Title	Y-chromosomal binary haplogroups in the Japanese population and their relationship to 16 Y-STR polymorphisms
Author(s) Alternative	Nonaka, I; Minaguchi, K; Takezaki, N
Journal	Annals of human genetics, 71(4): 480-495
URL	http://hdl.handle.net/10130/491
Right	The definitive version is available at www.blackwell-synergy.com

Y-chromosomal Binary Haplogroups in Japanese Population and their Relationship to 16 Y-STR Polymorphisms

I Nonaka[1], K Minaguchi[1]* and N Takezaki[2]

[1]Department of Forensic Odontology, Tokyo Dental College, 1-2-2 Masago, Mihama-ku, Chiba City, 261-0011, Japan.

[2]Information Technology Center, Kagawa University, 1750-1 Ikenobe, Mikicho, Kitagun, Kagawa 761-0793, Japan.

Tel : [81] (043) 270 3785

Fax : [81] (043) 270 3788

E-mail: minaguci@tdc.ac.jp

Running head: Y-chromosomal binary and STR polymorphisms

Key words: Y-chromosome, binary haplogroups, Y-STR haplotypes, Japanese population

Summary

We investigated Y chromosomal binary and STR polymorphisms in 263 unrelated male individuals from the Japanese population and further examined the relationships between the two separate types of data. Using 47 biallelic markers, we distinguished 20 haplogroups; four of which, D2b1/-022457, O3/-002611*, O3/-LINE1 del, and O3/-021354*, were newly defined in this study. Most haplogroups in the Japanese population are found in one of the three major clades, C, D, or O. Among these, two major lineages, D2b and O2b, account for 66% of Japanese Y chromosomes. Haplotype diversity of binary markers was calculated at 86.3%. The addition of 16 Y-STR markers increased the number of haplotypes to 225, yielding a haplotype diversity of 99.40%. A comparison of binary haplogroups and Y-STR type revealed a close association between certain binary haplogroups and Y-STR allelic or conformational differences, such as those at the DXYS156Y, DYS390m, DYS392, DYS437, DYS438, and DYS388 loci. Based on our data on the relationships between binary and STR polymorphisms, we estimated the binary haplogroups of individuals from STR haplotypes and frequencies of binary haplogroups in other Japanese, Korean, and Taiwanese Han populations. The present data will enable researchers to connect databases from binary haplogrouping in anthropological studies and Y-STR typing in forensic studies in East Asian populations, especially those in and around Japan.

Introduction

Y chromosome polymorphisms are of particular interest in human evolutionary studies, forensic genetics, and medical genetics (Underhill *et al.* 2000; YCC, 2002; Mitchell & Hammer, 1996; Gill *et al*. 2001; Jobling & Tyler-Smith, 2000, 2003). Y chromosomal biallelic markers have been studied by several research groups using different markers and nomenclature systems. To unify

this data, in 2002, the Y Chromosome Consortium (YCC) constructed a highly-resolved tree of binary haplogroups by genotyping the most published PCR-based markers on a common set of samples, which made it possible to compare the data of different studies on Y chromosomal binary polymorphisms (YCC, 2002; Jobling & Tyler-Smith, 2003). Recent studies on the non-recombining portion of Y chromosome (NRY) polymorphisms in the human population have shown a higher degree of population specificity than has been found using mtDNA or autosomal markers (Seielstad *et al.* 1998), making NRY markers more informative for comparing population relationships.

Y-STR polymorphisms are widely used in the field of forensic medicine because of their high level of diversity, and an enormous amount of Y-STR haplotype data has been accumulated (Y-STR Haplotype Reference Database –YHRD; www.ystr.org). Growth in such population data reflects its usefulness and importance for personal identification in this field.

Bosch *et al*. (1999) first described the relationship between Y-STR variability and Y binary haplogroups in populations from northwestern Africa and the Iberian Peninsula. Clarification of the relationship between Y chromosome biallelic markers and STR polymorphisms would enable mutual use of such data in both the evolutionary and forensic fields of study, and would greatly increase the utilitarian value of Y chromosome polymorphisms.

This study had two main purposes: to elucidate and enrich our knowledge of binary haplogroups in the Japanese population and to investigate 16-STR microsatellites to determine any possible relationship that might exist between the two, thus establishing a link between the two separate types of data in terms of use.

Materials and Methods

Samples

Genomic DNA was extracted from blood samples from 263 healthy unrelated Japanese male

individuals. Informed consent was obtained from the blood donors. Fig. 1 shows the areas of Japan and number of individuals studied in each prefecture, which also corresponds to their birthplace. Leukocyte preparations from the blood were digested with proteinase K (Sigma) at 55℃ overnight, followed by treatment with RNAse at 55℃ for 2 hrs. DNA was extracted with phenol/chloroform, precipitated with ethanol, and resuspended in TE buffer (10mM Tris-HCl, 1mM EDTA at pH 7.6).

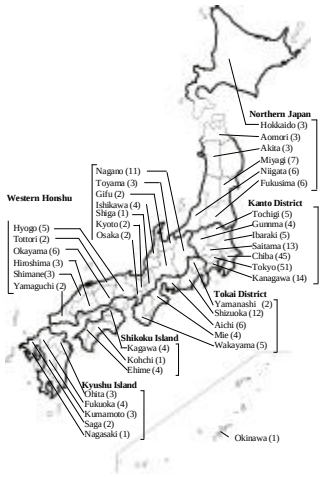

Figure 1

Detection of Biallelic Markers

PCR primer sequences, amplification product sizes, mutation positions, mutation patterns, annealing temperatures, and detection methods for 48 biallelic markers, including five kinds of marker quoted from the JSNP database (http://snp.ims.u-tokyo.ac.jp), are shown in Appendix 1 (Hammer & Horai 1995; Underhill *et al.* 1997; Zerjal *et al.* 1997; Shinka *et al.* 1999; Underhill *et al.* 2001; Santos *et al.* 2000; Hammer *et al.* 2001; Karafet *et al.* 2002; Haga *et al.* 2002; Hirakawa *et al.* 2002; Zegura *et al.* 2004). We first examined 26 markers from the JSNP database. Among these, 17 were polymorphic. However, we found no differences by SSCP gel electrophoresis at the other

9 loci. Among the 17 polymorphic loci, 3 were detected at multiple loci on the Y chromosome, 4 were not Y-specific when small amplification products were used for PCR amplification, 5 were associated with known NRY markers, and the other 5 are presented in this paper. Although it has been reported that these 5 markers were polymorphic in 24 Japanese individuals (Hirakawa *et al.* 2002), the relationship between these markers and NRY haplogroups remains to be elucidated. IMS-JST021355 and IMS-JST021354 are located at intron 3 of the zinc finger protein (ZFY) gene; IMS-JST022457 and IMS-JST022454 are located at intron 3 of the RBMY1A1 gene; and IMS-JST002611 is found in BAC clone RP11-105L10 from Y on Yp. All primer sequences in this study other than YAP, SRY465, P36, P39, and LINE1, were newly constructed referring to the sequence described by YCC (2002) and the GenBank database.

PCR amplification was performed in a 30 μl mixture containing 10 ng genomic DNA, 10 mM Tris-HCl at pH 8.3, 50 mM KCl, 2.5 mM $MgCl_2$, 0.02% gelatin, 200 μM dNTP, 400 nM of each primer, and 1.5U AmpliTaq Gold (Applied Biosystems). A 2-step PCR amplification process was used: 95℃ for 10 min, followed by 35 cycles of denaturation at 95℃ for 50 sec and annealing and extension at an appropriate temperature for 105 sec as shown in Appendix 1. After the 35th cycle, a final extension step was performed at annealing and extension temperature for 10 min. However, in the case of LINE1, we used 2 mM $MgCl_2$ and 1.25U AmpliTaq Gold in the amplification mixture and applied three-step PCR: 95℃-62℃-72℃ for 30 sec each. Most of the mutations found in the Japanese population were detected by SSCP gel electrophoresis. SSCP analysis was performed in a 17% polyacrylamide gel as described by Fujita & Kiyama (1995), modified so that some of the gel contained 5% glycerin; the gel and reservoir buffer were 0.475x TBE and 1x TBE, respectively, and a 16 x 36 cm gel of 0.4 cm thickness was used. Electrophoresis was performed at 55 V/cm constant voltage at 17.5℃. 47z was determined by PCR-RFLP, using restriction enzyme StuI (Shinka *et al.* 1999). Apart from M143, insertion or deletion polymorphisms, such as the YAP, M15, M175, LINE1, and M117, were separated by 8% or 12% polyacrylamide gels. All the

products were visualized by silver staining. Binary polymorphisms not observed in the Japanese population were compared by SSCP gel electrophoresis, as well as sequencing of the PCR products.

Detection of STRs

Three multiplexes were designed to detect 16 Y-STR polymorphisms (Appendix 2). Multiplex 1 included DYS19, DYS3891, DYS3892, DYS393, DYS435, and DYS437. Multiplex 2 included DYS390, DYS391, DYS392, DXYS156Y, and DYS388. Multiplex 3 included DYS385, DYS434, DYS436, DYS438, and DYS439. All the multiplexes were amplified under identical conditions. The reaction mixture contained 200 μM dNTPs, 1.5mM $MgCl_2$, 1.5U AmpliTaq Gold and 1-3 ng of template DNA in 20 μl final volume. The PCR conditions were as follows: 95℃ for 10 min, followed by 30 cycles of denaturation at 95℃ for 30 sec, annealing at 56℃ for 30 sec, extension at 72℃ for 50 sec, and final extension at 72℃ for 10 min. Primer sequences, dye-labeled primer concentrations, repeat numbers, and allele sizes are shown in Appendix 2 (Kayser *et al*. 1997; Ayub *et al*. 2000). For genetic typing, the ABI310 automatic sequencer (Applied Biosystems) and GeneScan 2.1 analysis software were used. The PCR products labeled with FAM and HEX were mixed and injected into a single run, using the GS400HDROX size standard. The PCR products labeled with TET were injected with the GS500TAMRA size standard.

The present samples included those previously analyzed for 10 Y-STR polymorphisms (Nonaka & Minaguchi, 2001). The primer sequences used in this study to amplify small PCR products for the DYS3891, DYS3892, DYS390, DYS391, DYS392, and DYS393 loci and large PCR products for the DYS385 locus, annealing temperatures, allele numbers, and allele sizes are also shown in Appendix 2. Two-step PCR was used to amplify these loci. PCR products were separated by 6% denaturing gels and visualized by silver staining.

The structural variations at the DYS390, DYS437, and DYS438 loci were analyzed by SSCP gel electrophoresis (Fujita & Kiyama, 1995) of the PCR products amplified using the same, but

non-labeled, primers as those used for multiplex PCR (Appendix 2). The 10% gel containing 5% glycerin was used for the DYS390 locus, and the 17% gel containing 5% glycerin was used for the DYS437 and DYS438 loci. The conditions for SSCP electrophoresis were the same as those described above. The PCR products were visualized by silver staining.

Sequence Analysis

PCR for sequencing was performed using the BigDyeTM Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems). Excessive dye was removed using DyeEx TM Spin Kit (Qiagen). Sequence analysis was performed on an ABI 310 DNA Sequencer.

Data Analysis

To determine binary haplogroups, we referred to the YCC NRY Tree 2003 (Jobling & Tyler-Smith, 2003). In this paper, in order to simplify descriptions, we have used + and - for the presence and absence of mutations. In addition, in designating new lineages, we have not changed the original binary haplogroup name used in the YCC NRY Tree 2003, as we did not want to confuse the present status of the tree. Therefore, nomenclature for the new binary haplogroups is tentative.

Simple haplotype diversity, *i.e.*, the probability of sampling two individuals with different haplotypes, was calculated as $1-\Sigma\ \chi_i^2$, where χ_i is the frequency of the i-th haplotype and n is the total number of haplotypes.

To measure Y-chromosome STR diversity within haplogroups, we employed the method described by Bosch *et al*. (1999).

Results and Discussion

Evaluation of Methods for Detecting Biallelic Markers

The PCR product sizes of biallelic markers P43, YAP, and P36, exceed 500 bp, which is not suitable for detecting binary markers from degraded DNA. However, in this study, we constructed primers to make the product sizes of the markers needed to determine the bottommost descendant of the YCC tree small enough to obtain amplification products from degraded DNAs in forensic and archeological studies. We first searched for sequences similar to the sequence surrounding the mutation position and constructed primers specific to the target sequence. The final PCR product sizes of these markers ranged from 90 bp to 172 bp. When PCR amplification was performed using female DNA as a template, very faint bands were detected in the amplification products at the M217, M214, P31, 47z, M88, and M7 loci in different positions to those of the target bands. Some of these faint bands were also amplified from male DNAs, although they did not affect the typing of these polymorphisms, either by SSCP or sequencing analysis. Because of the presence of sequences similar to those of the primers for M174 and M125, a comparatively clear band may also be amplified from female DNA when using an annealing temperature lower than those shown in Appendix 1. However, these bands were clearly distinguishable from the target bands in SSCP gel electrophoresis. The nucleotide sequences of all of the PCR products for the samples, with and without mutations, were confirmed by sequencing. These results showed that typing using the present primer pairs was reproducible and reliable.

Binary Haplogroups in the Japanese Population

Our first goal was to elucidate Y chromosomal binary haplogroups in the Japanese population referring to the YCC NRY Tree 2003 (Jobling & Tyler-Smith, 2003). We first examined 165 individuals to prepare a template for the bottommost haplogroup in the Japanese population, including the markers quoted from the JSNP database. The results of our analysis are shown in Appendix 3, and the process of defining the binary haplogroups is summarized in Table 1. We

examined biallelic markers YAP, IMS-JST022457 (022457), M9, SRY465, IMS-JST022454 (022454), and 47z in 165 individuals. However, because the 022454 mutation was completely associated with the SRY465 mutation (Appendix 3), all the individuals were divided into six groups using five kinds of biallelic marker: (1) YAP-/M9-, (2) YAP+/022457-, (3) YAP+/022457+, (4) M9+/SRY465+/47z-, (5) M9+/SRY465+/47z+, and (6) M9+/SRY465-. To determine the bottommost haplogroups corresponding to those of the YCC NRY tree and the relationship between these haplogroups and the new markers, we used hierarchical genotyping strategy (Underhill *et al*. 2001; Hammer *et al*. 2001). We essentially genotyped samples targeting at least those markers necessary to determine the bottom branch of the NRY tree.

[Table 1]

Among the six samples classified by YAP-/M9- (group 1), two were M105+, while the other four were M217+/M93-/P39-/M48-/M77-/M86-, showing that these belonged to the C1 and C3* haplogroups, respectively. The 34 YAP+/022457+ (group 3) individuals were M116a+/M125+, suggesting that this lineage belonged to D2b1. Two of the 32 YAP+/022457- samples (group 2) were also M116a+/M125+, corresponding to D2b1. Therefore, D2b1 was split into two haplogroups: M125+/022457+, designated D2b1/-022457, and M125+/022457-, designated D2b1/-M125*. Among the remaining 30 samples classified by YAP+/022457-, one sample was M15+, 12 were M55+/M116a-, and 17 were M116a+/M125-/M151-, which corresponded to the D1, D2a, and D2b* haplogroups, respectively. We also examined a new biallelic marker, IMS-JST021355. This mutation was completely associated with the YAP mutation (DE lineage) in the 118 samples investigated. It was also associated with the M174 mutation (D lineage), but not with the SRY4064 mutation (E lineage)(Appendix 3). Therefore we assigned this mutation to the same position as the M174 mutation (Fig. 2).

In the three groups classified by M9+, 12 M9+/SRY465+/47z- samples (group 4) could be assigned to O2b*, and 43 M9+/SRY465+/47z+ samples (group 5) could be assigned to O2b1. In

the remaining 38 samples of M9+/SRY465- (group 6), M175 was determined in all samples. Among these, 35 samples were M175+ and three samples were M175-. One of the three M175- samples (ST263 in Appendix 3) was M214-/M175-, suggesting that it did not belong in haplogroup O. This sample was M74+/P36+/M120+, showing that it belonged in haplogoup Q1. Because the other two (OS118 and AT119 in Appendix 3) were M9+/M214＋/M175-, they were considered to belong to macrohaplogroup N. However, these two samples did not share the M128, P43, or Tat mutations, and, as sequence information on LLY22g was not available, we classified them into the N/0 (xM175, xM128, xP53, xTat) haplogroup. The 35 M175+ samples were classified into 3 lineages, O1, O2, and O3, by the M119, P31, and M122 mutations, respectively. The 5 M119+ samples were M110-/M50-, showing that these belonged to O1*. The 2 P31+ samples were M95+/M88-, showing that these belonged to O2a* (Fig. 2). When 28 samples of M122+ (O3) were divided into two groups by the 021354 mutation, the M122+/021354- samples were further classified into 3 groups by the mutations 002611 and LINE1: LINE1-/002611-, LINE1+/002611+, and LINE1-/002611+. Because the LINE-/002611- sample did not share the M121, M164, M7, M159, or M134 mutations (Fig. 2), it was classified into O3*. Although the genotypes of the other groups, LINE1+/002611+ and LINE1-/002611+, suggested that the 002611 mutation occurred earlier than the LINE1 mutation, after increasing sample size, we found a controversial sample with LINE1+/002611- (TK221 in Appendix 3). This suggests that recurrent or back mutation occurred in 002611 or LINE1. Because the 002611 mutation was a point mutation, a recurrent or back mutation would be very unlikely. In contrast, it was reported that the LINE1 element was found integrated into the repetitive alphoid satellite array (Santos *et al*. 2000), which is known to undergo frequent expansion/contraction. Therefore, it is more likely that deletion occurred at the LINE1 locus. We found two more individuals with LINE1+/002611- in the Japanese population, and seven more samples in the Malay population later on. Therefore, we believe that the LINE1 insertion occurred as the earliest event (producing LINE1+/002611-=O3c*), followed by the

002611 mutation (producing LINE1+/002611+=O3/-002611*), with LINE1 finally being deleted (producing LINE1-/002611+=03/LINE1 del). Because all of these samples were M159-, the position where M159 split could not be determined in our samples. The M122+/021354+ samples contained the M134+ samples, but 021354+/M134- samples were also found, which showed that the 021354 mutation occurred before the M134 mutation. Therefore, the haplogroup of the 021354+/M134- samples was designated O3/-021354*.

We first established 20 binary haplogroups in 166 Japanese individuals. Then, we further determined an additional 97 individuals, and each of these individuals belonged to one of the 20 lineages described above (Appendix 3). The distribution and frequency of each haplogroup are summarized in Fig. 2. Among 263 combined samples, the frequencies of the D2, O2b, and O3 lineages were 38.8%, 33.5%, and 16.8%, respectively, which constituted approximately 90% of the Japanese population. Haplogroup diversity for the binary polymorphisms was calculated to be 86.3%.

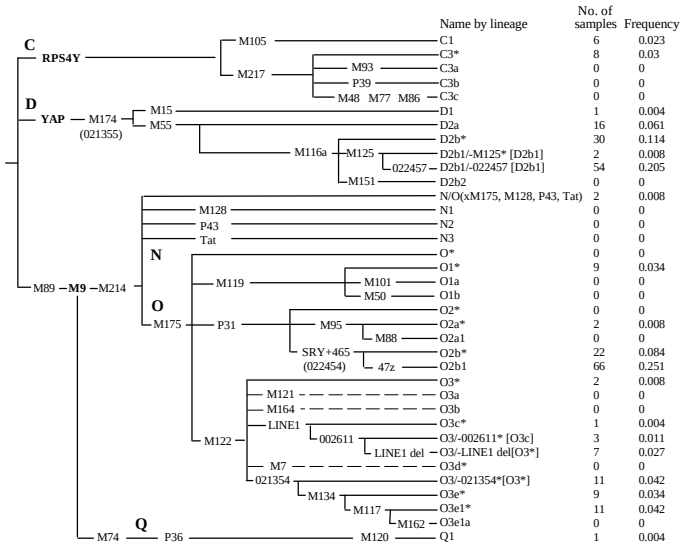

Figure 2

Determination of Y-STR Types and Relationship between Binary Haplogroups

Y-STR Polymorphisms

Next, we set out to determine Y-STR types. Although some of the present samples had already been typed for 10 Y-STR loci (DYS19, DYS3891, DYS3892, DYS390, DYS391, DYS392, DYS393, DYS385, DYS388, and DXYS156Y) (Nonaka & Minaguchi, 2001), six additional loci (DYS434, DYS435, DYS436, DYS437, DYS438, and DYS439) were further typed in the samples in this study, and the total number of samples was increased. We further examined conformational polymorphisms at 3 loci, DYS390, DYS437, and DYS438 (Appendix 4). These results were compared with binary haplogroups.

The allelic frequencies and gene diversities of the 16 Y-STR polymorphisms determined based on size variation in 263 individuals are shown in Appendix 5. The total number of STR haplotypes was 223, 203 of which were seen in only one individual each. The most frequent STR haplotype was observed in 3.04% of the population. Two kinds of STR haplotype were found in both the O2b* and O2b haplogroups. Therefore, when both types of information were combined,

the total number of haplotypes was 225, and haplotype diversity was calculated to be 99.40% (Appendix 4).

Conformational Polymorphisms of Y-STRs and Binary Haplogroups

The consensus repeat structure of the DYS390 locus is reported to be $(CTAT)_2(CTGT)_m(CTAT)_n(CTAT)_p(CTGT)_q$ (Forster *et al*. 2000). When we sequenced a total of 100 genes at the DYS390 locus and further compared them by SSCP gel electrophoresis, differences in repeat structure were completely associated with differences in SSCP gel pattern. Therefore, the typing of the remaining 163 samples was performed by SSCP gel electrophoresis (data not shown). Six kinds of allele, 22-27, could be classified into a total of 11 types, depending on the combination of the number of repeats of “m” and “n” (Appendix 4), but no samples were found with differences in the numbers of p or q in the Japanese population. Determination of conformational polymorphisms increased the gene diversity value of the DYS390 locus from 75.8% to 80.3%. The repeat number of the DYS390m segment was completely associated with certain binary haplogroups, except for one sample in C3* (KG7) (Appendix 6).

The consensus repeat structure of the DYS437 locus is reported to be $(TCTA)_m(TCTG)_2(TCTA)_4$, and differences in the number of “m” reflect differences in alleles (Gusmao *et al*. 2001, 2002). While determining the sequence of allelic products, two types of structural variation, $(TCTA)_m(TCTG)_2(TCTA)_4$ and $(TCTA)_m(TCTG)_1(TCTA)_4$, were found in the Japanese population. These types could be discriminated by SSCP gel electrophoresis (data not shown), allowing 198 samples to be examined by this method. Both types of structural variation were observed in alleles 14 and 15, respectively (Appendix 4). In contrast to the common repeat structure of $(TCTA)_m(TCTG)_2(TCTA)_4$, the $(TCTA)_m(TCTG)_1(TCTA)_4$ type was found in all of the samples in the O3 lineage, except in two samples which were classified into the O3* haplogroup. Although the relationships between this variation and the markers for O3a, O3b, and O3d (M121,

M164, and M7) are not known, it is evident that $(TCTA)_m(TCTG)_1(TCTA)_4$ diverged before the LINE1 and 021354 mutations.

The consensus repeat structure of the DYS438 locus is reported to be $(TTTTC)_1(TTTTA)_{0-1}(TTTTC)_m$ (Gusmao *et al.* 2001). It is also reported that the $(TTTTC)_1(TTTTA)_1(TTTTC)_8$ structure is present in the Chinese population (Hou *et al.* 2001). We compared 137 samples at the DYS438 locus by SSCP gel electrophoresis (data not shown). The structural variation with the repeat structure $(TTTTC)_1(TTTTA)_1(TTTTC)_8$ was observed in six (NG9, NN10, CB11, HO12, IR13, IK14) of the eight samples in haplogroup C3* (Appendix 4). The Y-STR haplotypes of the other two samples (KG7, CB8) differed from those of these six samples in the informative loci used to estimate binary haplogroups from STR type, as described below. One (KG7) differed in the repeat number of the DYS390m segment, and in the allelic combination of DYS385; the other (CB8) differed in the allele of DYS392, and the allelic combination of DYS385. We have already found two kinds of SNPs: One can discriminate these two samples and the six $(TTTTC)(TTTTA)(TTTTC)_8$ samples, and the other can further discriminate these two samples (unpublished observation). The sequence surrounding these SNPs, however, has multiple counterparts in the human genome database, all of which remains to be discriminated. These results suggest that individuals with the $(TTTTC)(TTTTA)(TTTTC)_8$ repeat constitute a different clade in the C3* haplogroup.

As described above, certain structural variations at STR loci offer fairly stable markers for phylogenetic comparison.

Size Variation in Y-STR Polymorphisms and Binary Haplogroups

As part of our second goal, we investigated potential relationships between size variation in Y-STR type and binary haplogroup in the Japanese population. When Y-STR types were compared within the same binary haplogroups, most of the haplotypes shared similar combinations

in the informative loci used to estimate binary haplogroups, as described below. However, individuals with different characteristics were also found. Six individuals in D2a (TK26, OY27, KG28, IR29, CB30, HG31) had different types at the DYS438 and DYS388 loci from the other samples in D2a, and one individual in O2b* (KN152) had a unique combination of DYS19, DYS3892, DYS390n, DYS436, and DYS438. The SNP which associated with the (TTTTC)(TTTTA)$(TTTTC)_8$ samples in the C3* haplogroup was also completely associated with the difference between these six individuals and the others in the D2a haplogroup. In addition, we found a SNP on Y chromosome in the other O2b* sample, although we cannot be sure whether it is an NRY polymorphism at this point (unpublished observation). Although appropriate biallelic markers that can define these types have yet to be established, these samples seem to constitute a different haplogroup from the common D2a or O2b* haplogroups. Therefore, we assigned new provisional haplogroup names for these six samples in D2a and one sample in O2b*, designating these as $D2a^n$ (new) and $O2b^{*n}$, respectively, (Appendix 4) and compared the relationships between STR type and binary haplogroup as described below.

We first compared diversity of STR haplotypes within and among binary haplogroups. Table 2 shows several parameters of STR diversity for haplogroups in the Japanese population. We compared haplogroups with a sample size of more than 3. O2b1 was the least diverse, and C3* and O3/-021354* were the most diverse haplogroups, as shown by average gene diversity, average number of loci with different alleles, and average difference in total repeat size. It is possible that these haplogroups with high diversity value contain old lineages with heterogenous STR-haplotypes. Although the most suitable biallelic marker has yet to be found, haplogroup C3* may contain 3 haplogroups, as mentioned above. Two samples (EH241 and TK242) in the O3/-021354* haplogroup, which differed in the allele sizes of DYS393, DYS434 and DYS385 combination, may constitute a different haplogroup, as we have already found a SNP in these samples (unpublished observation), although it remains to be established whether it is an NRY marker. The average

difference in total repeat size was also compared among different haplogroups (Table 3). The largest total repeat size difference within haplogroups was 11.57, found in the C3* haplogroup. Smaller corresponding values than this among haplogroups were obtained in each pair among D2a, D2a^{n}, D2b*, and D2b/-022457, between O2b* and O2b1, and in each pair among O3/-002611, O3/-LINE1 del, and O3/-021354. The value between O2b* and O2b1 (6.25) was the smallest among haplogroups, and was smaller than several values obtained within haplogroups. This suggests that divergence between O2b* and O2b1 was more recent than the divergence of STRs within these haplogroups. Average values in difference in total repeat size among haplogroups are expected to correlate with the divergent pattern of the binary NRY tree (Bosch *et al.* 1999). Although larger values were obtained between distant haplogroups than close ones, exceptions were also found. The value between O1* and O3e* was the smallest among those values between O1* and the other haplogroups. This was probably because of the similarity of the allele combination at the highly diverse DYS385 locus.

[Table 2, Table 3]

Estimation of Binary Haplogroups from Y-STR Haplotypes

Next, we tried to estimate binary haplogroups from STR haplotypes based on size difference, and further calculated frequency of binary haplogroups in other Japanese and nearby populations using STR haplotype data. In order to estimate binary haplogroups from STR haplotypes, we compiled data on distribution of allele frequency in individual haplogroups at each STR locus. The relationships between the STR alleles at the DXYS156Y, DYS390m, DYS392, DYS438, DYS388, DYS393, DYS434 DYS435 DYS437, DYS3891, DYS390, DYS19, DYS3892, DYS391, and DYS439 loci, and the binary haplogroups in this study are summarized in Appendix 6. Allelic variation at each locus showed striking differences among haplogroups. Some haplogroups displayed several alleles, while, in others, the number of alleles was highly restricted. Such alleles

which differ from the common alleles at a particular STR locus are highly informative for estimating corresponding haplogroups from STR databases. The two STR loci, DXYS156Y and DYS392, were very informative in discriminating the deep node of the Japanese haplogroups (C/D and O). Other informative loci for screening and estimating particular haplogroups can be found in Appendix 6. Information about the DYS385 locus is also useful in estimating binary haplogroups, as the allelic combination of this locus shows distinct tendencies within each haplogroup. However, this locus has a high gene diversity value (Kayser *et al.* 2000), and the combination of the alleles is not restricted to a single type of combination in each haplogroup. Therefore, they are not shown in Appendix 6. These results were used as a reference for estimating binary haplogroups from STR haplotypes.

We used STR haplotype data in other Japanese populations from the Northern, Central and Southern areas of Japan, the Korean population, and the Han population in Taiwan. We selected databases composed of as many STR loci as possible, as well as those including the DXYS156Y or DYS392 loci. The 200 samples from Asahikawa, a city situated in the central part of Hokkaido, the most northern island of Japan (Fig. 1) (Sasaki & Dehiya, 2000), were composed of 8 loci (DYS19-388-3891-3892-390-391-392-393); 207 from Nagoya, a city situated in Aichi Prefecture, and 87 from Okinawa, a southern island of Japan (Fig. 1) (Uchihi *et al*. 2003), were composed of 12 loci (DYS19-385-3892-390-391-392-393-435-436-437-438-439); 316 from Korea were composed of 10 loci (DXYS156Y-DYS19-385-388-3891-3892-390-391-392-393) (Shin *et al*. 2001); and 183 from Taiwanese Han were composed of 9 loci (DYS19-385-388-3891-3892-390-391-392-393) (Tsai *et al*. 2001). In order to estimate binary haplogroups from STR haplotypes, each database was first classified by DXYS156Y or DYS392, which are useful for discriminating macrohaplogroups. Then, the most similar STR haplotype to each sample was searched from the control database established in this study in consideration of other characteristic relationships between STR and binary haplogroups, as shown in Appendix 6. When the haplotype in question

differed in one repeat from the corresponding haplotype, we considered the sample to be a one-step neighbor to the corresponding haplotype. By doing such, we determined the distance in steps from the most similar haplotypes to each sample in the database. However, some samples were equally distant from more than two kinds of haplogroup.

The finding that binary haplogroups retain characteristic STR haplotypes suggests that we can estimate corresponding haplogroups from STR haplotypes. In this study, after establishment of binary haplogroups in 165 samples from the Japanese population, we further examined 98 samples. All but one sample (O3c*) corresponded to one of the established haplogroups. The results indicate that in estimating binary haplogroups from STR haplotypes in the Japanese population, we can assign all haplotypes to one of the present haplogroups. If the database does not contain similar haplotypes to a given sample in the corresponding haplogroup, the number of steps at which it differs may greatly increase. For example, if the haplotype of CB8 in the C3* haplogroup (Appendix 4) was not included in our database, and if this haplotype were found as a new one, it would differ from the other haplotypes in C3* at a minimum of 11 steps. Therefore, under the current method of estimating haplogroups, increasing the number of haplotypes in the control database might decrease the distance of differences from the control haplotype and chances of wrongly assigning a haplotype to a haplogroup. However, some haplotypes may be judged to be equally distant from two closely related haplogroups, no matter how much the control database increases.

In order to assign a STR haplotype to the most suitable haplogroup, it is necessary to determine cut-off values. It is not appropriate, in a sense, to uniformly apply the same cut-off value to all populations in this study, as the number and kinds of loci used to construct haplotypes differed in the population database. However, because differences in number of steps from the control database may vary due to scarcity of data or diversity of loci in a haplogroup, it is difficult to establish an ideal cut-off size for each population to obtain suitable results. When assignment of a

haplogroup was limited to within 4 steps, the frequencies of the samples to which haplogroup could not be assigned were 18.9%, 12.1% 12.6%, 25.9%, and 25.7% for the populations in Asahikawa, Nagoya, Okinawa, Korea and Taiwan, respectively. The numbers of samples that could not be assigned decreased by increasing cut- off size. By fixing the cut-off size to 7, only 0.5%, 1.9%, 1.2%, 3.8%, and 2.2% for each population could not be assigned. Because almost all the Japanese samples could be assigned to one of the present haplogroups, we provisionally used the same standard of a 7-step cut-off in this study. When the most similar haplotype differed within more than 8 steps, we concluded that no corresponding haplogroup was found.

In order to establish the accuracy of this method of prediction, we further determined STR haplotypes in a new dataset of 79 individuals from the Japanese population, predicted binary haplogroups, and compared the results determined by biallelic markers. In 71 samples (89.9%), the binary haplogroup determined coincided with a single predicted haplogroup. In 4 samples (5.1%), haplotype was equally distant from two kinds of closely related haplogroup. However, when the characteristics of association between haplogroup and allele size of STR were considered, all of these samples shared the characteristics of the inherent haplogroups shown in Appendix 6. In the remaining 4 samples (5.1%), the predicted haplogroups were closer to different haplogroups than the inherent ones. These samples shared the characteristic allele types of the inherent haplogroups, but bore slightly different haplotypes to those found in our control database. The haplotype of the sample belonging to O3c*, in particular, differed considerably from our control haplotype (there was only one sample in our database). The main reason these samples could not be assigned properly was the lack of control haplotypes in our database. Finally, using the present database, we could successfully predict binary haplogroups from STR haplotypes in 95% of the samples. Therefore, prediction of binary haplogroup from STR haplotype was fairly reliable.

[Table 4]

In calculating frequencies for binary haplogroups, some samples differed within the same

number of steps from 2 or 3 different haplogroups. Among 195 haplotypes from Asahikawa, 19.5% of the samples were equally distant from two (14.9%) or three (4.6%) haplogroups. In other populations from Nagoya, Okinawa, Korea, and Taiwan, 7.7%, 5.1%, 4.1%, and 8.6% of 183, 78, 295, and 162 haplotypes, respectively, were equally distant from two of the haplogroups. In these cases, a frequency of 1/2 or 1/3 was assigned to each haplogroup. The results of the final estimated frequencies for these binary haplogroups are shown in Table 4. In order to evaluate these estimated frequencies obtained from the Korean and Taiwanese Han populations, we compared them with the binary haplogroup frequencies reported by Karafet *et al*. (2001) and Jin *et al*. (2003). The estimated frequencies of the LINE1+ haplogroups (O3c* + O3/-002611) obtained in this study were 0.9% for Korean and 10.4% for Taiwanese. These were considerably lower than the frequencies reported by Karafet *et al*. (2001) and Jin *et al*. (2003), at 9.5% and 12.5%, for Korean, respectively, and 26.8% for Taiwanese Han (Karafet *et al*. 2001). In contrast, the estimated frequencies for O3/-LINE1 del were very high in both populations at 6.2% for Korean and 11.7% for Taiwanese. If the frequencies for O3/-LINE1 del were combined with the LINE1+ haplogroup in both populations, the combined frequencies at 7.1% for Korean and 22.1% for Taiwanese, were very similar to those of the LINE1+ haplogroups in those reports. Because the STR haplotypes of O3c, O3/-002611, and O3/-LINE1 del were similar, these results are not controversial, assuming that the LINE1 deletion has not occurred in the Korean and Taiwanese Han populations. Because the frequency of LINE1+ in the Japanese population was very low (1.5%), it is possible that back mutation of the LINE1 deletion occurred near Japan, and that this haplogroup is fairly restricted to the Japanese population. The frequencies of O3* in Korean and Taiwanese Han reported by Karafet *et al*. (2001), at 10.8% for Korean and 8.5% for Taiwanese, were higher than those found in our study, at 0.3% for Korean and 1.5% for Taiwanese. However, when the estimated frequencies for the O3/-021354* haplogroup were assigned to those for O3* in our study, the combined frequencies for O3* in both populations, at 10.1% for Korean and 7.4% for

Taiwanese, were very close to those reported by Karafet *et al.* (2001). Because the 021354* mutation occurred earlier than the M134 mutation, this mutation may be present in other East Asian populations. In our analysis of Korean data, the frequency of O2b* was higher than that of O2b1, which is consistent with the results obtained by Karafet *et al*. (2001) and Jin *et al*. (2003). However, the estimated frequency of O2b1 against O2b* in the Korean population in this study, at 17.2% for O2b* and 12% for O2b1, was relatively higher than their frequencies, at 32.4% and 4.1% for O2b* and O2b1, respectively, by Karafet *et al*. (2001) and 13.8% and 5.6% for O2b* and O2b1, respectively, by Jin *et al*. (2003). This was probably due to differences within the Korean population, because the number of samples differing within the same steps from O2b* or O2b1 in our analysis was only 8 out of the 96 samples classified into the O2b lineage, and the proper assignment of those 8 samples to O2b* or O2b1 would not make much difference to our estimated frequencies for the O2b* and O2b1 haplogroup in the Korean population. In the analysis of the Taiwanese Han population, the estimated frequency of O1* (22.4%) was relatively higher than that reported by Karafet *et al*. (2001) (11%). This was probably due to differences within the Taiwanese population, because the haplotypes of the samples estimated to be O1* in this study shared similar characteristics with regard to combination of STR types. However, because haplogroup O2* is not found in the Japanese population, and because this haplogroup has been found in 7.3% Taiwanese Han population (Karafet *et al*. (2001)), the possibility remains that the individuals in haplogroup O2* share a very similar combination of STR haplotypes to those of O1*, and they were, therefore, included in haplogroup O1*.

As described above, estimation of binary haplogroups from STR haplotypes in the Japanese population was highly reproducible, and estimated frequencies of binary haplogroups from STR haplotypes in Korean and Taiwanese Han populations showed reasonable values.

In the present study, we elucidated the phylogeny of Y-chromosomal binary haplogroups in the

Japanese population, including four new lineages, established a database connecting two types of data, binary polymorphisms and STR haplotypes, and estimated banary haplogroups from STR haplotype data. Using the present database, we can estimate and determine the very bottom binary haplogroup, without analyzing many biallelic markers from samples whose STR haplotype has been already determined. In addition, we can introduce STR haplotype data for comparison of binary haplogroup using estimated frequencies. The present data will enable researchers to connect databases from binary haplogrouping in anthropological studies and Y-STR typing in forensic studies in East Asian populations, especially those in and around Japan.

Acknowledgement

We thank Associate Professor Jeremy Williams, Laboratory of International Dental Information, Tokyo Dental College, for editing manuscript. This study was supported by the Grant-in Aid for Scientific Research (17390567 and 16659171) from the Ministry of Education, Science, Sports and Culture of Japan.

References

Ayub, Q., Mohyuddin, A., Qamar, R., Mazhar, K., Zerjal, T., Mehdi, S.Q. & Tyler-Smith, C. (2000) Identification and characterization of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acids Res* **28**, e8.

Bosch, E., Calafell, F., Santos, F.R., Perez-Lezaun, A., Comas, D., Benchemsi, N., Tyler-Smith, C., Bertranpetit, J. (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet* **65**, 1623-1638.

Forster, P., Rohl, A., Lunnemann, P., Brinkmann, C., Zerjal, T., Tyler-Smith, C. & Brinkmann, B. (2000) A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Gene* **67**, 182-196.

Fujita, T. & Kiyama, M. (1995) Identification of DNA polymorphism by asymmetric-PCR SSCP. *Biotechniques* **19**, 532-534.

Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M.A., de Knijff, P., Kayser, M., Krawczak, M., Mayr, W.R., Morling, N., Olaisen, B., Pascali, V., Prinz, M., Roewer, L., Schneider, P.M., Sajantila, A. & Tyler-Smith, C. (2001) DNA commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *Int J Legal Med* **114**, 305-309.

Gusmao, L., Alves, C. & Amorim, A. (2001) Molecular characteristics of four human Y-specific microsatellites (DYS434, DYD437, DYS438, DYS439) for population and forensic studies. *Ann Hum Genet* **65**, 285-291.

Gusmao, L., Alves, C., Costa, S., Amorim, A., Brion M., Gonzalez-Neira, A., Sanchez-Diz, P. & Carracedo, A. (2002) Point mutations in the flanking regions of the Y-chromosome specific STRs DYS391, DYS437 and DYS438. *Int J Legal Med* **116**, 322-326

Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y. & Tanaka, T. (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. *J Hum Genet* **47**, 605-610.

Hammer, M. F. & Horai, S. (1995) Y chromosomal DNA and the peopling of Japan. *Am J Hum Genet* **56**, 951-962.

Hammer, M.F., Karafet, T., Rasanayagam, A., Wood, E.T., Altheide, T.K., Jenkins, T., Griffiths, R.C., Templeton, A.R., Zegura, S.L. (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol* **15**, 427-441.

Hammer, M.F., Karafet, T.M., Redd, A.J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H. & Zegura, S.L. (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* **18**, 1189-1203.

Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., & Nakamura, Y. (2002) JSNP: a

database of common gene variations in the Japanese population. *Nucleic Acids Res* **30**, 158-162

Hou, Y.P., Zhang, J., Li, Y.B., Wu, J., Zhang, S.Z. & Prinz, M. (2001) Allele sequences of six new Y-STR loci and haplotypes in the Chinese Han population. *Forensic Sci Int* **15**, 147-152.

Jin, H.J., Kwak, K.D., Hammer, M.F., Nakahori, Y., Shinka, T., Lee, J.W., Jin, F., Jia, X., Tyler-Smith, C., Kim, W. (2003) Y-chromosomal DNA haplogroups and their implications for the dual origins of the Koreans. *Hum Genet* **114**, 27-35.

Jobling, M.A. & Tyler-Smith, C. (2000) New uses for new haplotypes the human Y chromosome, disease and selection. *Trends Genet* **16**, 356-362.

Jobling, M.A. & Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* **4**, 598-612.

Karafet, T.M., Osipova, L.P., Gubina, M.A., Posukh, O.L., Zegura, S.L. & Hammer, M.F. (2002) High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol* **74**, 761-789.

Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R.S., Redd, A.J., Zegura, S.L. & Hammer, M.F. (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet* **69**, 615-628.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez- Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J., Weichhold, G., de Knijff, P. & Roewer, L. (1997) Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med* **110**, 125-133, 141-149.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff, P., Stoneking, M. & Sajantila, A. (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as

revealed by direct observation in father/son pairs. *Am J Hum Genet* **66**, 1580-1588.

Mitchell, R.J. & Hammer, M.F. (1996) Human evolution and the Y chromosome. *Curr Opin Genet Dev* **6**, 737-742.

Nonaka, I. & Minaguchi, K. (2001) Allele frequencies and haplotype o ten Y-specific STRs in the Japanese population. *J Forensic Sci* **41**, 179-182.

Santos, F.R., Pandya, A., Kayser, M., Mitchell, R.J., Liu, A., Singh, L., Destro-Bisol, G., Novelletto, A., Qamar, R., Mehdi, S.Q., Adhikari, R., de Knijff, P. & Tyler-Smith, C. (2000) A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. *Hum Mol Genet* **9**, 421-430.

Sasaki, M. & Dahiya, R. (2000) The polymorphisms of various short tandem repeats on the Y chromosome in Japanese and German populations. *Int J Legal Med* **113**, 181-188.

Seielstad, M.T., Minch, E. & Cavalli-Sforza, L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nat Genet* **20**, 278-280.

Shin, D.J., Jin, H.J., Kwak, K.D., Choi, J.W., Han, M.S., Kang, P.W., Choi, S.K. & Kim, W. (2001) Y-chromosome multiplexes and their potential for the DNA profiling of Koreans. *Int J Legal Med* **115**, 109-117.

Shinka, T., Tomita, K., Toda, T., Kotliarova, S.E., Lee, J., Kuroki, Y., Jin, D.K., Tokunaga, K., Nakamura, H. & Nakahori, Y. (1999) Genetic variations on the Y chromosome in the Japanese population and implications for modern human Y chromosome lineage. *J Hum Genet* **44**, 240-245.

The Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Research* **12**, 339-348.

Tsai, L.C., Yuen, T.Y., Hsieh, H.M., Lin, M., Tzeng, C.H., Huang, N.E., Linacre, A. & Lee, J.C. (2001) Haplotype frequencies of nine Y-chromosome STR loci in the Taiwanese Han population. *Int J Legal Med* **116**, 179-183.

Uchihi, R., Yamamoto, T., Usuda, K., Yoshimoto, T., Tanaka, M., Tokunaga, S., Kurihara, R., Tokunaga, K. & Katsumata, Y. (2003) Haplotype analysis with 14 Y-STR loci using 2 multiplex

amplification and typing systems in 2 regional populations in Japan. *Int J Legal Med* **117**, 34-38.

Underhill, P.A., Jin, L., Lin, A.A., Mehdi, S.Q., Jenkins, T., Vollrath, D., Davis, R.W., Cavalli-Sforza, L.L. & Oefner, P.J. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* **7**, 996-1005.

Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Mirazon Lahr, M., Foley, R.A., Oefner, P.J. & Cavalli-Sforza, L.L. (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* **65**, 43-62.

Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J.R., Mehdi, S.Q., Seielstad, M.T., Wells, R.S., Piazza, A., Davis, R.W., Feldman, M.W., Cavalli-Sforza, L.L. & Oefner, P.J. (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**, 358-361.

Zegura, S.L., Karafet, T.M., Zhivotovsky, L.A. & Hammer, M.F. (2004) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol Biol Evol* **21**, 164-175.

Zerjal, T., Dashnyam, B., Pandya, A., Kayser, M., Roewer, L., Santos, F.R., Schiefenhovel, W., Fretwell, N., Jobling, M.A., Harihara, S., Shimizu, K., Semjidmaa, D., Sajantila, A., Salo, P., Crawford, M.H., Ginter, E.K., Evgrafov, O.V., Tyler-Smith, C. (1997) Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. *Am J Hum Genet* **60**, 1174-1183.

Figure Legends

Figure 1 Geographic location (prefecture) and number (in parentheses) of Japanese samples presented in this study. Prefectures were divided into six regional groups according to geography. Abbreviations for prefectures in Appendix 4 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; respectively.

Figure 2 Evolutionary tree based on YCC NRY Tree 2003 showing relationships and frequencies among 20 Y chromosomal binary haplogroups observed in Japanese population. In names by lineage, original states corresponding to YCC tree are shown in parentheses for new lineages.

Table 1 Summary of analysis of biallelic markers indispensable in determining bottommost haplogroups in 165 samples.

Name by lineage	Original state	No. of samples	M105 C1	M217 C3	M93 C3a	P39 C3b	M48 C3c	M77 C3c	M86 C3c	YAP DE	M15 D1	M55 D2	M116a D2b	M125 D2b1	M151 D2b2	022457 D2b1	M9 K→	M214 NO	M175 O	M119 O1	M101 O1a
C1		2	T(+)	A(-)						(-)						C(-)	C(-)				
C3*		4		C(+)	C(-)	G(-)	A(-)	C(-)	T(-)	(-)						C(-)	C(-)				
D1		1								(+)	(+)					C(-)	C(-)				
D2a		12								(+)		C(+)	A(-)			C(-)	C(-)				
D2b*		17								(+)			T(+)	T(-)	G(-)	C(-)	C(-)				
D2b1/-M125*	D2b1*	2								(+)			T(+)	C(+)		C(-)	C(-)				
D2b1/-022457	D2b1*	34								(+)			T(+)	C(+)		G(+)	C(-)				
N/O(xM175, M128, P43, Tat)	N*	2								(-)						C(-)	G(+)	C(+)	(-)		
O1*		5								(-)						C(-)	G(+)		(+)	C(+)	C(-)
O2a*		2								(-)						C(-)	G(+)		(+)		
O2b*		12								(-)						C(-)	G(+)				
O2b1		43								(-)						C(-)	G(+)				
O3*[b]		1								(-)						C(-)	G(+)		(+)		
O3c*[a]		(1)[a]																			
O3/-002611*	O3c	2								(-)						C(-)	G(+)		(+)		
O3/-LINE1 del	O3*	5								(-)						C(-)	G(+)		(+)		
O3/-021354*	O3*	6								(-)						C(-)	G(+)		(+)		
O3e*		6								(-)						C(-)	G(+)		(+)		
O3e1*		8								(-)						C(-)	G(+)		(+)		
Q1		1								(-)						C(-)	G(+)	T(-)	(-)		
total		165																			

Name by lineage			M50	P31	M95	M88	SRY+465	22454	47z	M122	M121	M164	LINE1	002611	021354	M159	M7	M134	M117	M162	M120
			O1b	O2	O2a	O2a1	O2b	O2b	O2b1	O3	O3a	O3b	O3c	O3	O3	O3c	O3d	O3e	O3e1a	O3e1a1	Q1
C1							C(-)	T(-)	G(-)												
C3*							C(-)	T(-)	G(-)												
D1							C(-)	T(-)	G(-)												
D2a							C(-)	T(-)	G(-)												
D2b*							C(-)	T(-)	G(-)												
D2b1/-M125*							C(-)	T(-)	G(-)												
D2b1/-022457							C(-)	T(-)	G(-)												
N/O(xM175, M128, P43, Tat)							C(-)	T(-)	G(-)												
O1*			T(-)				C(-)	T(-)	G(-)												
O2a*				C(+)	T(+)	A(-)	C(-)	T(-)	G(-)												
O2b*							T(+)	G(+)	G(-)												
O2b1							T(+)	G(+)	C(+)												
O3*[b]							C(-)	T(-)	G(-)	C(+)	(-)	T(-)	(-)	C(-)	T(-)	A(-)	C(-)	(-)			
O3c*[a]							C(-)	T(-)	G(-)	C(+)			(+)	C(-)	T(-)						
O3/-002611*							C(-)	T(-)	G(-)	C(+)			(+)	T(+)	T(-)						
O3/-LINE1 del							C(-)	T(-)	G(-)	C(+)			(-)	T(+)	T(-)						
O3/-021354*							C(-)	T(-)	G(-)	C(+)					C(+)			(-)			
O3e*							C(-)	T(-)	G(-)	C(+)					C(+)			delG(-	(-)	C(-)	
O3e1*							C(-)	T(-)	G(-)	C(+)					C(+)			delG(-	(+)	C(-)	
Q1							C(-)	T(-)	G(-)												C(+)
total																					

Note: Markers of YAP, 022457, M9, SRY465, 022454, and 47z were examined in all samples and samples were divided into six groups using five kinds of biallelic marker: (1) YAP-/M9- (C1, C3*), (2) YAP+/022457- (D1,D2a, D2b*, D2b1/M125*), (3) YAP+/022457+ (D2b1/-022457), (4) M9+/SRY465+/47z- (O2b*), (5) M9+/SRY465+/47z+ (O2b1), and (6) M9+/SRY465 (N/O, O1*, O2a*, O3*, O3c*, O3/-002611*, O3/-LINE1 del, O3/-021354*, O3e*, O3e1*, Q1).

[a] Sample in haplogroup O3c* was not included in original 165 samples; only types determined are shown.

Table 2 Parameters of STR diversity within haplogroups in Japanese population. Only haplogroups with more than 3 samples are shown.

Haplogroup	No.	No.H	No.P	HD	GD	ALDA	ACDA
C1	6	6	10	0.83	0.20 ± 0.19	4.13 ± 1.9(	5.13 ± 1.96
C3*	8	8	11	0.88	0.33 ± 0.28	6.43 ± 1.5(	11.57 ± 4.45
D2a	10	9	12	0.88	0.30 ± 0.24	5.67 ± 2.1(	7.84 ± 3.65
D2a(D2an)	6	6	8	0.83	0.24 ± 0.27	4.80 ± 1.4[	5.73 ± 1.84
D2b*	30	28	16	0.96	0.31 ± 0.24	5.51 ± 1.9[	7.66 ± 3.10
D2b1/-02245	54	44	12	0.97	0.21 ± 0.20	3.57 ± 1.5[	4.52 ± 2.20
O1*	9	7	8	0.81	0.18 ± 0.23	3.53 ± 1.6[	4.61 ± 2.26
O2b*	21	18	14	0.93	0.24 ± 0.22	4.32 ± 1.6[	5.11 ± 2.19
O2b1	66	47	12	0.96	0.18 ± 0.22	3.08 ± 1.3(	3.96 ± 2.15
O3/-002611*	3	3	9	0.67	0.26 ± 0.26	6.67 ± 2.6[	8.00 ± 3.56
O3/-LINE1de	7	7	11	0.86	0.28 ± 0.26	5.62 ± 2.1(	7.62 ± 2.98
O3/-021354*	11	11	15	0.91	0.39 ± 0.25	7.20 ± 2.2[	10.47 ± 3.93
O3e*	9	8	11	0.86	0.27 ± 0.25	5.17 ± 1.3[	7.56 ± 2.30
O3e1*	11	11	10	0.91	0.26 ± 0.26	4.84 ± 2.4(	6.15 ± 3.37

No.: Number of samples

No.H: Number of different STR haplotypes

No.P: Number of loci found to be variable in size within each haplogroup

HD: Haplotype diversity

GD: Average gene diversity at each locus within each haplogroup

ALDA: Average number of loci with different alleles in pairwise comparison within each haplogroup

ACDA: Average cumulative number of differences in repeat size in pairwise comparison within each haplogroup

Note: Allele type of DYS385 was treated as two different loci with larger and smaller sizes respectively in this calculation.

Table 3 Average cumulative number of differences in repeat size in pairwise comparison within each haplogroup

Haplogroups	C1	C3*	D2a	(D2an	D2b*	022457	O1*	O2b*	O2b1	002611*	LINE1 de	021354*	O3e*	O3e1*
C1	5.53													
C3*	15.6	11.57												
D2a	14.90	16.62	7.84											
D2a(D2an)	13.1	14.58	11.37	5.73										
D2b*	13.28	13.93	11.31	11.37	7.66									
D2b1/-022457	14.2	15.23	9.81	8.30	8.39	4.52								
O1*	15.9	17.67	17.48	19.30	17.63	18.40	4.61							
O2b*	21.1	17.46	16.38	17.48	18.32	17.83	16.91	5.11						
O2b1	19.2	18.76	16.59	17.81	19.27	18.81	18.74	6.25	3.96					
O3/-002611*	20.9	19.42	17.30	16.00	16.84	15.81	15.85	14.73	15.67	8.00				
O3/-LINE1 del	20.3	20.29	17.70	16.62	17.51	16.90	16.54	16.48	17.30	7.33	7.62			
O3/-021354*	23.7	23.23	20.92	19.79	20.36	20.24	19.77	20.03	20.45	10.76	10.92	10.47		
O3e*	17.3	19.11	16.37	18.70	17.66	18.26	12.25	17.22	19.19	13.63	14.92	16.07	7.56	
O3e1*	23.9	26.86	21.68	24.88	23.32	24.58	19.38	19.77	21.33	16.45	16.35	17.04	16.27	6.15

Note: Average differences in total repeat size within haplogroup are also indicated by underlined numbers.

Table 4 Freqencies of binary haplogroups (%) estimated from Y-STR haplotypes. (Frequencies of Kanto and wetern Japan were calculated from present data after selecting corresponding samples.)

Haplogroup		Asahikawa	Kanto	Nagoya	'estern Japa	Okinawa	Korea	Taiwan
	Total sample no	201	137	207	97	87	317	183
C1		0	1.5	4.8	3.1	8	0.3	0.3
C3*		4.8	2.2	1.4	4.1	2.3	8.8	6
D1		3.2	0	0	0	0	0.3	0
D2a		8.9	8	4.8	5.2	2.3	0.6	0
D2b*		35.4	13.1	8.9	8.2	16.7	2.2	0.3
D2b1/-M125*		1.7	1.5	1	0	1.2	0	0
D2b1/-022457		17.6	25.5	19.6	13.4	20.1	0.9	0
	Total of D2	63.6	48.1	34.3	26.8	40.3	3.7	0.3
N/O		0	0	0.7	1	2.3	3.5	1.1
O1*		0.5	2.2	2.9	4.1	1.2	4.1	22.4
O2a*		3.1	0	0	1	1.2	1.1	6.3
O2b*		5.6	6.6	11.4	9.3	3.4	17.2	1.1
O2b1		7.2	24.1	20.8	26.8	19.5	12	1.1
O3*		1	0.7	0	0	0	0.3	1.5
O3c*		0.6	0.7	0.5	0	0	0.3	1.9
O3/-002611*		3.1	0.7	3.4	2.1	6.3	0.6	8.5
O3/-LINE1 del		0.5	1.5	5.1	5.2	4	6.2	11.7
O3/-021354*		2.1	3.6	6	6.2	5.7	9.8	5.6
O3e*		4.4	2.9	3.9	5.2	2.3	14.7	7.1
O3e1*		0	4.4	2.9	5.2	2.3	12.6	21.9
Q1		0	0.7	0	0	0	0.6	1.1
Not determined		0.5	0	1.9	0	1.2	3.8	2.2
	Total of O2b	12.8	30.7	32.2	36.1	22.9	29.2	2.2
	Total of O3	11.7	14.5	21.8	23.9	20.6	44.5	58.2
Reference		Sasaki & Dehiy	This study	Jchihi *et al*	This study	Jchihi *et al*	Shin *et al* .	Tsai *et al* .