62% were included in Group V (haplogroup B). In the New World, 59% were included in Group III (haplogroup A). The high frequency for Group V (35%) was also observed in Southeast Asia, though it was lower than that in Oceania. Thus, the frequency distributions of groups III, IV and V indicated clear features specific to each geographical area.

Frequency of Group VI was low in most of the geographical areas. It was less than 5% in West Asia, Far East Asia, and Oceania, but showed more than 10% in Continental East Asia. It is thus one of the characteristics of Continental East Asians. Group II showed 33% in Far East Asia and 29% in Continental East Asia, whereas it showed 19% in Southeast Asia and 6% in Oceania. West Asia and Oceania had lower frequencies (7% and 6%, respectively) than Far East Asia and Continental East Asia.

Table 4. Frequency distribution for radiation groups (%)

		I	II	III	IV	V	VI
West Asia	[N = 45]	20	7	2	67	0	4
Far East Asia	[N = 227]	44	33	3	11	4	4
Continental East Asia	[N = 221]	31	29	5	14	10	13
New World	[N = 353]	11	10	59	2	17	0
Southeast Asia	[N = 136]	20	19	0	18	35	8
Australia	[N = 5]	80	0	0	20	0	0
Oceania	[N = 311]	17	6	0	12	62	2
Europe	[N = 519]	9	3	0	81	1	6
Africa	[N = 73]	79	5	0	16	0	0

Southeast Asia had higher frequency in Group II than Oceania (19% and 6%, respectively). Thus, the East Eurasian populations (Far East Asians and Continental East Asians) had a relatively high frequency of Group II.

To compare with the distribution in East Eurasian and the Circum-Pacific area, we retrieved 73 African and 519 European sequences from the DDBJ database (Vigilant et al., 1991; Di Rienzo and Wilson 1991; Sajantila et al., 1995; Arnason et al., 1996). They were divided into six radiation groups according to the characteristics of nucleotide substitutions and the frequencies were counted. In Africa, Group I was remarkably high (79%) and Group IV existed in low frequency (16%), whereas the other three groups (III, V and VI) were not found. In Europe, 81% was included in Group IV. The reason why African and European are occupied in only one group is that the radiation groups are characterized based on our original data set from populations in East Eurasian and the Circum-Pacific area.

Heterogeneity of substitution rate for the human mtDNA D-loop region has been known (e.g., Excoffier and Yang 1999; Meyer et al., 1999), and sites with high rates such as np223, np311, np362 in fact caused reticulations observed in the phylogenetic network of Fig. 3. Backward mutations occurred at those high-rate sites may cause erroneous classification among the six groups defined above. However, substitution rates of those sites are only four times higher than the average rate (Excoffier and Yang 1999), so their effects caused by backward mutations may be relatively small.

Discussion

We analyzed nucleotide sequences of mtDNAs deposited in the DDBJ/EMBL/GenBank International Nucleotide Sequence Database. A phylogenetic tree regarding East Eurasian populations was constructed by applying a series of criteria. We observed six major star-like clusters in this tree and characterized the corresponding putative six radiation groups (I–VI). In spite of the short DNA fragment used for the phylogenetic analysis, the frequency distributions of the radiation groups showed obvious feature specific to each geographical region. This indicates that the short region is suitable for ancient DNA analyses in East Eurasian populations, and implies that the groups characterized by the star-like clusters have some meanings.

The star-like phylogeny may provide possible evidence of a demographic expansion in past populations and the center of the star is the root lineage of expansion (Slatkin and Hudson 1991; Sherry et al., 1994). This idea was initially proposed by Di Rienzo and Wilson (1991); the center of star-like phylogeny found by Di Rienzo and Wilson (1991) corresponds to type 1 (CRS) in this study, which has the highest frequency (more than 20%) in Europe (Richards et al., 1996). This type was suggested to be the ancestral mtDNA type of the modern European population (Bandelt et al., 1995; Richards et al., 1996). In the tree based on 414 mtDNA types, there were nine centers of radiation (RC types) and more than 20 centers of near star-like phy-

logeny (tree not shown), implying that multiple demographic expansions might have also occurred in the East Eurasia and its surrounding region.

If those nine RC mtDNA types are the centers of demographic expansion, their appearance times correspond to the periods when these occurred. Assuming that is true, we estimated these times in the following way. We assumed that each descending lineage from one center (backbone type) started to diverge from the common ancestral type simultaneously. We further assumed existence of molecular clock, in other words, the mean of the nucleotide substitutions between extant lineages and the common ancestor (center) was expected to be proportional to the divergence time of these lineages from the ancestral type. Counting the number of lineage is a problem in this simple procedure. When we found a unique nucleotide substitution in one sequence from the ancestral type, the lineage to this extant sequence definitely existed. For example, RC type 173 had 14 such descendant lineages. When there are hierarchical structures for some lineages, nucleotide substitutions are averaged by taking into account the tree structure following Ishida et al. (1995).

The biggest problem of this procedure occurs when some extant sequences are identical with the center node. Because we have no information on the genealogical structure for those identical sequences, we considered the two extreme situations. One extreme is that all identical sequences are considered to represent independent lineages. Under this assumption we obtain a lower bound for the estimate of divergence time. Another extreme is the situation that all the identical sequences were derived only recently and give only one lineage from the center. This gives us the upper bound. For example, there are 135 individuals who had RC type 173 (Table 3). Therefore, the lower bound for the average number of substitution between the common ancestral type and the extant sequences is $14/(14 + 135) = 0.09$, while the upper bound becomes $14/(14 + 1) = 0.93$, more than ten times larger than the lower bound.

Furthermore, the evolutionary rate is necessary to obtain the time estimate. We used the rate estimated by Horai et al. (1996). Their estimate ($l = 8.6 \times 10\text{-}8$/site/year) was based on the 482-bp HV1 segment, and our 185-bp segment is overlapped with this. It is possible that the real evolutionary rate for the latter region may be different from that for the former region, but we assume that the difference is relatively small. In any case, when the evolutionary rate (l) and evolutionary distance (d) between extant and ancestral sequence is known, the divergence time (t) can be estimated as $t = d/l$ under the assumption of molecular clock. The assumption that gives a lower bound is more appropriate, because the nine backbone types are considered to be centers of demographic expansions and many lineages are expected to diverge independently under the population expansion. Therefore, we mostly relied on lower bound estimates in the following.

On the basis of the genealogy in the unrooted tree using 414 mtDNA types, the lower boundaries of coalescence times (years) of nine radiation center types were

estimated (Fig. 6). Group I probably originated in Africa; their coalescence time was estimated to be around 55,000 years ago. As well as Group I, Group IV also exists in all geographic regions. Group IV showed high frequency in West Asia and Europe, and its originating time was estimated to be around 64,000 years ago. This estimation agreed with the time estimated by Richards et al. (1996). The coalescence time estimation of Group IV is older than that of group I (55,000 yr). Because African populations are not contained in our original data set, the coalescence time of Group I may be a gross underestimation.

There were at least three major migrations into Oceania according to archaeological and linguistic studies (Bellwood 1985). First, an ancient Australoid (Sahulian) migrated from the Indo-Malaysian archipelago to Australia and New Guinea around 40,000 years ago. Second, Austronesians migrated from southern China and settled in Southeast Asia, and some of them migrated to islands 4,000–6,000 years ago. Third, migration of Austro-Asians began around 3,500 years ago. Groups V and VI are descendants of the backbone mtDNA type 1 of Group IV. Group V contained 77% of Polynesians, 53% of Aboriginal Taiwanese and 33% of Melanesians. Group V was not found in West Asia, and it existed in low frequency in Far East Asia (4%).

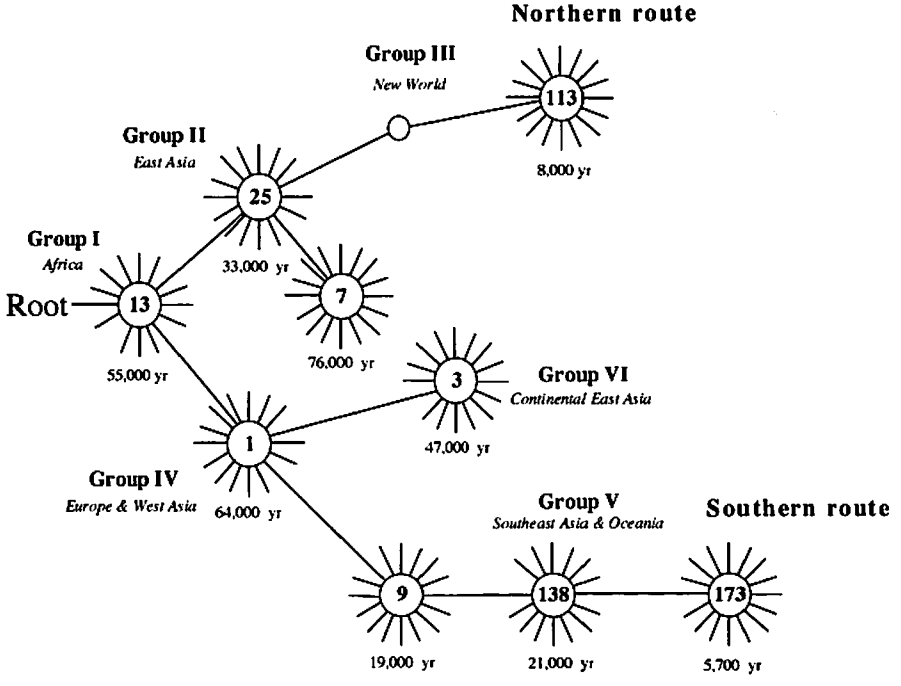

Figure 6. Proposed model of the peopling history of East Eurasia and the Circum-Pacific area. Lower limits of coalescence times (years) for the radiation groups are given for each radiation center.

Although it existed in Taiwan Han Chinese (14%) and Cantonese (10%), it was very low in Mongolia (7%) and was not found in Siberian Altai. Radiation centers of Group V are types 9, 138, and 173; the radiation cluster of type 9 contained 18% of Native Americans and several individuals from the Eurasia continental populations. In contrast, Native Americans were not found in the radiation of type 138, and Polynesians were exclusively found in the radiation of type 173. This implies that the radiation cluster of type 9 in Group V originated somewhere in the East Eurasian Continent, and the radiation clusters of types 138 and 173 originated in the southern part of the East Eurasian Continent or Oceania. The coalescence times of the radiation clusters in Group V overlapped with the periods of migration into Southeast Asia/ Oceania, which were based on archaeological and linguistic evidences. Thus the mtDNA genealogy of group V can be considered as the trace of human migration via the southern route crossing from India to Southeast Asia.

Group II may be the trace of expansion in East Asia, and its origination time was estimated to be around 33,000 years ago. Frequency of Group III, a descendant of Group II, was remarkably high in the New World (59%; Table4). In contrast, Group III haplotypes were not found in Southeast Asia, and Oceania as well as Africa and Europe. This implies that Group III originated in the northern part of the East Eurasian Continent. The population could have increased in size after ancestral people of Native Americans migrated to the New World. The Group III was estimated to originate around 8,000 years ago. From these arguments, the mtDNA genealogy of group III could be considered as the trace of human migration via the northern route across the mountains lying south and west of Tibet. The first human migration to the New World is thought to be around 15,000 years ago from the archaeological evidence (Morell 1990). Based on the analyses of mtDNA RFLP, Torroni et al. (1993b) estimated that the first migration occurred between 17,000 and 34,000 years ago. Though the archaeological estimation is still controversial, these dates above overlap with our estimations.

Although only 5% of the entire data set was included in Group VI, it was especially high (23%) in Taiwan Han Chinese; Group VI was one of the genetic characteristics of Taiwan Han Chinese. We found the high frequency of Group VI in human remains excavated from an archaeological site (around 2,000 years ago) at the Shandong Peninsula in Mainland China (Oota et al., 1999a). The coalescence time of Group VI estimated to be around 47,000 years ago suggests that the period when putative demographic expansion occurred in Group VI preceded those of groups II, III and V. So far, we have no modern data from the land region of Southeast Asia, where Group VI could be the major cluster. It is important to obtain data from these regions in the near future.

Saitou (1995) proposed a new terminology of human population clusters existed about 10,000 years ago: African, West Eurasian, and Circum-Pacificans including

East Eurasian, Sahulian, North American, and South American. We discuss the peopling history of Circum-Pacificans using this terminology in the following. In evolutionary genetic studies, population trees have been constructed using traditional markers. Based on the phylogenetic analyses, probably human populations can be subdivided into several major groups. By using polymorphic nuclear marker data, it is known that Circum-Pacifican was diverged from West Eurasian around 55,000 years ago after these two major groups diverged from African around 115,000 years ago (Nei and Roychoudhury 1974). Nei and Roychoudhury (1993) proposed that there were two migration routes to East Eurasia; one was the route crossing mountainous area in south and west of Tibet, and the other was the route from India to Southeast Asia.

Circum-Pacificans have been divided into northern (Sinodonts) and southern (Sundadonts) groups based on dental morphology (Turner 1990). Two major lineages that could be considered as traces of southern and northern routes were observed in our mtDNA phylogenetic tree. Because Group II contains some fraction of individuals in southern populations, this is not incompatible with the model that Circum-Pacificans originated in the southern part of Asia. But it does not strongly support the model for the following three reasons: 1) Group I considered as the root is more frequent in the northern area than in the southern area; 2) Group II that would have been the ancestral haplogroup of Group III was more frequent in the northern area than in the southern area; and 3) from the topology of our mtDNA tree, it is possible to conclude that both the southern and the northern routes existed. Therefore, it may be necessary to revise the southern origin model of all Circum-Pacificans.

Conclusion

Even a single locus using a short sequence data, the phylogenetic trees by a new series of procedure discussed in this study thus showed genetically remarkable characteristics in each population. This means that the short DNA fragment can show the genetic differentiation of Circum-Pacificans to the extent indicated in the frequency distributions (Table 4). Meanwhile, the result also suggests that the longer DNA region is necessary for studying further detailed genetic relationships between ethnical populations within each geographical group. Using many primer-sets in PCR amplification, the longer fragments (hopefully 300–400 bp) may be obtained by combining those short fragments. This way to elongate DNA fragments should be tried in all ancient DNA studies. However, in many cases it is impossible to obtain such long fragments from all individuals constantly, because the damaged state of DNA is extremely depending on individual states of preservation (Oota et al., 1999b). The large scale screening of mtDNA data, therefore, plays an important role in compensating missing data in ancient population studies.

Acknowledgments

We appreciate Siva Kumar Swaminathan and Brenda Bradley in Max-Planck Institute for Evolutionary Anthropology in Leipzig, Germany, and Peristera Paschon in Yale University School of Medicine in New Haven, U.S.A., for their comments on this manuscript. This study was partially supported by grants-in-aids to scientific research of Ministry of Education, Science, Sports, and Culture, Japan to H.O, S.U., and N.S.

References

Altschul S. F., Gish W., Mayers E. W., and Lipman D. J. (1990) Basic local alignment search tool. Journal Molecular Biology, 215: 403–410.

Anderson S., Bankier A. T., Barrel B. G., de Bruijn M.H.L., Coulson A. R., Drouin J., Eperon I. C., Nierlich D. P., Roe B. A., Sanger F., Schreier P. H., Smith A.J.H., Staden R., and Young I. G. (1981) Sequence and organization of the human mitochondrial genome. Nature, 290: 457–465.

Arnason U., Xu X., and Gullberg A. (1996) Comparison between the complete mitochondrial DNA sequences of Homo and the common chimpanzee based on nonchimeric sequences. Journal of Molecular Evolution, 42: 145–152.

Bandelt H.-J. (1994) Phylogenetic networks. Verh. Naturewiss. Ver. Hamburg (NF), 34: 51–71.

Bandelt H.-J., Forster P., Sykes B. C., and Richards M. B. (1995) Mitochondrial portraits of human populations using median networks. Genetics, 141: 743–753.

Batista O., Kolman C. J., and Bermingham E. (1995) Mitochondrial DNA diversity in Kuna Amerinds of Panama. Human Molecular Genetics, 4: 921–929.

Bellwood P. (1985) Prehistory of the Indo-Malaysian Archipelago. Academic Press, Sydney.

Bendall K. E. and Sykes B. C. (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. American Journal of Human Genetics, 57: 248–256.

Betty D. J., Chin-Atkins A. N., Croft L., Sraml M., and Easteal S. (1995) Multiple independent origins of the COII/tRNALys intergenic 9-bp mtDNA Deletion in Aboriginal Australians. American Journal of Human Genetics, 58: 428–433.

Comas D., Calafell F., Mateu E., Lezaun A. P., and Bertranpetit J. (1996) Geographic variation in human mitochondrial DNA control region sequence: The population history of Turkey and its relationship to the European populations. Molecular Biology and Evolution, 13: 1067–1077.

Di Rienzo A. and Wilson A. C. (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. Proceedings of National Academy of Sciences, USA, 88: 1597–1601.

Excoffier L. and Langaney A. (1989) Origin and differentiation of human mitochondrial DNA. American Journal of Human Genetics, 44: 73–85.

Excoffier L. and Smouse P. E. (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics, 136: 343–359.

Excoffier L. and Yang Z. (1999) Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. Molecular Biology and Evolution, 16: 1357–1368.

Felsenstein J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molcular Evolution, 17: 368–376.

Felsenstein J. (1993) PHYLIP: Phylogeny Inference Package, Version 3.5c. University of Washington, Seattle.

Forster P., Harding R., Torroni A., and Bandelt H.-J. (1996) Origin and evolution of Native American

mtDNA variation: a reappraisal. American Journal of Human Genetics, 59: 935–945.

Handt O., Richards M., Trommsdorff M., Kilger C., Simanainen J., Georgiev O., Bauer K., Stone A., Hedges R., Schaffner W., Utermann G., Sykes B., and Pääbo S. (1994) Molecular genetic analyses of the Tyrolean Ice Man. Science, 264: 1775–1778.

Horai S. and Hayasaka K. (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. American Journal of Human Genetics, 46: 828–842.

Horai S., Hayasaka K., Murayama K., Wate N., Koike H. and Nakai N. (1989) DNA amplification from ancient human skeland others, remains and their sequence analysis. Proceedings of Japanese Academy, Series B, 65: 229–233.

Horai S., Murayama K., Hayasaka K., Matsubayashi S., Hattori Y., Fucharoen G., Harihara S., Park K. S., Omoto K., and Pan I.-H. (1996) MtDNA polymorphism in East Asian populations, with special reference to the peopling of Japan. American Journal of Human Genetics, 59: 579–590.

Ishida N., Oyunsuren T., Mashima S., Mukoyama H., and Saitou N. (1995) Mitochondrial DNA sequences of various species of the Genus Equus with special reference to the phylogenetic relationship between Przewalskii's Wild Hourse and Domestic Horse. Journal of Molecular Evolution, 41: 180–188.

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

Krings M., Stone A., Schmitz R. W., Krainitzki H., Stoneking M., and Pääbo S. (1997) Neandertal DNA sequences and the origin of modern humans. Cell, 90: 19–30.

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

Kurosaki K., Matsushita T., and Ueda S. (1993) Individual DNA identification from ancient human remains. American Journal of Human Genetics, 53: 638–643.

Meyer S. G., Weiss G., and von Haeseler A. (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics, 152: 1103–1110.

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

Nei M. and Roychoudhury A. K. (1974) Genetic variation within and between the three major races of man, caucasoids, negroids, and mongoloids. American Journal of Human Genetics, 26: 421–443.

Nei M. and Roychoudhury A. K. (1993) Evolutionary relationships of human populations on a global scale. Molecular Biology and Evolution, 10: 927–943.

Oota H., Saitou N., Matsushita T., and Ueda S. (1995) A genetic study of 2,000-year-old human remains from Japan using mitochondrial DNA sequences. American Journal of Physical Anthropology, 96: 133–145.

Oota H., Saitou N., Matsushita T., and Ueda S. (1999a) Molecular genetic analysis of remains of a 2,000-year-old human population in China — and its relevance for the origin of the modern Japanese population. American Journal of Human Genetics, 64: 250–258.

Oota H., Saitou N., Kurosaki K., Pookajorn S., Ishida T., Matsushita T., and Ueda S. (1999b) Ancient DNA: a new strategy for studying population history. In: Omoto K. (ed.), "Interdisciplinary perspectives on the origin of the Japanese," International Research Center for Japanese Studies, Kyoto, pp. 25–41.

Pääbo S., Gifford J. A., and Wilson A. C. (1988) Mitochondrial DNA sequences from a 7000-Year old brain. Nucleic Acids Research, 16: 9775–9787.

Redd A. J., Takezaki N., Sherry S. T., McGravey S. T., Sofro A. S. M., and Stoneking M. (1995) Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Molecular Biology and Evolution, 12: 604–615.

Richards M., Corte-Real H., Forster P., Macaulay V., Wilkinson-Herbot H., Demaine A., Papiha S., Hedge R., Bandelt H.-J., and Sykes B. (1996) Paleolithic and Neolithic lineages in European mitochondrial gene pool. American Journal of Human Genetics, 59: 185–203.

Saitou N. (1995) A genetic affinity analysis of human populations. Human Evolution, 10: 17–33.

Saitou N. (2000) Programs for constructing phylogenetic trees and networks of closely related sequences. In: Iwatsuki K. (ed.), "IIAS International Symposium on 'Biodiversity,'" International Institute for Advanced Studies, Kyoto, pp.45–50.

Saitou N. and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4: 406–425.

Saitou N. and Ueda S. (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequence of primate. Molecular Biology and Evolution, 11: 504–512.

Sajantila A., Lahermo P., Anttinen T., Lukka M., Sistonen, P., Savontaus M. L., Aula P., Beckman L., Tranebjaerg L., Gedde-Dahl T., Issel-Tarver L., Di Rienzo A., and Pääbo S. (1995) Genes and langua ges in Europe: an analysis of mitochondrial lineages. Genome Research, 5: 42–52.

Sherry S. T., Rogers A. R., Harpending H., Soodyall H., Jenkins T., and Stoneking M. (1994) Mismatch distributions of mtDNA reveal recent human populations expansions. Human Biology, 66: 761–775.

Shields G. F., Schmiechen A. M., Frazier B. L., Redd A., Voevoda M. I., Reed J. K., and Ward R. H. (1993) MtDNA sequences suggest a recent evolutionary divergence for Beringian and northern north American populations. American Journal of Human Genetics, 53: 549–562.

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

Sykes B., Leiboff A., Low-Ber J., Tetzer S., and Richars M. (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. American Journal of Human Genetics, 57: 1463–1475.

Thompson J. D., Higgins D. G., and Gibson T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22: 4673–4680.

Torroni A., Schurr T. G., Cabell M. F., Brown M. D., Neel J. V., Larsen M., Smith D. G., Vollo C. M., and Wallace D. C. (1993a) Asian affinities and continental radiation of the four founding Native American mtDNAs. American Journal of Human Genetics, 53: 563–590.

Torroni A., Sukernik R. I., Schurr T. G., Starikovskaya Y. B., Cabell M. F., Crawford M. H., Comuzzie A. G., and Wallence D. C. (1993b) MtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. American Journal of Human Genetics, 53: 591–609.

Turner C.G.I. (1990) Major features of Sundadonty and Sinodonty, including suggestions about East Asian miroevolution, population history, and late Pleistocene relationships with Australian Aboriginals. American Journal of Physical Anthropology, 82: 295–317.

Vigilant L. (1990) Control region sequences from African populations and the evolution of human mitochondrial-DNA. Thesis paper (PhD). Univ. of California, Berkeley.

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

Ward H. R., Frazier B. L., Dew-Jager K, and Pääbo S. (1991) Extensive mitochondrial diversity within a single Amerindian tribe. Procedings of National Academy of Sciences, USA, 88: 8720–8724.

Ward H. R., Redd A., Valencia D., Franzier B., and Pääbo S. (1993) Genetic and linguistic *differentia*tion in the Americas. Proceedings of National Academy of Sciences, USA, 90: 10663–10667.

Editor-in-Charge: Takafumi Ishida