

Sayaka Maruyama · Kiyoshi Minaguchi · Naruya Saitou

## Sequence polymorphisms of the mitochondrial DNA control region and phylogenetic analysis of mtDNA lineages in the Japanese population

Received: 22 February 2002 / Accepted: 3 April 2003 / Published online: 5 July 2003

© Springer-Verlag 2003

**Abstract** Sequence polymorphisms of the hypervariable mitochondrial DNA (mtDNA) regions HVI and HVII, and coding region polymorphisms were investigated in 211 unrelated individuals from the Japanese population. Sequence comparison of the HVI and HVII regions led to the identification of 169 mitochondrial haplotypes defined by 147 variable positions. Among them 145 types were observed in only 1 individual; the other 24 types were shared by 2 or more individuals. The gene diversity was estimated at 0.9961, and the probability of two randomly selected individuals from the population having identical mtDNA types was 0.86%. We also established phylogenetic haplogroups in the Japanese population based on the coding and control region polymorphisms and compared the haplotypes with those in other Japanese, Korean and Chinese populations. As a result, three new subhaplogroups, G4a, G4b, and N9b, and several haplotypes specific for the Japanese and Korean populations were identified. The present database can be used not only for personal identification but also as an aid for geographic or phenotype (race) estimation in forensic casework in Japan.

**Electronic Supplementary Material** Supplementary material is available for this article if you access the article at <http://dx.doi.org/10.1007/s00414-003-0379-2>

**Keywords** Mitochondrial DNA · Polymorphism · Hypervariable regions · Coding region · Phylogeny

### Introduction

Knowledge of the frequencies with which certain mitochondrial DNA sequences occur in a given population is of crucial importance for the application of mitochondrial markers to forensic studies, because mtDNA sequence types are strongly correlated with geographic origin or phenotype (race) (Forster et al. 2002). Recently, classifications of mtDNA lineages in east Asian populations using control region and coding region information have been developed (Yao et al. 2002; Kivisild et al. 2002) and a more reliable phylogenetic tree for mtDNA differentiation has been presented. These studies provide very useful information for estimating geographic origin or phenotype from biological evidence by DNA analysis.

With the increase in the number of foreign workers in Japan, the need to estimate geographic origin in forensic casework is growing. However, there is a limit to distinguishing closely related ethnic groups using morphological information alone.

In this study, to further our understanding of mtDNA polymorphisms in the Japanese population and to approach the possibility of geographic or phenotype estimation from individual biological evidence using mtDNA polymorphism data, we have carried out a population study of a large number of Japanese individuals that examined the sequence variations in the HVI and HVII region and coding region information specific for mtDNA haplogroups. We further inferred individual HVI and HVII sequence data sets from closely related populations such as other Japanese, Koreans, and Chinese, and examined differences among these populations.

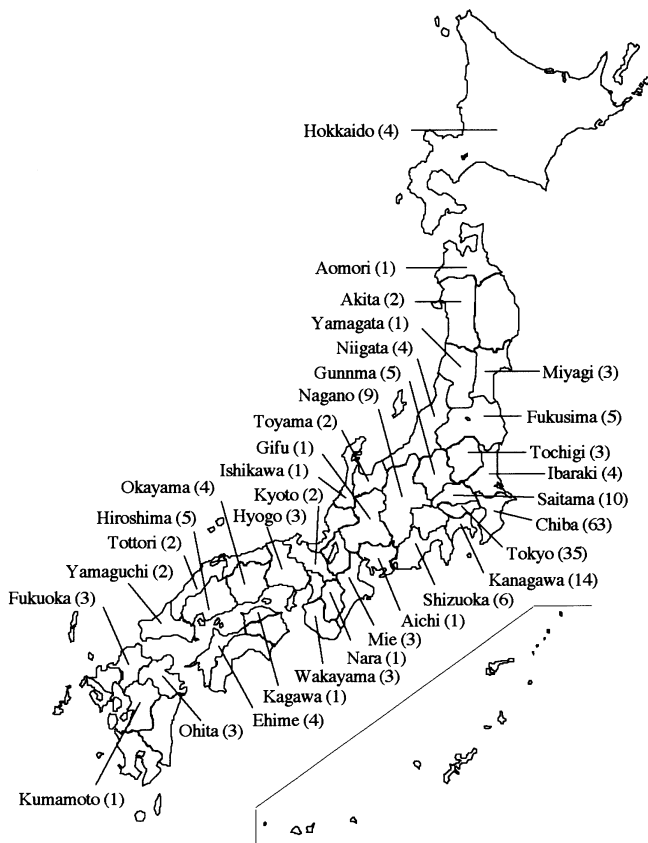
### Materials and methods

#### Samples

Genomic DNA was extracted from blood samples of 211 healthy unrelated Japanese individuals. Informed consent was obtained from the blood donors. The names of the prefectures corresponding to the birthplaces and the numbers of individuals used in this

S. Maruyama (✉) · K. Minaguchi  
Department of Forensic Odontology, Tokyo Dental College,  
1-2-2 Masago Mihama-ku, 261-0011 Chiba City, Japan  
Tel.: +81-43-2703786, Fax: +81-42-2703788,  
e-mail: smaruyam@tdc.ac.jp

N. Saitou  
Laboratory of Evolutionary Genetics,  
National Institute of Genetics,  
Mishima-shi, 411-8540 Shizuoka-ken, Japan



**Fig. 1** Geographic location and the number (in parenthesis) of the present Japanese samples. Abbreviations for prefectures in Table S1 are *HO* Hokkaido, *AM* Aomori, *AT* Akita, *YGA* Yamagata, *MG* Miyagi, *FS* Fukushima, *TG* Tochigi, *IR* Ibaraki, *NG* Niigata, *GM* Gunma, *ST* Saitama, *TK* Tokyo, *CB* Chiba, *KN* Kanagawa, *NN* Nagano, *SO* Shizuoka, *AC* Aichi, *TY* Toyama, *IK* Ishikawa, *GF* Gifu, *ME* Mie, *KT* Kyoto, *NR* Nara, *WK* Wakayama, *HG* Hyogo, *OY* Okayama, *HS* Hiroshima, *TT* Tottori, *YGU* Yamaguchi, *KG* Kagawa, *EH* Ehime, *FO* Fukuoka, *OI* Ohita, and *KM* Kumamoto

study are shown in Fig. 1. Leukocyte preparations from blood were digested with proteinase K (Sigma) at 55°C overnight, followed by treatment with RNase at 55°C for 2 h. DNA was extracted with phenol/chloroform, precipitated with ethanol, and resuspended in TE buffer (10 mM Tris-HCl, 1 mM EDTA at pH 7.6) (Sambrook et al. 1989).

#### PCR amplification and sequencing of control regions

We determined the nucleotide sequences of 211 samples from 16024 to 16398 for HVI and from 49 to 369 for HVII. We used different sized PCR products to determine the sequences of HVI and HVII. Among the 211 samples, the nucleotide sequences of 120 samples and 91 samples were determined using the different PCR products. In 120 samples, 3 partial segments (185 bp, 324 bp, and 361 bp) of the control region were amplified using the primer pairs L15978 and H16162, and L16093 and H16416 for HVI and L29 and H389 for HVII, respectively (Table 1). In 91 samples, the complete non-coding control region (1,150 bp) was amplified using the primer pair L15978 and H601. Smaller sized products are more suitable for use in actual forensic cases, however to expand the database, amplification of the whole control region minimizes the chances of making errors in the experiments, because one kind of template can be used to determine the sequences of both regions, and sequencing regions can be overlapped depending on the primers (Bandelt et al. 2001).

Each amplification was performed in a 30 µl mixture containing 10 ng of genomic DNA, 10 mM Tris-HCl at pH 8.3, 50 mM KCl, 2.5 mM MgCl<sub>2</sub>, 0.02% gelatin, 200 µM dNTP, 400 nM of each primer, and 1.5 U of AmpliTaq Gold (Perkin-Elmer Cetus). PCR amplification was done by 2-step or 3-step methods. PCR primers and PCR conditions used are shown in Table 1. Amplicons using the primer pairs L16093 and H16416 for HVI and L29 and H389 for HVII were purified for sequencing by one of the following three methods:

1. Elution from silver-stained 8% acrylamide gels using the “crush and soak” method (Sambrook et al. 1989), followed by re-amplification by PCR and purification using Centricon 100 (Amicon).
2. Purification using the QIAquick gel extraction kit (Qiagen) or the StrataPrep™ DNA gel extraction kit (Stratagene) according to the manufacturer’s instructions.
3. Purification using QIAquick PCR purification kit (Qiagen) according to the manufacturer’s instructions.

Amplicons of the complete control region using the primer pairs L15978 and H601 were purified with QIAquick PCR purification kit (Qiagen). PCR for sequencing was performed using Dye Terminator or BigDye™ Terminator cycle sequencing ready reaction kits with AmpliTaq DNA Polymerase FS (Perkin Elmer Cetus). Excess dye was removed using Centri-Sep Spin columns (Princeton) or DyeEx™ spin kits (Qiagen). Sequence analysis was performed on an ABI 373A DNA sequencer (Applied Biosystems) or an ABI 310 DNA sequencer (Applied Biosystems).

DNA sequences of partial fragments in 120 samples were determined from both strands using original primer pairs, and primers used for sequencing complete 1,193 bp fragments in 91 samples were the original PCR primer pairs, H16420, L16381, H194, and L155 (Table 1). For samples with length heteroplasmy in HVI and HVII regions, primers L16194 and L317 were used for additional sequencing. When ambiguous electrophoretograms for 20–30 nucleotides near the primers were obtained in the 29–389 fragments, complete 1,193 bp fragments were further amplified and primer L16381 was used for sequencing. While performing phylogenetic analysis, the haplogroup motif and private mutations were checked and reconfirmed by sequencing additional PCR products, even if there was little doubt of their accuracy in comparison with the sequences for other samples in the same haplogroup and the database for neighbors (Bandelt et al. 2001).

#### Typing of other polymorphisms

We examined the nucleotide variation between np 10305 and 10420 by SSCP (single strand conformation polymorphism) analysis of all of the samples (Table 1). SSCP analysis was performed in 17% polyacrylamide gels as described by Fujita and Kiyama (1995) except that the gel contained 5% glycerine, the gel and reservoir buffer were 0.475×TBE and 1×TBE, respectively, and a 16×36 cm gel with 0.4 cm thickness was used. Electrophoresis was performed at 55 V/cm constant voltage at 15°C. All the products were visualized by silver staining. When a different pattern was obtained, the PCR product was sequenced to confirm the mutation. The intergenic COII/tRNALys 9 bp deletion (Horai et al. 1996), and 3010 and 5178 mutations (Umetsu et al. 2001) were also determined from all of the samples as described in those studies. Other coding region polymorphisms specific for each lineage were selected by referring to the data of Yao et al. (2002) and Kivisild et al. (2002). Typing for these polymorphisms was performed by SSCP analysis after restriction enzyme digestion, or by direct sequencing. PCR primers, annealing temperatures, and PCR methods used for SSCP analysis are shown in Table 1. SSCP analysis and sequencing identified unexpected mutations in addition to the expected ones. Some of these unexpected mutations were novel ones for the new haplogroup. When the characteristics of sequences did not match those of any known subhaplogroups, we further extended sequencing to find coding region mutations for the new subhaplogroup. Primers to sequence other regions than those shown in Table 1 are available from the authors on request.

**Table 1** PCR primers, annealing temperature, and conditions for sequencing and SSCP analysis

Primers	Annealing temperature and PCR method <sup>a</sup>	Polymorphic sites or region	Detection method
L15978/H16162 5'-CACCATTAGCACCCAAAGCTA-3' 5'-TATGTACTACAGGTGGTCAAG-3'	58, 2-step	HVI	Sequencing
L16093/H16416 5'-CACCATTAGCACCCAAAGCT-3' 5'-TTCACGGAGGATGGTGGT-3'	52, 2-step	HVI	Sequencing
L029/H389 5'-CTCACGGGAGCTCTCCATGC-3' 5'-CTGGTTAGGCTGGTGTAGG-3'	58, 2-step	HVII	Sequencing
L15978/H601 <sup>c</sup> 5'-CACCATTAGCACCCAAAGCTA-3' 5'-CTTTGAGGAGGTAAGCTACATA-3'	58, 3-step	16024-576	Sequencing
L550/H698 5'-AACCAAACCCCAAAGACAC-3' 5'-GCATGTGTAATCTTACTAAGA-3'	54, 2-step	663	SSCP
L3143/H 3294 5'-TAAGGCCTACTTCACAAAGC-3' 5'-AGGAATTGAACCTCTGACTGT-3'	58, 2-step	3172.1, 3206	SSCP
L3143/H3419 5'-TAAGGCCTACTTCACAAAGC-3' 5'-TTGGGGCCTTTGCGTAGTT-3'	58, 2-step	3172.1, 3206, 3338, 3394	NciI-SSCP <sup>b</sup>
L4688/H4856 5'-TATCCTCTTCAACAATATACTC-3' 5'-AAGCAGGCCGGATGTCAG-3'	56, 2-step	4715, 4761, 4793, 4833	SSCP
L5109/H5253 5'-ACCGCATTCTACTACTCA-3' 5'-GCAAAAAGCCGGTTAGCG-3'	56, 2-step	5231, (5147), (5178) <sup>d</sup>	SSCP
L8211/H8310 5'-TCGTCCTAGAATTAATTCCC-3' 5'-AGTTAGCTTTACAGTGGGCT-3'	58, 2-step	9-bp Deletion	8% Native gel
L9803/H9972 5'-AGCCACAGGCTTCCACGGA-3' 5'-TGGAGACATACAGAAATAGTC-3'	58, 2-step	9824, 9923	SSCP
L10285/H10441 5'-CATGAGCCCTACAAACAAC-3' 5'-GAGTCGAAATCATTTCGTTTTG-3'	58, 2-step	10310, 10321, 10345, 10373 10362, 10397, 10398, 10400 10410	SSCP
L10285/H10682 5'-ACATAGAAAAATCCACCCCTTACG-3' 5'-TGCTTCGCAGGCGGCAAAGACTA-3'	58, 2-step	10589, 10607, 10609, 10646 (10654) <sup>d</sup>	SpeI-SSCP <sup>b</sup>

<sup>a</sup>The 2-step PCR was 95°C for 10 min, followed by 35 cycles of denaturation at 95°C for 50 s and annealing and extension at an appropriate temperature for 105 s. After the 35th cycle, a final extension step was performed at annealing and extension temperature for 10 min. The 3-step PCR was 95°C for 10 min, followed by 35 cycles of denaturation at 95°C for 1 min, annealing at 58°C for 1 min, and extension at 72°C for 3 min. After the 35th cycle, a final extension step was performed at 72°C for 10 min.

<sup>b</sup>SSCP gel electrophoresis after digestion with restriction enzyme.

<sup>c</sup>The following primers were also used for sequencing in addition to the PCR primers; L16194: 5'-ATGCTTACAAGCAAGTACAG-3', H16420: 5'-TGATTTACGGAGGAT-3', L16381: 5'-TCAGATAGGGTCCCTTGA-3', H194: 5'-GTATGTTTCGCTGTAATATT-3', and L155: 5'-TATTTATCGCACCTACGTTC-3', L317: 5'-CTTCTGGCCACAGCACTTAA-3'.

<sup>d</sup>Migration of samples with mutation sites in parentheses differed slightly from the samples with no mutations. To detect such polymorphisms, careful inspection is required.

#### Data analysis

The mtDNAs were assigned to the haplogroups according to the classification proposed by Yao et al. (2002) and Kivisild et al. (2002). Using their data and those obtained in this study, we classified the published mtDNA sequence data of other Japanese, Korean, and Chinese populations by motif search and near matching in HVI and HVII region, as described in those studies, to compare the frequency distributions of these haplogroups.

## Results and discussion

### Determination of sequences

We studied possible influences on sequencing profiles that might depend on the purification procedure of the PCR products as described in the Materials and Methods section: re-amplification from the acrylamide gel (method 1), purification from agarose gel (method 2), or direct purifi-

cation from the amplification mixture (method 3). Ambiguous electrophoretograms for 20–30 nucleotides near the primers were sometimes obtained because of high background noise, especially in cases of direct purification from the amplification mixture (method 3). Almost all of the samples purified from agarose gels and all of the samples purified by re-amplification from acrylamide gels gave satisfactory results with no background. When we compared the electrophoretograms from the same samples after purification from agarose gel and direct purification from the amplification mixture, the former gave better electrophoretograms with no background in the first 20–30 nucleotides. Because we used Centri-Sep to remove the remaining dye in all samples, ambiguous results for the background in the first 20–30 nucleotides should be attributable to the purity of the original PCR products. Similarly, sequencing of the PCR products of the complete 1,150 bp control region using original PCR primers sometimes gave electrophoretograms with high background in the first 10–30 nucleotides; however, semi-nested PCR sequencing using internal primers almost always gave satisfactory results with no background in the first 10–30 nucleotides.

### Length heteroplasmy

In HVI, the transition of the T to a C at position 16189 leads to the formation of a C-stretch, always associated

**Table 2** Distributions of the sequence variations in the region from 16180 to 16193 in HVI and from 303 to 315 in HVII

Region	Sequence variations	Number observed
HVI	AAAA CCCCCTCCCC-	Anderson sequence
	AAAA CCCCCTCCCC-	128
	AAAA CTCCCTCCCC-	2
	AAAA TCCCCTCCCC-	3
	AAAA CCCTCTCCCC-	9
	AAAA CCCTCCCCCC-	4
	AAAA CCCCCTCCCT-	1
	AAAA CCCCCTCCTC-	3
	AAAA CCCCCCTCCCCC	5
	AAAA CCCCCCCCCC-	5
	AAAA CCCCCCCCCCCC	4
	AAAC CCCCCCCCCC-	9
	AAAC CCCCCCCCCCCC	14
	AAAC CCCCCCCCCCCC	1
	AACC CCCCCCCCCCCC	6
	AACC CCCCCCCCCC-	14
	AACC CCCCCCCCCC--	3
HVII	CCCCCCC--TCCCCC-	Anderson sequence
	CCCC-----TCCCCC	1
	CCCCCC----TCCCCC	5
	CCCCCCTC--TCCCCC	1
	CCCCCCC--TCCCCC	76
	CCCCCCCC--TCCCCC	86
	CCCCCCCCC-TCCCCC	39
	CCCCCCCCCCTCCCCC	1
	CCCC-----CCCCC	2

with pronounced heteroplasmy and was observed in 27% of the individuals from the Japanese population. The number of bases in the C-stretch was determined from the L-strand by the following rule: when the height of the C peak suddenly declined and the height of the corresponding background peak attained approximately 50% of the height of the C peak, the preceding C peak was considered to be the final peak of the C-stretch. The results of the sequence variations and their distributions in the region from 16180 to 16193 are shown in Table 2. Among the 54 samples having a C-stretch without T in the sequence, 85% had 2 or 3 As preceding the homopolymeric C-tract (16180–16183). When the number of As was 4, the C repeat was 10 or 11 bp. When the number of As was 3, the length of the C repeat was mostly 12 or 11 bp. When the number of As was 2, the most common type had 12 Cs, and the second most frequent type had 13 Cs. Thus the overall length of the A+C repeat sequence was usually 14 or 15 bp. It appears that the length of this repeat is under some selective constraint, as proposed by Bendall and Sykes (1995), Howell and Smejkal (2000), and Malik et al. (2002). Although we determined the number of Cs in the C-stretches from 16184 to 16193 using the abovementioned rule, the total number of bases in the C-stretch was fixed as 10 for the actual comparison of individual sequence variations and for haplogrouping, as has been done in other similar reports.

In HVII a heteroplasmic situation often occurs with an insertion of Cs between positions 303 and 309. In this case, the length of the C-stretch was determined from the H-strands, regarding the highest T peak as the base at position 302. The patterns of sequences from 300 to 315 in the Japanese population are shown in Table 2. A pronounced heteroplasmic situation was observed in the majority of the samples with more than 9 Cs between positions 303 and 309, in 76% of the samples with 8 Cs, and in 1 sample with 7 Cs. In the samples with 8 Cs without pronounced heteroplasmy, low peaks corresponding to other groups of length variants were usually observed. Although these samples might actually be heteroplasmic, this could also be an artifact that arises during the PCR. Thus we could not clearly discriminate all the samples with and without length heteroplasmy in this region. In two samples (nos.25 and 191 in Table S1), the transition of the T to a C at position 310 leads to the formation of a C-stretch associated with a high level of length heteroplasmy. Both samples showed additional high peaks corresponding to the T peak at position 302 in the H-strand. Judging only from electrophoretograms, the major haplotype may have 11 Cs, but haplotypes with 10 and 12 Cs can also be present in high proportions in both samples. Because these two haplotypes belong to different haplogroups, this mutation may have occurred independently. It appears that the total number of Cs in this region may also be under some selective constraint, similar to the HVI C-stretch.

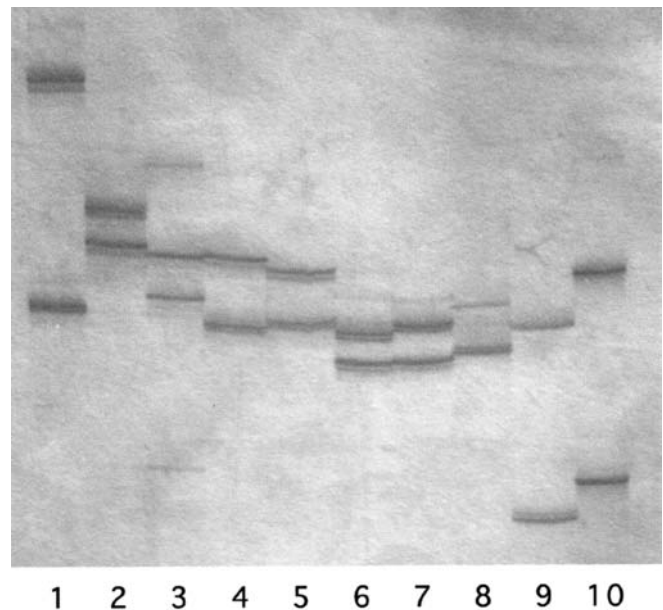
## Population data

We have determined the nucleotide sequences of the HVI and HVII control regions from 211 persons who were unrelated in their maternal lineages (Table S1). In 375 nucleotides in HVI (16024–16398), 93 (24.8%) sites were variable and in 321 nucleotides in HVII (49–369), 52 (16%) sites were variable. A total of 112 and 96 different haplotypes were observed in HVI and HVII, respectively, in which 24 haplotypes of HVI and 29 haplotypes of HVII were shared by more than 1 individual. The most common haplotypes of HVI and HVII were found in 22 (10.4%) and 25 (11.8%) individuals, respectively. Among the combinations of HVI and HVII, 169 haplotypes were identified: 145 types were unique, and the most common type was observed in 9 individuals, 1 type in 5 individuals, 2 types in 4 individuals, 4 types in 3 individuals, and 16 types in 2 individuals. Gene diversity calculated from our data using the equation  $h = n(1 - \sum x_i^2) / (n - 1)$ , (where  $n$  is sample size and  $x_i$  is the frequency of  $i$ -th mtDNA type) (Tajima 1989) was 0.9765 for HVI, 0.9685 for HVII, and 0.9961 for both regions combined. The probability of two randomly selected individuals from a population having identical mtDNA types ( $P = \sum x_i^2$ ) was 2.8% for HVI, 3.6% for HVII, and 0.86% for both regions combined.

## Phylogenetic studies

Phylogenetic differentiation of mtDNA in the Japanese population was investigated using sequence variation in HVI and HVII, and specific coding region mutations. We referred to the classification tree for the east Asian population proposed by Yao et al. (2002) and Kivisild et al. (2002). Because mutations between 10305 and 10420 provide useful information for analyzing haplogroups in Asian populations, we used SSCP analysis to detect those mutations. SSCP electrophoresis of the PCR products 10285–10441, identified 10 different types in 211 individuals (Fig. 2). Each type had characteristic sequence differences, which are informative for defining the macrohaplogroups, M and N, and some of the other subhaplogroups. Since it was not difficult to discriminate each SSCP band pattern, this method is not time-consuming when analyzing large numbers of samples. In addition, the small PCR product size is suitable for analyzing degraded DNA samples in forensic cases. Analysis of 3,010 and 5,178 mutations was also useful in determining haplogroups in the Japanese population, because haplogroup D was very common in the population, and some samples in the haplogroup G were difficult to discriminate from haplogroup D by only control region sequences.

The sequence variation of HVI and HVII and those at other polymorphic sites are shown in Table S1. All of the present samples could be placed in macrohaplogroups M or N. The present data suggest new subhaplogroups, G4a, G4b, and N9b. Altogether, we distinguished 33 haplogroups; 1 sample with a G\* haplotype was not specified further.



**Fig. 2** SSCP gel electrophoresis of the PCR products, 10285–10441. Ten types were observed. Each lane shows the type possessing the following mutations. Lane 1 10397, 10398, 10400, lane 2 10345, 10398, 10400, lane 3 10398, 10400, 10410, lane 4 10398, 10400, lane 5 10321, 10398, 10400, lane 6 no mutation, lane 7 10310, lane 8 10398, lane 9 10362, 10398, 10400, 10410, lane 10 10373, 10389, 10400

In this study 12 individuals (5.7%) shared the 16223–16325–16362–150 motif (1 lacked 150) and 4793 and 4833 mutations. Representative samples shared additional sites for the haplogroup G but lacked sites specific for other M haplogroups, including G2. Thus we designated this haplogroup as G4a. The control region motif of G4a has also been found in other Japanese (Seo et al. 1998; Imaizumi et al. 2002) and Korean samples (Lee et al. 1997; Pfeiffer et al. 1998), and in one Taiwanese Han (Tsai et al. 2001), but it was not found in Chinese Han samples (Yao et al. 2002). Two individuals (0.9%) shared the 16184–16214–16362–200 motif and 7867 and 15323 mutations in addition to specific sites for the haplogroup G except 4833, but lacked specific sites for other M haplogroups including G2. Because these individuals shared the mutation 7867 with those in haplogroup G4a, we tentatively designated this haplogroup as G4b. The control region motif of G4b was also found in one Korean sample (Pfeiffer et al. 1998). Another sample also shared coding region mutations at positions specific for haplogroup G, suggesting that this sequence belongs to a new subhaplogroup. However, because a similar sequence was found in only one sample from Korea (# P226 of Lee et al. 1997) whose coding region information had not been obtained, we indicated this haplogroup as G\*. Although 4 individuals did not share characteristic sites other than simple motifs of recurrent sites, 16189–16223 in the control region, they shared 5147 and 10607 mutations in addition to a specific site for haplogroup N. Because they lacked 14178 and 5231 mutations specific for N9a and Y, we designated this haplogroup as N9b. A similar haplotype was also found

in four Japanese individuals reported by Imaizumi et al. (2002), and one Korean (Lee et al. 1997), but was not found in Chinese (Tsai et al. 2001; Yao et al. 2002). Among these new haplogroups, G4a was also found in Ainu, and N9b-like sequences were also found in Ainu and Ryukyans (Horai et al. 1996).

Other findings in this study were as follows:

1. M10 haplotypes found in this study shared the 3172.1 mutation in addition to the 10646 mutation. This site is possibly a specific mutation of this haplogroup.
2. The 16093-16188.1-16193.1-16362-16390-146-150-152-182-217 motif was found in 4 individuals in haplogroup D5. Although this haplotype seems to belong to a new haplogroup, it was found only in the samples from Chiba prefecture. A similar HVI motif has been found in one sample from Japanese (Yoshii et al. 1995), but so far has not yet been found in other Japanese (Seo et al. 1998; Imaizumi et al. 2002), Korean (Lee et al. 1997; Pfeiffer et al. 1998), or Chinese Han populations (Tsai et al. 2001; Yao et al. 2002; Kivisild et al. 2002).
3. Most of the D4a haplotypes observed in this study shared the 10410 mutation, suggesting the presence of a subclade of D4a in the Japanese population, as indicated by Yao et al. (2002).
4. In the D4 haplotype 4 individuals shared the 16224-16245-16292-19362 motif, which also occurred in 5 out of 502 Japanese individuals in other studies (Yoshii et al. 1995; Seo et al. 1998; Nishimaki et al. 1999; Imaizumi et al. 2002; Koyama et al. 2002), in 2 Koreans (Pfeiffer et al. 1998), in 1 Han Chinese (Yao et al. 2002), and in 1 central Asian (Comas et al. 1998). In the D4 haplotype 6 individuals shared the 16245-16362-191.1-194-199-207 motif which also occurred in 3 Japanese and 1 Japanese individual in other studies (Seo et al. 1998; Imaizumi et al. 2002), and in 3 Koreans (Lee et al. 1997; Pfeiffer et al. 1998), but not in Chinese Hans (Tsai et al. 2001; Yao et al. 2002). This motif seems to define subclades of D4, although no coding region information is available at the moment.
5. Most of the individuals in the B4b1 group shared the 16284-199-202-207 motif in addition to the haplogroup specific motif 16136-16189-16217. This motif is also frequent in other Japanese populations (Seo et al. 1998; Imaizumi et al. 2002), but occurred in only one Korean (Lee et al. 1997; Pfeiffer et al. 1998), in one Chinese (Yao et al. 2002), and in two Korean individuals with similar sequences (Lee et al. 1997; Pfeiffer et al. 1998). Thus this haplotype tends to be relatively restricted to the Japanese population.
6. All of the F1b haplotypes observed in this study shared the 16129-16232-16249-16311-152 motif in addition to the haplogroup-specific motif 16189-16304-249del. The complete motif was common to almost all of the other reported Japanese populations. Identical or very similar motifs were found in Koreans (Lee et al. 1997; Pfeiffer et al. 1998) and in one Chinese (Yao et al. 2002). The sequence in three Chinese individuals (Yao et al. 2002) was slightly different from that of the Japanese; thus the complete motif tends to be restricted to the Japanese or Korean population.
7. One individual bore 12705C suggesting R status and an 16189-16217 motif suggesting B4 but did not bear the 9 bp deletion. Similar HVI motifs were found in one Korean (Lee et al. 1997) and one Japanese (Imaizumi et al. 2002), and slightly different motifs were found in one Ainu, two Ryukyans (Horai et al. 1996), and one Korean (Lee et al. 1997). Because we had not yet specified the haplogroup using coding region information, we tentatively classified this sample as R\*.

Haplogroup frequencies were compared with those of other Japanese, Korean, and Chinese populations (Table 3). The sequences of FO19 in D4 (16129, 16223, 16294, 16362, 16152, 16195, 16309.1C) and of TK24 in D4a (16129, 16223, 16362, 16152, 16195) were both very similar, and discrimination of each haplogroup was very difficult only by near matching with the previously established sequences in the control region. Thus D4 and D4a were combined as D4\* for comparison. From the study of Seo et al. (1998) 11 samples were excluded from comparisons, as well as 2 samples of Imaizumi et al. (2002), 18 samples of Lee et al. (1997), 2 samples of Pfeiffer et al. (1998), and 3 samples of Tsai et al. (2001) because we could not assign them to proper haplogroups. In the comparison under these circumstances, the most common haplogroup in the Japanese population was D4\*, which was found in 33–36% of the population. The Korean population had slightly lower frequencies than the Japanese, and Chinese had frequencies less than half of these values (12–14%). The haplogroup with the second highest frequency in the Japanese population was M7a, which included 6–14% of the population. The frequency of M7a in the Korean population was approximately 3%, and this haplogroup so far has not been observed in Chinese. The M7b1 haplotype, which is of Chinese origin, was not found in this study. The frequency of macrohaplogroup M in the Japanese population was the highest (70–74%) among the three population studied, that in the Korean population was slightly lower (65–71%), and that in the Chinese population was the lowest (52–56%). Other features observed in the Japanese and Korean population were a restricted distribution of A5, in addition to the haplogroups G4a and N9b, as mentioned above.

Consequently, Japanese, Koreans and Chinese have similar haplogroups, but the distribution of frequencies is somewhat different. Several haplogroups appear to be restricted to the Japanese and Korean populations. Moreover, other haplotypes which might be restricted to Japan and Korea were also found. These results suggest that determination of clades of the mtDNA phylogeny may help us to estimate geographic origin or phenotype (race) in forensic cases. Although we estimated most of the haplogroups from the information in the control region, there were sequences that were difficult or ambiguous to assign to specific haplogroups only by control region information. It is time-consuming to examine coding region information to determine phylogenetic clades. However, when much more data are compiled, mtDNA haplogrouping

**Table 3** Comparison of the estimated frequencies of mtDNA haplogroups in the Jaoanese, Korean and Chinese populations

Haplo-group	Estimated frequency (%)						
	Japanese-1 (n=211)	Japanese-2 (n=89)	Japanese-3 (n=160)	Korean-1 (n=288)	Korean-2 (n=58)	Taiwanese (n=152)	Chinese (n=263)
M7a1*	10.0	13.5	6.3	2.8	3.4		
M7b1				0.7		2.6	3.4
M7b2	2.8	5.6	5.6	2.8	5.2	0.7	1.5
M7b						3.9	0.8
M7c1a	0.5	1.1	0.6	1.0	5.2	0.7	0.8
M7c1b				2.1			
M7c1c						1.3	
M7c1				0.7		1.3	
M7c			0.6			1.3	1.5
M7		1.1		1.0		1.3	1.1
M8a	0.5	2.2	0.6	1.0	5.2	3.9	4.9
C	0.9		1.3	1.7		2.6	3.0
CZ				0.7		0.7	
Z	1.4	1.1		2.1		1.3	1.9
M9*	0.9	2.2	3.1	3.1	1.7	1.3	2.7
M10	1.4	1.1	1.3	4.2	8.6	2.6	1.9
M		1.1	1.9	0.7		2.6	4.2
G2	0.5		0.6			2.6	0.4
G2a	2.8	1.1	4.4	1.7	3.4	1.3	3.0
G4a	5.7	4.5	3.1	3.5	3.4	1.3	
G4b	0.9				1.7		
G*	0.5						
D5a	0.5		1.9	0.7		2.6	3.8
D5	4.7	1.1	3.1	1.0	3.4	5.3	2.7
D4b	0.5	3.4	3.8	1.4		0.7	0.8
D4*	35.5	33.7	32.5	31.9	27.9	12.5	14.2
D		1.1	0.6			1.3	0.4
A5	5.2	3.4	5.0	2.8	1.7		
A	3.8		1.9	6.9	3.4	6.6	7.2
N9a	1.9	2.2	3.8	3.5	1.7	2.6	4.6
N9b	1.9		2.5				
N*				2.1			0.4
Y	0.5			3.1	1.7	1.3	1.1
Y1			0.6				
B4a	0.5	2.2	1.3	1.4		7.9	4.6
B4b1	6.6	3.4	1.3	2.1	1.7	3.3	3.4
B4*	0.5	1.1	3.8	1.0	5.2	2.0	2.3
B5a	0.9	1.1		4.2	3.4	1.3	2.3
B5b	1.9	3.4	1.9	1.4	3.4	0.7	1.5
B5	0.5			0.3		1.3	0.4
B						0.7	1.5
R9a						2.0	1.1
F1a	0.9	3.4	1.9	1.0	3.4	5.9	5.7
F1b	4.3	3.4	4.4	0.3	1.7	2.0	3.0
F1c					1.7	0.7	1.1
F1				1.7		0.7	
F2a				0.3		1.3	1.9
F2				0.7			2.3
F		2.2		1.4		2.0	0.8
R+R10+R*	0.5		0.6	0.7		2.0	1.1
T1							0.4
HV							0.4

*Japanese-1* present study

*Japanese-2* Seo et al. (1998)

*Japanese-3* Imaizumi et al. (2002)

*Korean-1* Lee et al. (1997)

*Korean-2* Pfeiffer et al. (1998)

*Taiwanese* Tsai et al. (2000)

*Chinese* Yao et al. (2002)

M7a1 and M7a, M9 and M9a, D4 and D4a, N and N9, and B4 and B4c were combined and shown as M7a\*, M9\*, D4\*, N\*, and B4\*, respectively. For comparison with the database, 9 samples from Seo et al. (1998) (Nos.

24, 25, 31, 46, 48, 59, 63, 72, 89), 1 sample from Imaizumi et al. (2002) (JP003), and 3 samples from Lee et al. (1997) (P29, P33, H117) were excluded, because they seemed to be artefactual recombinations. Transitions 16242C/T-16248C/T and the transversion 16239C/G in the database of Seo et al. (1998) were omitted from variable positions (Bandelt et al. 2001), because we could not find any such mutations in our samples. The 11 samples and 2 samples of Koreans which bore very similar sequences to those of D4 and G3 but lacked the 16362 mutation were regarded as D4 and G3 types, respectively, 7 samples with the motif 16223-16362-150 in Koreans were regarded as D, although these sequences were similar to those of D5.

will become more useful in forensic science. We are also confident that making efforts to minimize errors in the database of the control region will increase the usefulness of mtDNA polymorphism in any research fields such as anthropology, forensic science, and medical genetics.

**Acknowledgements** This study was supported by the Grant-in-Aid for Scientific Research (B) (14370701) from the Ministry of Education, Science, Sports and Culture of Japan.

## References

- Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23:147
- Bandelt HJ, Lahermo P, Richards M, Macaulay V (2001) Detecting errors in mtDNA data by phylogenetic analysis. *Int J Legal Med* 115:64–69
- Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. *Am J Hum Genet* 57:248–256
- Comas D, Calafell F, Mateu E et al. (1998) Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *Am J Hum Genet* 63:1824–1838
- Forster P, Cali F, Rohl A et al. (2002) Continental and subcontinental distributions of mtDNA control region types. *Int J Legal Med* 116:99–108
- Fujita T, Kiyama M (1995) Identification of DNA polymorphism by asymmetric-PCR SSCP. *Biotechniques* 19:532–534
- Horai S, Murayama K, Hayasaka K et al. (1996) mtDNA polymorphism in East Asian populations, with special reference to the peopling of Japan. *Am J Hum Genet* 59:579–590
- Howell N, Smejkal CB (2000) Persistent heteroplasmy of a mutation in the human mtDNA control region: hypermutation as an apparent consequence of simple-repeat expansion/contraction. *Am J Hum Genet* 66:1589–1598
- Imaizumi K, Parsons TJ, Yoshino M, Holland MM (2002) A new database of mitochondrial DNA hypervariable regions I and II sequences from 162 Japanese individuals. *Int J Legal Med* 116:68–73
- Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R (2002) The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* 19:1737–1751
- Koyama H, Iwasa M, Maeno Y et al. (2002) Mitochondrial sequence haplotype in the Japanese population. *Forensic Sci Int* 125:93–96
- Lee SD, Shin CH, Kim KB, Lee YS, Lee JB (1997) Sequence variation of mitochondrial DNA control region in Koreans. *Forensic Sci Int* 87:99–116
- Malik S, Sudoyo H, Pramoongjago P et al. (2002) Nuclear mitochondrial interplay in the modulation of the homopolymeric tract length heteroplasmy in the control (D-loop) region of the mitochondrial DNA. *Hum Genet* 110:402–411
- Nishimaki Y, Sato K, Fang L, Ma M, Hasekura H, Boettcher B (1999) Sequence polymorphism in the mtDNA HV 1 region in Japanese and Chinese. *Legal Med* 1:238–249
- Pfeiffer H, Steighner R, Fisher R, Mornstad H, Yoon CL, Holland MM (1998) Mitochondrial DNA extraction and typing from isolated dentin – experimental evaluation in a Korean population. *Int J Legal Med* 111:309–313
- Sambrook J, Fritsch EF, Maniatis T (1989) Isolation of DNA fragment from polyacrylamide gels. *Molecular cloning: a laboratory manual*, 2nd edn. Cold Spring Harbor, New York
- Seo Y, Stradmann-Bellinghausen B, Rittner C, Takahama K, Schneider PM (1998) Sequence polymorphism of mitochondrial DNA control region in Japanese. *Forensic Sci Int* 97:155–164
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
- Tsai LC, Lin CY, Lee JCI, Chang JG, Linacre A, Goodwin W (2000) Sequence polymorphism of mitochondrial D-loop DNA in the Taiwanese Han population. *Forensic Sci Int* 119:239–247
- Umetsu K, Tanaka M, Yuasa I et al. (2001) Multiplex amplified product-length polymorphism analysis for rapid detection of human mitochondrial DNA variations. *Electrophoresis* 22:3533–3538
- Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP (2002) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* 70:635–651
- Yoshii T, Takeda E, Akiyama K, Ishiyama I (1995) Sequence polymorphism of mitochondrial DNA and its forensic application (in Japanese). *Nippon Hoigaku Zasshi* 49:242–250