

Hwan Young Lee · Ji-Eun Yoo · Myung Jin Park ·
Ukhee Chung · Kyoung-Jin Shin

Mitochondrial DNA control region sequences in Koreans: identification of useful variable sites and phylogenetic analysis for mtDNA data quality control

Received: 21 October 2004 / Accepted: 30 May 2005 / Published online: 22 September 2005
© Springer-Verlag 2005

Abstract We have established a high-quality mtDNA control region sequence database for Koreans. To identify polymorphic sites and to determine their frequencies and haplotype frequencies, the complete mtDNA control region was sequenced in 593 Koreans, and major length variants of poly-cytosine tracts in HV2 and HV3 were determined in length heteroplasmic individuals by PCR analysis using fluorescence-labeled primers. Sequence comparison showed that 494 haplotypes defined by 285 variable sites were found when the major poly-cytosine tract genotypes were considered in distinguishing haplotypes, whereas 441 haplotypes were found when the poly-cytosine tracts were ignored. Statistical parameters indicated that analysis of partial mtDNA control region which encompasses the extended regions of HV1 and HV2, CA dinucleotide repeats in HV3 and nucleotide position 16497, 16519, 456, 489 and 499 (HV1ex+HV2ex+HV3CA+ 5SNPs) and the anal-

ysis of another partial mtDNA control region including extended regions of HV1 and HV2, HV3 region and nucleotide position 16497 and 16519 (HV1ex+HV2ex+HV3+2SNPs) can be used as efficient alternatives for the analysis of the entire mtDNA control region in Koreans. Also, we collated the basic informative SNPs, suggested the important mutation motifs for the assignment of East Asian haplogroups, and classified 592 Korean mtDNAs (99.8%) into various East Asian haplogroups or sub-haplogroups. Haplogroup-directed database comparisons confirmed the absence of any major systematic errors in our data, e.g., a mix-up of site designations, base shifts or mistypings.

Keywords mtDNA · Control region · Statistical analysis · Haplogroup · Koreans

Electronic supplementary material Supplementary material is available for this article at <http://dx.doi.org/10.1007/s00414-005-0005-6> and accessible for authorised users.

H. Y. Lee · J.-E. Yoo · M. J. Park · U. Chung · K.-J. Shin (✉)
Department of Forensic Medicine and Brain
Korea 21 Project for Medical Science, College of Medicine,
Yonsei University,
134 Shinchon-Dong, Seodaemun-Gu,
120-752 Seoul, South Korea
e-mail: kjshin@yumc.yonsei.ac.kr
Tel.: +82-2-22282481
Fax: +82-2-3620860

U. Chung · K.-J. Shin
Biometrics Engineering Research Center,
Yonsei University,
134 Shinchon-Dong, Seodaemun-Gu,
120-749 Seoul, South Korea

K.-J. Shin
Human Identification Research Institute,
Yonsei University,
134 Shinchon-Dong, Seodaemun-Gu,
120-752 Seoul, South Korea

Introduction

Analysis of the human mitochondrial DNA control region has become a powerful tool for forensic identity testing. Patterns of mutations which have accumulated over time along the transmission of mtDNA lineage give rise to the differences among individuals, and have become the basis for forensic discrimination. Other properties of mtDNA that make it valuable for human identification studies include its high copy number, maternal inheritance, and the absence of recombination [1–3]. Current mtDNA forensic testing consists primarily of sequence analysis of portions of the control region, and most often targets the first and second hypervariable regions (HV1 and HV2). However, HV1/HV2 haplotypes show an uneven distribution; such mtDNA haplotypes are quite common, while others are quite rare [4]. Therefore, other regions within the control region and variations within the mtDNA coding region of forensic utility have sometimes been targeted to better differentiate mtDNA [5–8].

On the other hand, the frequencies of certain mtDNA sequences in a given population can reflect geographic or

historical origin. Recently, phylogenetic analyses based on the classifications of mtDNA lineages in East Asian populations using control region and coding region information have been reported [9–14], and a more reliable phylogenetic tree for mtDNA differentiation has been presented. A posteriori phylogenetic analysis of sequences can help to eliminate most systematic errors (e.g., a mix-up of site designations, base shifts or mistypings), and thus stringent quality control of mtDNA data by haplogroup-directed database comparisons has become available [15, 16].

In this study, we established a high-quality mtDNA control region sequence database for Koreans. We carried out a population study that examined sequence variations in the mtDNA control region, and evaluated the usefulness of various regions within the mtDNA control region by assessing the contributions of respective region to reduced random match probability. In addition, we classified mtDNA lineages (haplogroup) using control region information, and compared haplogroup frequencies with those reported by other East Asian studies.

Materials and methods

DNA samples

Buccal swabs or blood samples were obtained from 593 unrelated Koreans for the present study. DNA was extracted using a QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

PCR amplification and mtDNA sequencing

To amplify mitochondrial control region, PCR amplification was carried out using 1–2 ng of template DNA, and F15971 and R638 as primers (Table 1). Thermal cycling was conducted using a PTC-200 DNA engine (MJ Research, Waltham, MA, USA) under the following conditions: 95°C for 11 min; 35 cycles of 95°C for 30 s, 56°C for 30 s, 72°C for 90 s; and a final extension at 72°C for 7 min. The resultant amplified PCR products were purified with exonuclease I (USB, Cleveland, OH) and shrimp alkaline phosphatase (USB, Cleveland, OH, USA), and sequenced using a Big Dye Terminator Cycle Sequencing v2.0 Ready

Reaction Kit (Applied Biosystems, Foster City, CA, USA). Sequencing reactions were analyzed either using an ABI PRISM 3100 or an ABI PRISM 310 Genetic Analyzer (Applied Biosystems). Primers used for sequencing, in addition to the amplification primers, were F16328, F015, F314, R16509, R240 and R569 (see Table 1). The results were analyzed using Sequencing Analysis Software Version 3.4 and Sequence Navigator 1.01 (Applied Biosystems). To ensure sequencing data quality, a redundant approach to data generation and analysis was used. Duplicate amplifications were sequenced in both the forward and reverse directions, and the resultant consensus sequences were used for further analyses.

Identification of the major length variants of the C tracts in HV2 and HV3

To determine the major length variants of the C tracts (the major C-tract genotypes) in HV2 and HV3 region heteroplasmic individuals, PCR amplification was carried out using a previously described method [17] with 50–100 pg of template DNA and the following set of primers: F291 (5'-ATT TCC ACC AAA CCC CTC C) and R389 (5'-HEX-CTG GTT AGG CTG GTG TTA GG) for HV2 length heteroplasmy and F447 (5'-CAT TAT TTT CCC CTC CCA CTC C) and R638 (5'-FAM-GGT GAT GTG AGC CCG TCT AAA C) for HV3 length heteroplasmy. The resultant PCR products were analyzed by capillary electrophoresis on an ABI PRISM 310 Genetic Analyzer (Applied Biosystems) using GeneScan analysis software version 3.1 (Applied Biosystems).

Pairwise comparisons

Pairwise comparisons within the population were performed using the program mtDNAmanager (Shin, unpublished data), which permits hypervariable length differences in C tracts of HV2 and HV3 to be either ignored or considered in comparisons. However, length variations in the C tract of the HV1 region were ignored in all comparisons.

Genetic characteristics of the control region

The genetic diversity, random match probability and average number of nucleotide differences were calculated as described by Tajima [18], Stoneking et al. [19] and Nei [20] using the program mtDNAstar (Shin, unpublished data). Major C-tract length variants of HV2 and HV3 were either considered or ignored in terms of distinguishing haplotypes to calculate the above statistical parameters.

Haplogroup determination of Korean mtDNA

Korean mtDNA haplotypes were classified into East Asian haplogroups and sub-haplogroups based on patterns of

Table 1 Amplification and sequencing primers for mtDNA control region

Primers	Sequences
F15971	5'-TTA ACT CCA CCA TTA GCA CC-3'
F16328	5'-CGT ACA TAG CAC ATT ACA GT-3'
F015	5'-CAC CCT ATT AAC CAC TCA CG-3'
F314	5'-CCG CTT CTG GCC ACA GCA CT-3'
R16509	5'-AGG AAC CAG ATG TCG GAT ACA G-3'
R240	5'-TAT TAT TAT GTC CTA CAA GCA-3'
R569	5'-GGT GTC TTT GGG GTT TGG TTG-3'
R638	5'-GGT GAT GTG AGC CCG TCT AAA C-3'

shared haplogroup-specific or haplogroup-associated polymorphisms as reported previously using the program mtDNAmanager (Shin, unpublished data) [9–12, 14]. In addition, a phylogenetic analysis at the level of sub-haplogroups was performed using Network software version 4.1.0.8 (<http://www.fluxus-engineering.com>) [21, 22].

Results

Statistical parameters

We determined the nucleotide sequences of the complete mitochondrial DNA control region in 593 unrelated Koreans. This mtDNA haplotype database search is available at the authors' website (<http://forensic.yonsei.ac.kr/>). After

comparing complete control region sequences with the major length variants of the C tracts in HV2 and HV3 being considered in distinguishing haplotypes, 494 haplotypes (83.3%) defined by 285 variable sites were found. Among these, 424 sequences (71.5%) were observed only once, 50 twice (16.9%), 15 three times (7.6%), 2 four times (1.3%), 2 five times (1.7%) and 1 six times (1.0%). The most frequently observed haplotype was 16187T, 16223T, 16290T, 16319A, 73G, 235G, 263G, 309.1C, 315.1C, 523d, 524d in comparison to revised Cambridge reference sequence (rCRS).

In the case of the major C-tract genotypes ignored for distinguishing haplotypes, 441 sequences (74.4%) were observed. From these, 362 haplotypes (61.0%) were observed once, 43 twice (14.5%), 17 three times (8.6%), 10 four times (6.7%), 6 five times (5.1%), and the other 3 sequences were observed six, eight, and ten times, re-

Table 2 Genetic diversity, random match probability, average number of nucleotide differences and observed numbers of haplotypes in mtDNA of 593 Koreans

	Genetic diversity		Random match probability (%)		Average number of nucleotide differences		Observed number of haplotypes	
	Neglecting poly-C ^a	Considering poly-C ^a	Neglecting poly-C ^a	Considering poly-C ^a	Neglecting poly-C ^a	Considering poly-C ^a	Neglecting poly-C ^a	Considering poly-C ^a
HV1	0.9920	–	0.96	–	6.92	–	320	–
HV2	0.9429	0.9785	5.87	2.31	2.74	3.52	155	212
HV3	0.7666	0.7981	23.47	20.32	1.78	2.18	33	40
7S-SP ^b	0.6251	–	37.60	–	0.82	–	36	–
HV1ex ^c	0.9925	–	0.92	–	7.04	–	330	–
HV2ex ^c	0.9463	0.9794	5.53	2.23	2.83	3.61	164	223
HV3ex ^c	0.7723	0.8035	22.90	19.78	1.82	2.22	40	47
HV1+HV2	0.9967	0.9985	0.50	0.32	9.66	10.44	408	466
HV1+HV2+HV3	0.9975	0.9989	0.42	0.28	11.43	12.62	420	477
HV1ex+HV2ex	0.9970	0.9986	0.47	0.31	9.87	10.65	419	473
HV1ex+7S-SP+HV2ex	0.9979	0.9990	0.38	0.27	10.48	11.26	428	483
HV1ex+HV2ex+HV3CA ^d	0.9977	0.9989	0.40	0.28	10.93	11.71	424	477
HV1ex+HV2ex+HV3	0.9977	0.9990	0.39	0.27	11.64	12.83	430	483
HV1ex+HV2ex+5SNPs ^e	0.9978	0.9990	0.38	0.27	11.06	11.84	428	482
HV1ex+HV2ex+HV3CA+5SNPs	0.9982	0.9991	0.35	0.25	12.13	12.91	434	488
HV1ex+HV2ex+HV3+2SNPs ^f	0.9982	0.9992	0.35	0.25	12.22	13.40	438	491
Complete control region	0.9983	0.9992	0.34	0.25	12.30	13.48	441	494

^aStatistical parameters were calculated by either neglecting or considering the major poly-C tract genotypes in HV2 and HV3 regions. The major C-tract genotype of HV1 heteroplasmy were not considered here because it could not be determined using the present method

^bThe 7S-SP region is situated between n.p. 16366 and 72

^cThe extended regions of HV1, HV2, and HV3 refer to the regions between nucleotide position (n.p.) 16024 and 16400, n.p. 41 and 340, and n.p. 341 and 576, respectively

^dThe CA dinucleotide repeat polymorphism in the HV3 region was considered in distinguishing haplotypes

^eFive SNP sites include n.p. 16497, 16519, 456, 489 and 499, all of which showed transition rates of 3% or more

^fTwo SNP sites include n.p. 16497 and 16519 in the 7S-SP region

spectively (1.0, 1.3, and 1.7%, respectively). The most abundant haplotypes (16187T, 16223T, 16290T, 16319A, 73G, 235G, 263G, 523d, 524d and 16223T, 16362C, 73G, 263G, 489C) were observed in 3.0% of the samples.

To better categorize the data obtained, several terms were used, as described previously [8]. The region spanned by 7S DNA during the replication of the H strand, situated between nucleotide positions (n.p.) 16366 and 72, was called 7S DNA spanned region (7S-SP), and the region between n.p. 341 and 576 containing HV3 was called the HV3 extended region (HV3ex). In addition, we used the HV1 extended region (HV1ex) and the HV2 extended region (HV2ex), which refer to the regions spanning n.p. 16024 and 16400, and n.p. 41 and 340, respectively.

Based on the observed frequencies of the mtDNA control region haplotypes and the data from partial areas of this region (i.e., HV1, HV2, HV3, 7S-SP, HV1ex, HV2ex, HV3ex, HV1+HV2, HV1+HV2+HV3, HV1ex+HV2ex, HV1ex+7S-SP+HV2ex, HV1ex+HV2ex+HV3CA, HV1ex+HV2ex+HV3, HV1ex+HV2ex+5SNPs, HV1ex+HV2ex+HV3CA+5SNPs, HV1ex+HV2ex+HV3+2SNPs), genetic diversity, random match probability, the average number of nucleotide differences, and the observed number of haplotypes were calculated (Table 2). Here, HV3CA refers to the CA dinucleotide repeats in HV3, five SNPs refer to n.p. 16497, 16519, 456, 489, and 499 located in 7S-SP and HV3ex with 3.0% or more transition rates, and two SNPs refer to n.p. 16497 and 16519 in 7S-SP.

The genetic diversities of HV1+HV2, HV1ex+HV2ex+HV3 and the complete control region were estimated to be 0.9967, 0.9977, and 0.9983, respectively. Consideration of the major length variants of the HV2 and HV3 C tracts increased these genetic diversities to 0.9985, 0.9990, and 0.9992, respectively.

The probability of a random match between two unrelated individuals was calculated to be 1:200 (0.50%) for HV1+HV2, 1:256 (0.39%) for HV1ex+HV2ex+HV3, and 1:294 (0.34%) for the complete control region. When considering the major C-tract genotype in length heteroplasmy, this parameter decreased to 1:313 (0.32%), 1:370 (0.27%), and 1:400 (0.25%) for HV1+HV2, for HV1ex+HV2ex+HV3, and for the complete control region, respectively.

The average number of nucleotide differences increased as the combined partial control regions were expanded, up to 12.30, and in the case of the major C-tract genotype of length heteroplasmy considered in distinguishing haplotypes, this increased to 13.48.

Also, the observed number of haplotypes increased when regions were added to the calculation. The different sequences were counted as 408, 430, 438 and 440 for HV1+HV2, HV1ex+HV2ex+HV3, HV1ex+HV2ex+HV3+ 2SNPs, and the complete control region, respectively. The addition of the major C-tract length variant to the calculation resulted in 466, 483, 491, and 494 haplotypes for HV1+HV2, HV1ex+HV2ex+HV3, HV1ex+HV2ex+HV3+2SNPs, and the complete control region, respectively.

Variable sites in the mtDNA control region

The observed mutational events including nucleotide substitutions, insertions, and deletions compared to rCRS are shown in Table 3. The most polymorphic region was HV1, which contained 142 variable positions (49.82%), whereas the HV2 region showed 79 mutational sites (27.72%). In addition, 7S-SP and HV3ex displayed 29 and 35 variable positions, respectively (10.18 and 12.28%, respectively).

Substitutions were observed at 246 positions with a total of 5,487 differences. Of 246 substitution sites, transition and transversion were observed in 236 and 28 sites, respectively. Substitution sites that showed both transition and transversion reached up to 18. Of the 5,487 differences, 5,164 occurrences (94.1%) were transitions, mainly of T to C and C to T (63.0%). Transversions were relatively rare (323 occurrences; 5.9%), but A to C and C to A predominated in the HV1 region (93.5%). The high rate of A to C transversion in Koreans (236 occurrences) was mainly due to a change at n.p. 16182 (71 individuals) or at 16183 (143 individuals).

Deletions were found at 21 sites with a total of 318 occurrences mainly in the HV3ex and HV2 regions. In particular, CA dinucleotide deletion at n.p. 523 and 524 in the HV3 region was observed in 219 individuals.

Insertions included two types of nucleotide addition, i.e., homoplasmic and heteroplasmic. Homoplasmic insertions were found mainly in HV2 (593 occurrences) and in HV3ex regions (17 occurrences). The high frequencies of homoplasmic insertion mutations in the HV2 and HV3ex regions were mainly due to the insertion of nucleotide C at n.p. 315 (589 individuals) and CA dinucleotide insertion at n.p. 524 (15 individuals). Other insertion sites with relatively high frequencies included 309.1 (375 individuals) and 573.1 (38 individuals), which were often found to be related to length heteroplasmy.

Length heteroplasmy also occurred in the poly-C tract of the HV1 region with a T-to-C transition at n.p. 16189 and in CA dinucleotide repeats of the HV3 region. One hundred eighty-eight individuals (31.7%) and 377 individuals (63.6%) showed length heteroplasmy in the poly-C tract of the HV1 and HV2 regions, respectively. In addition, three individuals (0.5%) and 39 individuals (6.6%) displayed length heteroplasmy in the CA dinucleotide repeats and the poly-C tract of the HV3 region, respectively. Point heteroplasms were observed in 14 individuals; at 16093T/C twice, 16103A/G once, 16129G/A once, 16172T/C once, 16176C/T once, 16342T/C once, 146T/C once, 152T/C three times, 194 T/C once, 195T/C once and 204T/C once.

In addition, all mutable nucleotide positions were ranked into three categories according to the observed mutation frequencies. In Table 3, data in italics indicate all mutable sites observed in one to five individuals (<1%). Underlined data represent mutable sites observed in six to 17 individuals (<3%), and those in bold indicate mutable sites in >18 individuals (>3%). The nucleotide positions that

Table 3 Mutable sites observed in the sequence analysis of mtDNA control region with respect to revised Cambridge reference sequence

rCRS	Korean	Number	rCRS	Korean	Number	rCRS	Korean	Number	rCRS	Korean	Number
16046	T C	1	16221	C T	1	16360	C T	2	211	A G	1
16048	G A	1	16222	C T	1	16362	T C	274	214	A G	7
16051	A G	10	16223	C T	433	16362	T A	1	215	A G	6
16066	A G	7	16224	T C	6	16365	C T	1	217	T C	1
16072	C T	1	16227	A G	15	16366	C T	1	222	C T	1
16075	T C	5	16231	T C	13	16368	T C	2	228	G A	8
16078	A T	1	16232	C T	2	16380	C T	2	234	A G	3
16086	T C	10	16232	C A	13	16390	G A	17	235	A G	52
16092	T C	11	16234	C T	20	16391	G A	2	236	T C	1
16092	T A	2	16235	A G	4	16399	A G	9	237	A G	2
16093	T C	27	16239	C T	2	16400	C T	1	242	C T	2
16094	T C	1	16240	A G	2	16428	G A	2	247	G del	2
16102	T C	1	16241	A G	1	16463	A G	1	249	A G	1
16104	C T	2	16242	C T	1	16482	A G	1	249	A del	65
16111	C T	14	16243	T C	20	16497	A G	24	252	T C	2
16126	T C	14	16244	G A	1	16519	T C	306	257	A G	4
16129	G A	111	16245	C T	8	16523	A G	2	260	G A	2
16131	T C	2	16248	C T	2	16524	A G	2	263	A G	587
16134	C T	2	16249	T C	24	16524	A C	1	279	T C	1
16136	T C	23	16250	C T	3	16527	C T	2	281	A G	1
16140	T C	28	16256	C T	12	41	C T	1	284	A G	1
16145	G A	4	16257	C T	2	44.1	ins	2	290	A del	1
16148	C T	11	16257	C A	43	47	G A	2	291	A del	1
16148	C G	1	16258	A G	1	56	A del	2	298	C T	1
16150	C T	7	16258	A C	1	58	T A	2	307	C T	2
16154	T C	3	16259	C T	2	61	C T	1	309	C del	3
16158	A G	2	16260	C T	8	62	G A	1	309.1	ins	375
16162	A G	13	16261	C T	66	63	T C	4	309.2	ins	115
16164	A G	10	16262	C T	1	64	C T	6	309.pC ^a	ins	1
16166	A G	2	16263	T C	3	65	T G	1	310	T del	4
16166	A C	1	16264	C T	2	66	G A	2	311	C del	3
16167	C T	4	16265	A C	6	68	G A	1	315.1	ins	589
16168	C T	3	16266	C T	31	71	G A	1	316	G A	5
16169	C T	1	16266	C G	4	71.1	ins	2	318	T C	4
16170	A G	2	16266	C A	5	73	A G	590	326	A G	4
16171	A G	2	16268	C T	7	73	A C	3	333	T C	1
16172	T C	74	16269	A G	1	75	G A	1	334	T C	2
16173	C T	2	16270	C T	3	89	T C	3	338	C T	2
16174	C T	10	16271	T C	4	93	A G	10	368	A G	1
16175	A G	1	16274	G A	18	94	G A	6	373	A G	3
16176	C T	3	16278	C T	34	103	G A	17	373	A T	1
16178	T C	1	16284	A G	14	106	G del	1	376	A G	1
16179	C A	6	16287	C T	3	107	G del	1	385	A G	2
16181	A C	1	16288	T C	15	108	A del	1	408	T A	2
16182	A C	71	16290	C T	55	109	G del	1	431	C T	2
16182	A T	4	16291	C T	10	110	C del	1	455.1	ins	1
16183	A G	3	16292	C T	4	111	A del	1	456	C T	18
16183	A C	143	16293	A G	5	131	T C	1	459	C del	1
16184	C T	15	16293	A C	2	143	G A	11	459.1	ins	1
16185	C T	9	16294	C T	15	146	T C	60	460	T C	2
16186	C T	4	16295	C T	19	150	C T	127	466	T C	3
16186	C A	1	16296	C T	2	151	C T	13	471	T C	2
16187	C T	28	16297	T C	24	152	T C	126	482	T C	3
16188	C T	2	16298	T C	56	153	A G	6	489	T C	346

Table 3 (continued)

rCRS	Korean	Number	rCRS	Korean	Number	rCRS	Korean	Number	rCRS	Korean	Number
16188.1	<i>ins</i>	1	16299	A <i>G</i>	2	178	A <i>G</i>	1	493	A <i>G</i>	1
16189	T C	199	16300	A <i>G</i>	2	179	T <i>C</i>	1	499	G A	22
16189	T <i>del</i>	4	16302	A <i>G</i>	2	182	C <i>T</i>	1	502	C <i>A</i>	1
16192	C <u>T</u>	11	16303	G <i>A</i>	1	183	A <u>G</u>	10	509	C <i>T</i>	1
16193	C <i>T</i>	3	16304	T C	49	184	G <i>A</i>	1	513	G <i>A</i>	1
16193.1	<i>ins</i>	1	16309	A <i>G</i>	3	185	G <u>A</u>	8	513	G <u>del</u>	6
16193.pC ^a	ins	188	16311	T C	74	188	A <i>G</i>	4	514	C <i>T</i>	3
16194	A <i>G</i>	4	16316	A <u>G</u>	13	189	A G	20	514	C <u>del</u>	6
16194	A <u>C</u>	6	16317	A <i>G</i>	3	191.1	<i>ins</i>	4	521	A <i>del</i>	1
16195	T <u>C</u>	7	16318	A <i>T</i>	1	193	A <i>G</i>	2	523	A del	220
16201	C <i>T</i>	1	16319	G A	85	194	C T	29	524	C del	219
16203	A <i>G</i>	1	16320	C <i>T</i>	1	195	T C	39	524.1	<u>ins</u>	11
16204	G <i>A</i>	1	16324	T <u>C</u>	9	197	A <i>G</i>	1	524.2	<u>ins</u>	11
16206	A <i>G</i>	1	16325	T C	18	198	C <i>T</i>	2	524.3	<i>ins</i>	4
16206	A <i>T</i>	1	16327	C <u>T</u>	15	199	T C	61	524.4	<i>ins</i>	4
16209	T <u>C</u>	15	16335	A <i>G</i>	2	200	A G	23	533	A <i>G</i>	1
16213	G <i>A</i>	5	16342	T <i>C</i>	1	202	A <u>G</u>	9	546	A <i>G</i>	2
16214	C <i>T</i>	5	16344	C <u>T</u>	10	203	G <u>A</u>	8	548	C <i>T</i>	1
16215	A <i>G</i>	1	16354	C <i>T</i>	2	204	T C	33	573.1	ins	38
16216	A <i>G</i>	4	16355	C <u>T</u>	6	205	G <i>A</i>	4	573.2	ins	30
16217	T C	68	16355	C <i>A</i>	1	207	G A	27	573.3	ins	30
16218	C <i>T</i>	5	16356	T <i>C</i>	2	209	T <i>C</i>	1	573.pC ^a	ins	28
16220	A <i>C</i>	1	16357	T <u>C</u>	13	210	A <u>G</u>	9			

Data in italics indicate mutable sites in one to five individuals (<1%). Underlined data represent mutable sites in six to 17 individuals (<3%), and those in bold indicate mutable sites in more than 18 individuals (>3%)

^apC indicates insertion of multitudes of cytosines, which generally makes it difficult to determine the major length variant of homopolymeric C tract

showed high mutation percentages in Koreans (>20%) are 16183C (24.1%), 16189C (34.2%), 16193.pC (31.7%), 16223T (73.0%), 16362C (46.2%), 16519C (51.6%), 73G (99.5%), 152C (21.6%), 263G (99.0%), 309.1 (63.2%), 315.1 (99.3%), 489C (58.3%), 523d (37.1%) and 524d (36.9%).

haplogroups include D4a (5.1%), D4b2 (3.4%), M7b2 (3.7%), B4 (4.0%), B4a (3.7%), B4b1 (3.4%), B5b (3.0%), A4 (3.0%), A5 (3.2%) and N9a2 (4.7%).

Haplogroup determination and haplogroup distribution

We classified 592 Korean mtDNAs (99.8%) into East Asian haplogroups or sub-haplogroups by referring to the East Asian classification tree proposed by Yao et al. [9], Kivisild et al. [10], Kong et al. [11, 12] and Tanaka et al. [14]. The phylogenetic backbone used was in accord with Yao et al. [9], Kivisild et al. [10] and Kong et al. [11, 12], and the haplogroups D, G and N9 were subdivided by referring to the classification tree proposed by Tanaka et al. [14]. The basic informative SNPs in the control region used to identify haplogroups are listed in Table 4 (for more details, see ESM).

All of the present samples could be placed in macro-haplogroups M or N, where 489C in HV3 defined the haplogroup M. The observed haplogroup frequencies are shown in Fig. 1. The D4* haplogroup was observed at the highest frequency (15.7%), and other commonly observed

Discussion

All statistical parameters favors entire control region analysis and the consideration of the major C-tract genotype for distinguishing haplotypes. The low values of random match probability and high values for the average number of nucleotide differences were observed across the entire mtDNA control region in Koreans. However, the samples recovered from crime scenes are often highly degraded, which leads to poor PCR amplification of the larger-sized complete mtDNA control region. Therefore, the targeting and preferential amplification of mtDNA using small amplicon products and the selective scoring of highly informative variable sites is likely to be a more effective method for mtDNA analysis. The contributions of partial areas in the mtDNA control region to reduced random match probability can be assessed by comparing the values obtained from various combinations of respective areas. Statistical parameters indicated that the analysis of HV1ex+HV2ex is more effective than that of HV1+HV2, and that the analyses of HV1ex+HV2ex+HV3+2SNPs and HV1ex+HV2ex+HV3CA+5SNPs are as efficient as entire control

Table 4 List of important nucleotide positions that identify East Asian mtDNA haplogroups and haplogroup frequencies in Koreans

Haplogroup	HV1 ^a	HV2 ^a	HV3, etc. ^a	Sample
D4*	16223–16362		489	93
D4a	16129–16223–16362	152	(16519)–489	30
D4b	16223–16319–16362		489–523d–524d	4
D4b1	16223–16319–16362	152	489–523d–524d	6
D4b2	16223–16362	194	16519–489–523d–524d	20
D5	16189–16223–16362	150	489	3
D5a	16182Y–16183–16189–16223–16266–16362	150	489–523d–524d	12
D5b	16189–16223–16362	150	456–489	16
D*	16362		489	7
G1a	16223–16325–16362	150	16519–489	14
G1b	16184–16214–16223–16362		489	2
G2a1	16189–16223–16278–16362		489	4
G2a1a	16223–16227–16278–(16362)		489	15
G2a2	16051–16150–16223–16278–16362		489	6
G3a	16223–16274–16362	143–152	489	10
M7a	16209–16223		489	1
M7a1	16209–16223–16324		489–(523d–524d)	8
M7b1	16129–16192–16223–16297	150–199	489	1
M7b2	16129–16189–16223–16297–16298	150–199	489	22
M7c	16223	146–199	16519–489–523d–524d	1
M7c1	16223–16295	(146)–199	16519–489–523d–524d	15
M8	16223–16298		489	1
CZ	16223–16298	249d	489	1
C	(16223)–16298–16327	249d	16519–489	15
Z	16185–16223–16260–16298	152–249d (or 247d) ^b	489	5
M8a	16184–16223–16298–16319		489	6
M9a	16223–16234–16316–16362		489	10
M10	16223–16311		489–573.pC	9
M11	16223	215–318–326	489	4
M*	16223		489	5
F	16304	249d		3
F1	16304	249d	16519–523d–524d	6
F1ac	16129–16304	249d	16519–523d–524d	2
F1a	(16129)–16162–16172–(16304)	249d	(16519)–523d–524d	13
F1c	16111–16129–16304	152–249d	16519–523d–524d	2
F1b	(16182)–16183–16189–16304	249d	(16519)–523d–524d	16
F2a	16291–16304	249d		3
R11	16189–16311	185–189		2
B4	16183–16189 (or 16189d) ^b –16217			24
B4a	16182–16183–16189–16217–16261		(16519)–523d–524d	22
B4b1	16136–16183–16189–16217		16519–499	20
B5a	16129–16140–16187–16189–16266R	93–210	16519–523d–524d	9
B5b	16140–16183–16189–16243		16519–523d–524d	18
A	16223–16290–16319	235		13
A4	16223–16290–16319–16362	235	523d–524d	18
A5	16187–16223–16290–16319	235	523d–524d	19
N9a	16223–16257A–16261	150		4
N9a1	16223–16129–16257A–16261	150		11
N9a2	16172–16223–16257A–(16261)	150		28

Table 4 (continued)

Haplogroup	HV1 ^a	HV2 ^a	HV3, etc. ^a	Sample
N9b	16182–16183–16189–16223		16519	2
Y	16126–16231–16311		482	2
Y1	16126–16231–16266	146	16519	9
ND				1

ND Not determined

^aFor each haplogroup, 73, 263, 315.1 and heteroplasmic insertions were ignored in haplogroup classification; suffixes A, Y and R indicate transversion and d indicates deletion; exceptionally, all 16182 and 16183 indicate A-to-C transversion mutations; mutation sites which are missed once in a given haplogroup of the present study are given in parentheses. Haplogroups, which seem to lack control region information for exact haplogroup determination, are marked by an asterisk

^bUnpublished data

region analysis (Table 2). Accordingly, the analyses of HV1ex+HV2ex+HV3+2SNPs and HV1ex+HV2ex+HV3CA+5SNPs can be used as efficient alternatives for the analysis of the entire mtDNA control region in Koreans. This seems to be mainly due to the relatively high genetic diversity of the CA dinucleotide repeats in the HV3 region, and the relatively high number of n.p. 16497 and 16519 changes in the Korean population. These findings imply that the analysis of some variable sites outside HV1 and HV2 can play an important role in the forensic discrimination in certain populations.

On the other hand, the cumbersome determination of the major C-tract genotype by PCR amplification with fluorescence labeled primers limits the use of length heteroplasmy in forensic discrimination in spite of their capability to increase the power of discrimination. Similarly, point heteroplasmy has some defects in a point that the observed number of point heteroplasmy tends to depend on the detection methods, and the majority of heteroplasmic variants can not be detected by direct sequencing of PCR products [23]. However, 14 point heteroplasmy (11 heteroplasmic

sites) were observed from direct sequencing of the mtDNA control region in the present study, and n.p. 16093 and n.p. 152, which showed point heteroplasmy in two and three individuals, respectively, were found to be one of most frequently observed heteroplasmic positions as described in previous reports [23–25].

Also, we classified Korean mtDNA into East Asian haplogroups (Table 4 and ESM). Reference to a worked-out phylogeny and comparisons with published data sets of East Asian populations were performed to avoid potential errors [9, 10, 14]. The basic informative SNPs in the control region used to identify haplogroups in the present study are similar to those used by Allard et al. [13] and Yao et al. [15], and more than 99% of the Korean mtDNA lineages can be allocated to specific mtDNA haplogroups according to their mutation motifs (Table 4). In addition to the HV1 and HV2 mutation motifs for haplogroup assignment, we collated the important mutational positions in HV3ex and 7S-SP. We also added mtDNA control region mutation motifs for the assignment of 20 another haplogroups and sub-haplogroups to those of Allard et al. [13] and Tanaka

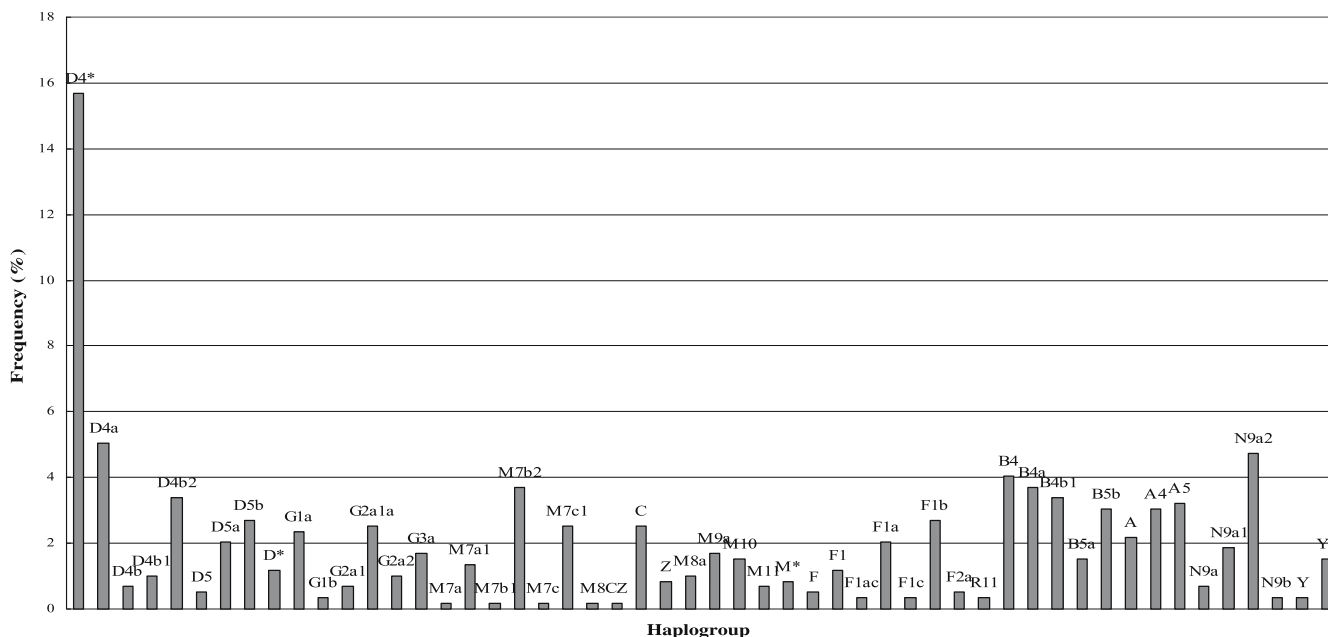


Fig. 1 Haplogroup frequencies observed in 593 unrelated Koreans. Number and letters refer to the haplogroups defined in the text

et al. [14]: D4b1, D4b2, D5b, D*, G1b, G2a1, G2a1a, G2a2, G3a, M7a, M8, M*, CZ, F1, F1ac, A4, A5, N9a1, N9a2 and N9b.

As compared with the mutation motifs presented by Yao et al. [15], we did not include 73G and 263G in all mutation motifs to define haplogroups because of their lack of discriminative power. Also, unlike Yao et al. [15], the A haplogroup mutation motif in this study did not include 152C. The site 152C was found to be variable within this haplogroup, which concurs with the findings of Allard et al. [13]. On the other hand, the M10 haplotypes in this study did not share 16519C, in contrast to the results of Allard et al. [13] and in agreement with the report of Yao et al. [15]. The F1a and F1b haplotypes found in this study shared the 16162G and 16182C–16183C mutations, respectively, in addition to the mutation motifs proposed by Yao et al. [15]. B4a haplotypes were found to share 16182C mutation in addition to the mutation motif proposed by Allard et al. [13]. Both B4 and B4b1 haplotypes in this study were also found to share 16183C in addition to the mutation motifs described by Yao et al. [15], and the 16189d mutation observed in B4 in this study occasionally alternated with 16189C (confirmed by unpublished data). 247d mutation in the Z haplogroup also alternated with 249d (confirmed by unpublished data). In addition to the mutation motif proposed by Allard et al. [13] and Yao et al. [15], B5a haplotypes shared 16129A–16187T–16519C–93G mutations. These sites possibly represent specific mutations of this haplogroup. In addition, the sequence 16093C–16188.1C–16193.1C–16362C–16390A–146C–150T–152C–182T–217C, which was reported to be found in some Japanese individuals in haplogroup D5, was observed in one Korean individual, and was assigned to haplogroup D5, as described by Maruyama et al. [26]. Also, in reference to Kong et al. [12], G4a in Maruyama et al. [26] corresponds to G1a in the present study. The distribution pattern of Korean mtDNA haplogroup frequencies generally parallels to that of the Japanese [13, 26], but showed slight differences versus that of the Chinese [9, 13]. The D4* haplogroup occurred at highest frequency in Koreans (15.7% in this study, 16.5% according to Allard et al. [13], and 31.9% by Maruyama et al. [26]) and in Japanese (19.6% Allard et al. [13] and 35.5% Maruyama et al. [26]), and was also common in Chinese (6.2% Allard et al. [13] and 14.2% Maruyama et al. [26]). However, the G haplogroup and its sub-haplogroups were observed in relatively high frequencies in Koreans (8.6% in the present study and 5.2% by Maruyama et al. [26]) and Japanese populations (10.4% Maruyama et al. [26]), but occurred sparsely in the Chinese (4% Allard et al. [13] and 3.4% Maruyama et al. [26]). Differences between the haplogroup frequencies of populations in the same geographic location shown in previous reports and in the present study are considered to be mainly due to different sample collection and classification categories. In addition, some haplogroups defined on the basis of very small number of mutation sites in control region might make their precise haplogroup determination difficult, occasionally leading to different distribution of haplogroups.

In conclusion, all statistical and phylogenetic results and values in the present study indicate that our mtDNA data are relatively well suited for application to forensic casework and that they contribute to a better characterization of Korean mtDNA types. Our examination of sequence variations in the mtDNA control region, and the evaluation of the usefulness of various partial areas within the mtDNA control region in terms of reducing random match probability, suggest that variable sites lying outside HV1 and HV2 (e.g., HV3 and five highly informative SNPs, n.p. 16497, 16519, 456, 489 and 499) can play an important role in a better differentiation of Korean mtDNA. Also, stringent quality control of mtDNA data by haplogroup-directed database comparisons confirmed the absence of major systematic errors in our data (e.g., a mix-up of site designations, base shifts, or mistypings). However, the high frequencies of haplogroup D4* in the Korean and Japanese populations, but not in the Chinese population, suggest the need for finer subdivisions within this haplogroup to better characterize Korean and Japanese mtDNA lineages.

Acknowledgements This work was supported by a grant from the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

References

- Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A* 77:6715–6719
- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290:457–465
- Holland MM, Parsons TJ (1999) Mitochondrial DNA sequence analysis—validation and use for forensic casework. *Forensic Sci Rev* 11:21–50
- Coble MD, Just RS, O’Callaghan JE, Letmanyi IH, Peterson CT, Irwin JA, Parsons TJ (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int J Leg Med* 118:137–146
- Lutz S, Wittig H, Weisser HJ, Heizmann J, Junge A, Dimo-Simonin N, Parson W, Edelmann J, Anslinger K, Jung S, Augustin C (2000) Is it possible to differentiate mtDNA by means of HVIII in samples that cannot be distinguished by sequencing the HV1 and HVII regions? *Forensic Sci Int* 113: 97–101
- Parsons TJ, Coble MD (2001) Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. *Croat Med J* 42:304–309
- Lee SD, Lee YS, Lee JB (2002) Polymorphism in the mitochondrial cytochrome B gene in Koreans. An additional marker for individual identification. *Int J Leg Med* 116:74–78
- Vanecek T, Vorel F, Sip M (2004) Mitochondrial DNA D-loop hypervariable regions: Czech population data. *Int J Leg Med* 118:14–18
- 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
- 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

11. Kong QP, Yao YG, Liu M, Shen SP, Chen C, Zhu CL, Palanichamy MG, Zhang YP (2003) Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. *Hum Genet* 113:391–405
12. Kong QP, Yao YG, Sun C, Bandelt HJ, Zhu CL, Zhang YP (2003) Phylogeny of east Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet* 73:671–676
13. Allard MW, Wilson MR, Monson KL, Budowle B (2004) Control region sequences for East Asian individuals in the Scientific Working Group on DNA Analysis Methods forensic mtDNA data set. *Leg Med (Tokyo)* 6:11–24
14. Tanaka M, Cabrera VM, Gonzalez AM, Larruga JM, Takeyasu T, Fuku N, Guo LJ, Hirose R, Fujita Y, Kurata M, Shinoda K, Umetsu K, Yamada Y, Oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res* 14:1832–1850
15. Yao YG, Bravi CM, Bandelt HJ (2004) A call for mtDNA data quality control in forensic science. *Forensic Sci Int* 141:1–6
16. Bandelt HJ, Salas A, Lutz-Bonengel S (2004) Artificial recombination in forensic mtDNA population databases. *Int J Leg Med* 118:267–273
17. Lee HY, Chung U, Yoo JE, Park MJ, Shin KJ (2004) Quantitative and qualitative profiling of mitochondrial DNA length heteroplasmy. *Electrophoresis* 25:28–34
18. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
19. Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am J Hum Genet* 48:370–382
20. Nei M (1987) *Molecular evolutionary genetics*. Columbia University Press, New York, NY, USA, p 257
21. Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37–48
22. Foster P, Torroni A, Renfrew C, Röhl A (2001) Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol Biol Evol* 18:1864–1881
23. Tully LA, Parsons TJ, Steighner RJ, Holland MM, Marino MA, Prenger VL (2000) A sensitive denaturing gradient-Gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am J Hum Genet* 67:432–443
24. Tully G, Barritt SM, Bender K, Brignon E, Capelli C, Dimo-Simonin N, Eichmann C, Ernst CM, Lambert C, Lareu MV, Ludes B, Mevag B, Parson W, Pfeiffer H, Salas A, Schneider PM, Staalstrom E (2004) Results of a collaborative study of the EDNAP group regarding mitochondrial DNA heteroplasmy and segregation in hair shafts. *Forensic Sci Int* 140:1–11
25. Calloway CD, Reynolds RL, Herrin GL Jr, Anderson WW (2000) The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. *Am J Hum Genet* 66:1384–1397
26. Maruyama S, Minaguchi K, Saitou N (2003) Sequence polymorphisms of the mitochondrial DNA control region and phylogenetic analysis of mtDNA lineages in the Japanese population. *Int J Leg Med* 117:218–225