Hum Genet (2006) 118: 695–707
DOI 10.1007/s00439-005-0106-9

ORIGINAL INVESTIGATION

Shi-Lin Li · Toshimichi Yamamoto
Takashi Yoshimoto · Rieko Uchihi · Masaki Mizutani
Yukihide Kurimoto · Katsushi Tokunaga · Feng Jin
Yoshinao Katsumata · Naruya Saitou

Phylogenetic relationship of the populations within and around Japan using 105 short tandem repeat polymorphic loci

Received: 1 July 2005 / Accepted: 1 November 2005 / Published online: 29 November 2005
© Springer-Verlag 2005

Abstract We have analyzed 105 autosomal polymorphic short tandem repeat (STR) loci for nine East and Southeastern Asian populations (two Japanese, five Han Chinese, Thai, and Burmese populations) and a Caucasian population using a multiplex PCR typing system. All the STR loci are genomewide tetranucleotide repeat markers of which the total number of observed alleles and the observed heterozygosity were 756 and 0.743, respectively, for Japanese populations. Phylogenetic analysis for these allele frequency data suggested that the Japanese populations are more closely related with southern Chinese populations than central and/or northern ones. STRUCTURE program analysis revealed the almost clearly divided and accountable population structure at $K=2$–6, that the two Japanese populations always formed one group separated from the other populations and never belong to different groups at $K \geq 3$. Furthermore, our new allele frequency data for 91 loci were analyzed with those for 52 worldwide populations published by previous studies. Phylogenetic and multidimensional scaling (MDS) analyses indicated that Asian populations with large population size (six Han Chinese, three Japanese, two Southeast Asia) formed one distinct cluster and are closer to each other than other ethnic minorities in east and Southeast Asia. This pattern may be the caviar of comparing populations with greatly differing population sizes when STR loci were analyzed.

Keywords Short tandem repeat · Population genetics · East Asian · Japanese · Phylogenetic tree · Polymorphism

Introduction

The Japanese archipelago was geographically dissociated from Asian continent around 12,000 years ago after the last glacial period (Aikens and Higuchi 1982). In the process of formation of the modern Japanese, there have been many migration events into Japan from continental Asia. There are many studies comparing mitochondrial DNA polymorphisms of Japanese and surrounding populations(e.g., Horai et al. 1996; Tanaka et al. 2004; Tajima et al. 2004). However, maternally inherited mitochondrial DNA has different characteristics in the context of its sex-specific modes of transmission compared to nuclear DNA. This restriction also applies to Y-chromosomes that are transmitted paternally (e.g., Hammer and Horai 1995).

Genotyping technologies have remarkably improved for many types of DNA markers recently. Especially, the numerous short tandem repeats (STRs), also known as microsatellites, have been used in phylogenetic analyses of extant human populations (e.g., Bowcock et al. 1994; Brinkmann et al. 1996; Perez-Lezaun et al. 1997; Chu et al. 1998; Rosenberg et al. 2002; Ayub et al. 2003; Zhivotovsky et al. 2003). These STR loci are clearly useful for studying the genetic relationships of closely related populations. However, there has been little study on the phylogenetic relationship around the Japanese population using these genome-wide STR DNA markers on autosomal chromosomes to date.

S.-L. Li · T. Yamamoto (✉) · T. Yoshimoto · R. Uchihi
M. Mizutani · Y. Kurimoto · Y. Katsumata
Department of Legal Medicine and Bioethics, Graduate School of Medicine, Nagoya University, Nagoya, Japan
E-mail: yamachan@med.nagoya-u.ac.jp
Fax: +81-52-7442121

K. Tokunaga
Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, 113-0033 Tokyo, Japan

N. Saitou
Division of Population Genetics, National Institute of Genetics, 411-8540 Mishima, Japan

F. Jin
Chinese Academy of Sciences, Institute of Genetics and Developmental Biology, 100101 Beijing, China

Takezaki and Nei (1996), by using computer simulation with microsatellite DNA loci, showed that Cavalli-Sforza and Edwards' (1967) chord distance D_C and Nei et al.'s (1983) D_A distance generally showed the higher correct topology (P_C) values than other distance measures in both the infinite-allele model (IAM) and the stepwise mutation model (SMM), whether the bottleneck effect exists or not. In addition, in case of SMM such as microsatellites, they showed that high P_C values (more than 80%) are obtained under the condition of small branch length, small sample size (~30), and a large number of loci (~100) with high heterozigosity (~0.8). Therefore, in the present study, we genotyped STR loci under conditions as close to these as possible, and examined the genetic relationship among human populations in east and Southeast Asia. Additionally, we examined the genetic relationship of the worldwide human populations by adding our data into the allele frequency data available at a web site (http://www.cmb.usc.edu/people/noahr//diversity.html#data).

Moreover, a novel computer program, STRUCTURE, was recently developed for an extensive analysis of population substructure and to identify population outliers (Pritchard et al. 2000). We applied this program to infer the population structure among human populations in east and Southeast Asia using the genotype data.

Materials and methods

DNA samples

Blood samples were collected with informed consent from Japanese living in the middle part of Honshu, the main island of Japan (Nagoya), Han Chinese living in five provinces (Shaanxi, Hunan, Guangdong, Fujian, and Beijing), Thai living in Bangkok, and Burmese living in Yangon (see Fig. 1). The DNA was extracted from blood samples by the usual organic extraction method or using some commercially available kits. Japanese DNA samples in Okinawa were previously collected for studying HLA alleles and haplotypes (Hatta et al. 1999). The DNA samples of Caucasian living in UK were kindly provided by Dr. Yuri E. Dubrova at the University of Leicester. Thirty-two DNA samples of each population were utilized for STR genotyping. In the DNA samples of these Han Chinese populations, a part of those of Guangdong (Huizhou region) and Fujian (Putian region), and those of Shaanxi (Xi'an region) and Hunan (Changsha region) were used for the studies reported previously by Roubinet et al. (2004) and Oota et al. (2002), respectively.

DNA amplification and genotyping

One hundred and five tetranucleotide STR markers on autosomal chromosome were selected from 168 STR loci in the screening set 8A (Research Genetics, Huntsville, AL, USA) by removing tri- and dinucleotide STRs, and STRs on X/Y-chromosome, as shown in Table 1, with their common motifs of repeat unit. Each one of their primer sets is labeled with any one of three different fluorescent-colored dyes (FAM, TET, HEX). The PCR amplification was performed using 48 sets of temporary multiplex, and the PCR products were multi-loaded with 15 panels by gel electrophoresis with a ABI PRISM 377 DNA Sequencer (PE Applied Biosystems, Foster city, CA, USA) based on a method described previously

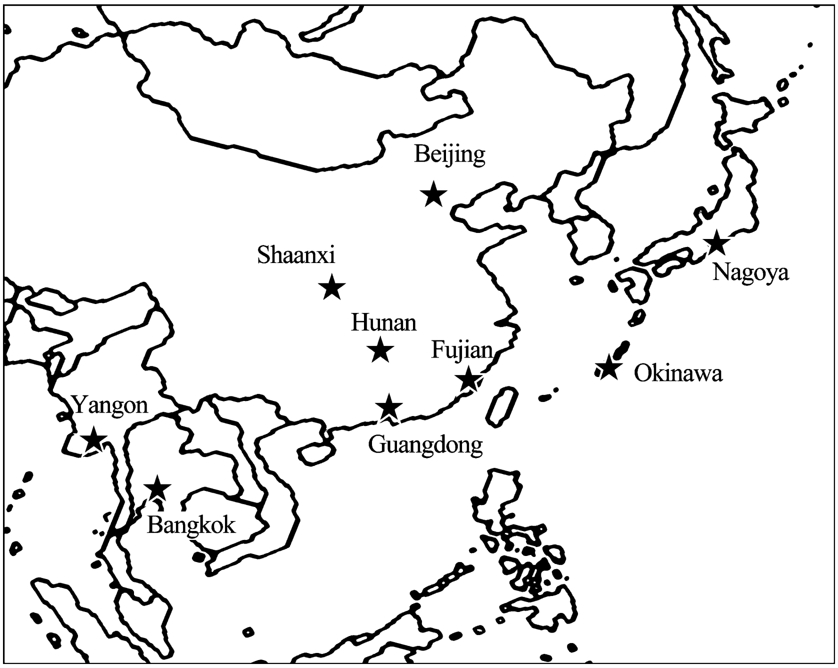


Fig. 1 Geographical location of the nine Asian populations analyzed in the present study

Table 1 Marker designation, type of repeat unit, number of alleles, G_{ST}, and heterozygosity values (Ht) for the 105 STR markers used in the present study

Repeat		9 Asian populations			9 Asian + English populations		
Marker	Unit	Alleles	G_{ST}	Ht	Alleles	G_{ST}	Ht
Chromosome 1							
D1S1612	GGAA	10	0.030	0.798	12	0.040	0.812
D1S1597	GATA	8	0.019	0.706	8	0.020	0.710
D1S552	GGAT	8	0.025	0.685	8	0.027	0.692
D1S2134	GATA	13	0.028	0.765	13	0.050	0.785
D1S1665	GATA	11	0.019	0.722	11	0.020	0.713
D1S534	GATA	15	0.022	0.791	15	0.028	0.804
D1S1679	GGAA	9	0.013	0.839	9	0.013	0.839
D1S518	GATA	9	0.015	0.780	9	0.025	0.794
D1S1660	GATA	8	0.019	0.802	8	0.023	0.805
D1S549	GATA	10	0.018	0.752	10	0.018	0.755
Chromosome 2							
D2S2976	GATA	15	*0.091	0.617	16	*0.101	0.658
D2S1400	GGAA	6	0.030	0.515	7	0.037	0.534
D2S1394	GATA	8	0.023	0.727	8	0.021	0.724
D2S2972	GATA	12	0.018	0.748	12	0.019	0.741
D2S1328	GATA	8	0.018	0.595	9	*0.072	0.649
D2S1399	GGAA	14	0.011	0.869	14	0.015	0.866
D2S1391	GATA	8	0.018	0.662	8	0.022	0.677
D2S1384	GATA	8	0.031	0.767	8	0.042	0.779
Chromosome 3							
D3S2387	GATA	19	0.015	0.863	20	0.018	0.866
D3S4545	GATA	13	0.017	0.777	15	0.032	0.793
D3S2432	GATA	10	0.017	0.791	12	0.019	0.792
D3S1766	GATA	9	0.024	0.725	9	0.024	0.732
D3S2460	GATA	9	0.015	0.745	9	0.016	0.748
D3S2427	GATA	18	0.022	0.888	19	0.025	0.889
Chromosome 4							
D4S2366	GATA	8	0.019	0.769	8	0.030	0.779
D4S2639	GATA	10	0.013	0.814	10	0.032	0.831
D4S1627	GATA	8	0.021	0.781	8	0.041	0.795
D4S1625	GATA	8	0.020	0.721	8	0.022	0.730
D4S1652	GATA	7	0.013	0.586	7	0.036	0.612
Chromosome 5							
D5S2845	GATA	9	0.027	0.772	9	0.026	0.770
D5S1470	GATA	10	0.015	0.797	10	0.017	0.802
D5S2500	GATA	11	0.018	0.780	11	0.018	0.785
D5S1505	GATA	8	0.018	0.818	8	0.021	0.821
D5S820	GATA	8	0.021	0.771	8	0.024	0.777
D5S1456	GATA	6	0.022	0.778	6	0.023	0.783
Chromosome 6							
D6S1053	GATA	8	0.016	0.796	8	0.015	0.796
D6S1056	GATA	10	0.014	0.843	10	0.017	0.844
GATA184A08	GATA	11	0.014	0.839	12	0.017	0.835
D6S1277	GATA	9	0.019	0.726	9	0.021	0.732
Chromosome 7							
D7S3056	GATA	8	0.023	0.742	8	0.024	0.734
D7S3051	GATA	11	0.013	0.788	11	0.017	0.790
D7S2846	GATA	6	0.027	0.700	6	0.025	0.702
D7S3046	GATA	12	0.027	0.844	12	0.028	0.846
D7S1842	GGAA	10	0.017	0.820	10	0.022	0.828
D7S1823	GATA	11	0.018	0.821	11	0.022	0.828
Chromosome 8							
D8S1106	GATA	8	0.017	0.653	9	0.023	0.663
D8S1477	GGAA	11	0.020	0.800	14	0.028	0.810
D8S1113	GGAA	8	0.036	0.701	8	0.035	0.710
D8S1132	GATA	10	0.017	0.851	10	0.016	0.851
D8S373	GATA	10	0.027	0.847	10	0.028	0.845
Chromosome 9							
D9S2169	GATA	7	0.015	0.675	7	0.014	0.679
D9S925	GATA	10	0.012	0.773	11	0.014	0.777
D9S1118	GATA	11	*0.055	0.850	11	*0.053	0.848
D9S934	GATA	10	0.016	0.786	10	0.017	0.788

Table 1 (Contd.)

Repeat		9 Asian populations			9 Asian + English populations		
Marker	Unit	Alleles	G_{ST}	Ht	Alleles	G_{ST}	Ht
Chromosome 10							
D10S1435	GATA	9	0.016	0.748	9	0.015	0.746
D10S1426	GATA	7	0.026	0.733	7	0.028	0.738
D10S1432	GATA	8	0.022	0.706	8	0.021	0.709
D10S677	GGAA	9	0.013	0.837	10	0.016	0.840
D10S1239	GATA	8	0.014	0.687	8	0.023	0.700
D10S1213	GGAA	12	0.023	0.723	12	0.024	0.734
D10S1248	GGAA	9	0.016	0.759	9	0.018	0.764
Chromosome 11							
D11S1984	GGAA	11	0.017	0.838	12	0.034	0.848
D11S1999	GATA	10	0.015	0.716	10	0.033	0.743
D11S2000	GATA	22	0.022	0.892	22	0.023	0.896
D11S4464	GATA	8	0.016	0.743	9	0.017	0.745
Chromosome 12							
D12S372	GATA	7	0.013	0.730	7	0.013	0.728
D12S391	GATA	14	0.012	0.846	14	0.017	0.853
D12S375	GATA	7	0.023	0.752	7	0.028	0.757
D12S1064	GATA	9	0.016	0.768	9	0.017	0.772
PAH	TCTA	9	0.029	0.745	9	0.028	0.750
D12S395	GATA	10	0.019	0.668	10	0.030	0.687
Chromosome 13							
D13S894	GATA	7	0.021	0.623	8	0.024	0.637
D13S317	GATA	8	0.017	0.801	8	0.027	0.807
D13S796	GATA	11	0.012	0.808	11	0.014	0.808
Chromosome 14							
D14S1280	GATA	7	0.013	0.648	7	0.011	0.650
D14S306	GATA	9	0.012	0.770	9	0.016	0.772
D14S617	GGAA	10	0.025	0.763	10	0.027	0.771
D15S822	GATA	19	0.018	0.838	19	0.031	0.851
Chromosome 15							
D15S643	GATA	14	0.011	0.835	14	0.012	0.837
D15S657	GATA	8	0.013	0.830	8	0.016	0.829
D15S642	GATA	12	0.016	0.755	13	0.030	0.774
Chromosome 16							
D16S764	GATA	8	*0.111	0.704	8	*0.102	0.700
D16S753	GGAA	10	0.015	0.762	10	0.021	0.772
D16S3253	GATA	9	0.016	0.741	9	0.022	0.745
D16S2624	GATA	8	0.023	0.706	8	0.030	0.715
D16S539	GATA	7	0.038	0.789	7	0.039	0.789
Chromosome 17							
D17S1308	GTAT	6	0.030	0.577	7	0.037	0.596
D17S1303	GATA	9	0.014	0.724	9	0.020	0.723
D17S1293	GGAA	11	0.014	0.845	11	0.016	0.848
D17S1290	GATA	16	0.012	0.833	16	0.016	0.835
D17S1301	GATA	7	0.013	0.697	7	0.013	0.698
Chromosome 18							
D18S877	GATA	7	0.037	0.697	7	0.035	0.696
D18S1364	GATA	10	0.019	0.843	10	0.022	0.840
Chromosome 19							
D19S591	GATA	7	0.020	0.743	7	0.027	0.752
D19S586	GATA	8	0.018	0.657	8	0.016	0.661
D19S433	GGAA	13	0.034	0.814	13	0.033	0.810
D19S246	GATA	13	0.016	0.788	13	0.027	0.803
Chromosome 20							
D20S482	GATA	10	0.023	0.716	11	0.022	0.720
D20S470	GGAA	15	0.021	0.872	16	0.024	0.872
D20S481	GATA	9	0.022	0.723	9	0.028	0.741
D20S480	GATA	10	0.015	0.808	10	0.015	0.806
Chromosome 21							
D21S1432	GATA	9	0.013	0.741	9	0.016	0.740
D21S2055	GATA	17	0.022	0.885	18	0.024	0.888
Chromosome 22							
D22S689	GATA	9	0.027	0.798	9	0.026	0.798
D22S683	GATA	23	0.019	0.838	24	0.021	0.848

*$G_{ST} > 0.05$

The numbers and both values were calculated separately without/with the Caucasian population. The markers are arranged according to their location on the chromosomes

(Mizutani et al. 2001) with minor modification. Fragment sizes for each STR loci were determined on the basis of known internal lane size standards using software GeneScan Analysis (Version 3.1), and their genotypes were determined by comparing the size data analyzed by Mizutani et al. (2001). These 105 STR markers were widely distributed in all autosomes as listed in Table 1; varying 10 loci on chr. 1–2 loci on chr. 18, 21, and 22.

Statistical analyses

Tests for Hardy–Weinberg equilibrium (HWE) were performed using a homozygosity test (Weir 1992), a likelihood ratio test (Chakraborty et al. 1991), and an exact test (Guo and Thompson 1992). The observed Ht and the unbiased estimates of expected Ht were calculated according to Edwards et al. (1992). The G_{ST} values and Ht were estimated using DISPAN (downloaded from http://mep.bio.psu/downlods/dispan.zip).

The genetic distances were calculated from the allele frequency data at all the 105 STR loci by D_{A} (Nei et al. 1983) distance with the NJBAFD (downloaded from http://iubio.bio.indiana.edu/soft/molbio/evolve/njbafd/), and D_{C} (Cavalli-Sforza and Edwards 1967) and θ_{W} (F_{ST}) distance (Reynold et al. 1983) with the PHYLIP 3.5c (Felsenstein 1995), and then phylogenetic trees were constructed by using the Neighbor-Joining (NJ) method (Saitou and Nei 1987) using the MEGA Version 2.1 (Kumar et al. 2001). Bootstrap values were obtained based on 1,000 replications. A phylogenetic tree by using the D_{A} distance and NJ method with the NJBAFD was also constructed from the allele frequency data at 91 STR loci, which are in common among 105 loci in the present study, in total 61 worldwide populations by adding our ten populations into 52 world populations obtained from the literature reported previously (Rosenberg et al. 2002). However, when the allele frequency data were downloaded from http://www.cmb.usc.edu/people/noahr//diversity.html, those for 51 populations were obtained because the data for Han population were calculated by combining those in US Han and northern China Han populations.

Since the data sizes were slightly different between both the data sets, they were matched between both databases without contradiction by comparing the allele frequency distribution at each locus in Japanese and Chinese in our database to those in Japanese and Han populations in the database of Rosenberg et al. (2002).

The MDS analysis based on a 10×10 matrix of pairwise D_{A} distance values calculated above and F_{ST} (Latter 1972) calculated with the MICROSAT Version 2.0 software at 105 STR loci in our ten populations was performed using the SPSS 12.0 software package (SPSS Inc., Chicago, IL, USA). Similarly, the MDS analyses on 61×61 D_{A} and F_{ST} distance matrix at 91 STR loci in the 61 worldwide populations was performed.

To conduct an extensive analysis of population substructure and to identify population clusters, we also applied the computer program STRUCTURE 2.0 (Pritchard et al. 2000) which examines the populations at the level of the individual on the basis of genotype data at 105 loci. The program was run for 20,000 iterations after a burn in of length 20,000 with an admixture model of correlated allele frequencies. Models in which there are K population (where K may be the unknown number clustering those populations) assigned from 2 to 7 in this study are assumed, which are estimated on the basis of their individual genotypes in each population. To more thoroughly display results produced by the genetic clustering program STRUCTURE, we used program DISTRUCT (downloaded from http://www-hto.usc.edu/»noahr/distruct.html) to make detailed graphical figure, having used STRUCTURE to generate the population Q-matrix, which was created by averaging membership coefficients of each cluster across individuals for each population.

Like the phylogenetic tree analysis mentioned above, the genotype data at the 91 STR loci in our ten populations were combined with those in the 18 East Asia populations and Uygur population out of the Central/South Asia populations (downloaded from http://www.cmb.usc.edu/people/noahr//diversity.html) (Rosenberg et al. 2002) by matching the size data, and STRUCTURE and DISTRUCT analyses were also performed in the total 29 populations.

Results

We genotyped all these 105 tetranucleotide STR loci as described previously (Mizutani et al. 2001). A total of 320 individuals were typed from the following ten human populations: two Japanese, five Han Chinese, one Burmese, one Thai, and one English population. The number of alleles at each STR loci observed in these ten populations was counted as shown in Table 1, and the total number of alleles and number of unique alleles observed in each population and average heterozygosities (±SE) calculated in each population are also shown in Table 2.

Deviation from the HWE was checked using three kinds of statistical tests: homozygosity (Homo) test, likelihood ratio (LR) test, and exact (Ex) test. The numbers of loci at which the allele frequency distributions were significantly deviated from HWE ($P < 0.05$) with the three kinds of tests are summarized in Table 3. Only two loci at which the distributions were significantly deviated from HWE ($P < 0.05$) with all the three tests were observed in more than one population: D2S1400 in the Okinawa and Bangkok populations, and D14S617 in the Shaanxi and Hunan populations. The mean of the total number of loci significantly deviated from HWE was 25.5 loci (8.1%).

These values were comparatively low even though the number (64) of chromosomes analyzed in each

Table 2 Total number of alleles, number of unique alleles, and average heterozygosities (±SE) for the ten human populations

Population	Total alleles	Unique alleles	Average heterozygosity (±SE)
Japan			
Nagoya	750	12 (1.60%)	0.7698±0.0082
Okinawa	763	17 (2.23%)	0.7627±0.0082
China			
Beijing	765	16 (2.09%)	0.7800±0.0071
Shaanxi	760	6 (0.79%)	0.7667±0.0078
Hunan	747	8 (1.07%)	0.7706±0.0078
Fujian	751	16 (2.13%)	0.7648±0.0093
Guangdong	766	6 (0.78%)	0.7637±0.0098
Southeast Asia			
Bangkok	761	18 (2.37%)	0.7774±0.0081
Yangon	787	20 (2.54%)	0.7774±0.0081
Europe			
England	795	27 (3.40%)	0.7988±0.0059
10 populations	1084		0.7732

Values in *parentheses* indicate percentage of unique alleles in each population

Table 3 The number of loci observed as significant deviations from Hardy-Weinberg equilibrium (HWE) ($P < 0.05$) with three tests at the 105 STR loci in the ten populations in the present study

Test	Japanese		Chinese					Southeast Asian		Caucasian
	Nagoya	Okinawa	Beijing	Shaanxi	Hunan	Fujian	Guangdong	Bangkok	Yangon	England
H only	1	5	5	2	3	6	4	3	3	6
L only	1	4	1	1	1	5	0	2	2	0
E only	2	0	0	2	1	0	0	3	0	0
H and L	0	0	0	0	0	0	2	1	0	0
H and E	0	1	0	1	2	2	1	1	1	2
L and E	4	1	1	8	4	1	5	5	3	2
H and L and E	2	3	2	4	7	1	3	4	51	
H	3	9	7	7	12	9	10	9	9	9
L	7	8	4	13	12	7	10	12	10	3
E	8	5	3	15	14	4	9	13	9	5
H or L or E	18 (5.7%)	22 (7.0%)	14 (4.4%)	35 (11.1%)	38 (12.1%)	20 (6.3%)	29 (9.2%)	34 (10.8%)	28 (8.9%)	17 (5.4%)

H Homozigosity test, *L* Likelihood ratio test, *E* Exact testValues in *parentheses* indicate percentage of total loci significantly deviated from HWE in each population

population was relatively small compared with the number of alleles (locus mean = 10.3). These findings indicate that there was no contingent event, sampling error, or population substructure. Accordingly, it was considered that these allele frequency data would be reliable and the whole data were used for the following analyses.

For statistical properties for the 105 STR loci analyzed in the present study, their G_{ST} as a measure of allelic diversity and Ht for each locus were calculated as shown in Table 1. The G_{ST} value averaged over all loci was 0.0254 in the ten populations, and 0.0209 in the nine Asian populations, close to the values reported previously in a study on northern Pakistan populations (G_{ST} values: 0.03; Mansoor et al. 2004), but about six times less than the values on worldwide populations (0.15; Ayub et al. 2003) using microsatellite markers. In comparison between the nine and the ten populations, G_{ST} values of the latter were slightly higher than those of the former at almost all the loci (0.004 on the average), but extremely high at only one locus (0.054 higher; D2S1328), which indicates the extremely different allele frequency distribution between East Asians and Caucasians at the locus. The 105 STR loci showed significantly higher heterozygosities in the English population than those in all the East Asian populations. The Southeast Asian populations had a higher level of variation than the Japanese and Chinese populations.

A phylogenetic tree, based on D_A distance using the NJ method, provides strong evidence for the closest relationship among the two regional Japanese populations (bootstrap: 100%), and also for closer relationship between the Japanese and Southern Han Chinese populations (Fujian and Guangdong) (bootstrap: 84%) as shown in Fig. 2. The Beijing population located in Northern China is clustered with the two central Han Chinese including Hunan and Shaanxi populations. The branching pattern in which the Southern Chinese and Japanese populations were in different clusters from the Northern and Central Chinese populations was supported by the somewhat high bootstrap values of 67%, suggesting that authenticity of this pattern might be considered. The Southeast Asian populations were not clustered with each other, while the Thai population was close to the Chinese and Japanese populations with high bootstrap values of 98%.

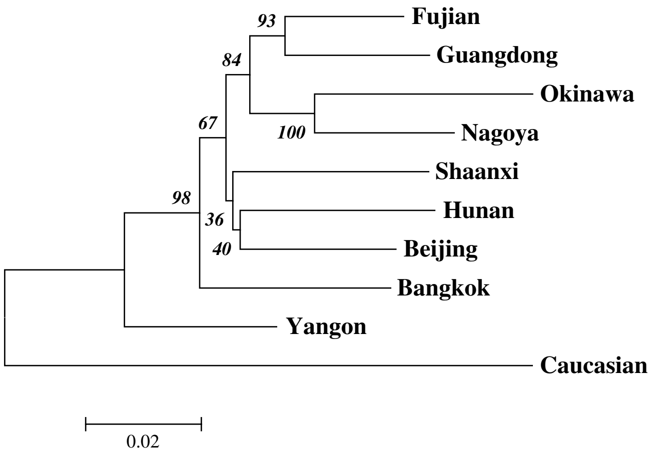


Fig. 2 A Neighbor-joining (*NJ*) tree showing the relationships of ten human populations examined in the present study on the basis of D_A distances calculated from the allele frequencies at 105 tetranucleotide STR loci. The scale for the distance is shown *bottom left*. Bootstrap values are provided at each fork of branches as *italic numbers*

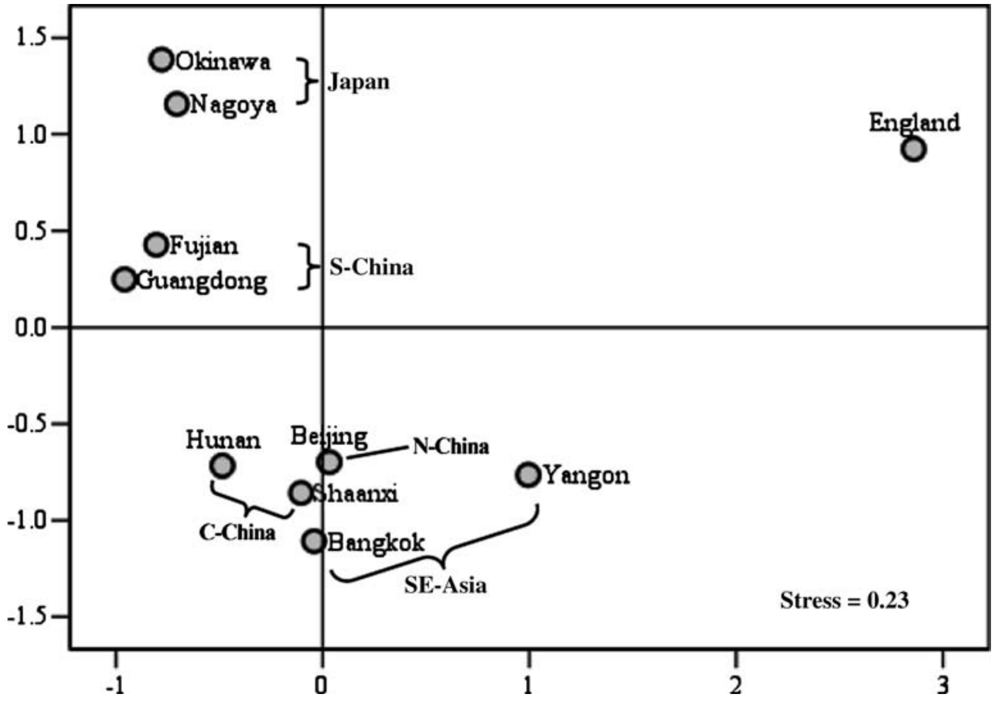
We also constructed two other NJ trees using different genetic distances, D_C and F_{ST}. These phylogenetic trees showed the same topology as that on the basis of D_A distance at all with slightly different branch lengths between each population (trees not shown), which suggested that the phylogenetic relationship among these ethnic and/or regional populations is authentic.

Multidimensional scaling analysis (MDS) of pairwise D_A distance values revealed groups of genetically related populations (Fig. 3), same as the phylogenetic tree as mention above. One of the most characteristic findings of the present analysis was that the Japanese groups had somewhat closer genetic affinities with the southern Chinese populations than northern Chinese populations, and another was that the distribution was slightly different from their geographical relationship. The low "stress" value (0.23) of the MDS plot indicated a good fit between the two-dimensional graph and the original distance matrix. The MDS of pairwise F_{ST} distance showed two groups, the Japanese populations and the Chinese and Southeast Asian populations. The distance between the former group and the latter group was almost equal to that between the former group and the English, and that between the latter and the English, and the shape linking the two groups and English was an almost regular triangle despite the lower stress value (0.16) (figure not shown).

K values (number of "populations") were assigned from 2 to 6 for STRUCTURE and DISTRUCT programs in Fig. 4a. At $K=2$, the two clusters, namely English and Asians, can be clearly seen. At $K=3$, however, the two Japanese populations were primarily separated from the other Asian populations. At $K=4$, the Beijing population is separated from other non-Japanese Asian populations (yellow color), but this "population" component also slightly appeared in other Asian populations.

The new "population" (light blue color) coming into view at $K=5$ occupied the major portion in Central Chinese populations (Hunan and Shaanxi), and distributed at some degrees in Southern Chinese and Southeast Asian populations. At $K=6$, Southern Chinese (Guangdong and Fujian) was newly occupied by new "population" (purple color), and it also constituted a little part of the other Asian populations. Interestingly, the "population" (green color) dominated in English

Fig. 3 Multidimensional scaling (*MDS*) plot of ten human populations analyzed in the present study, based on D_A genetic distances calculated from 105 autosomal tetranucleotide STR loci



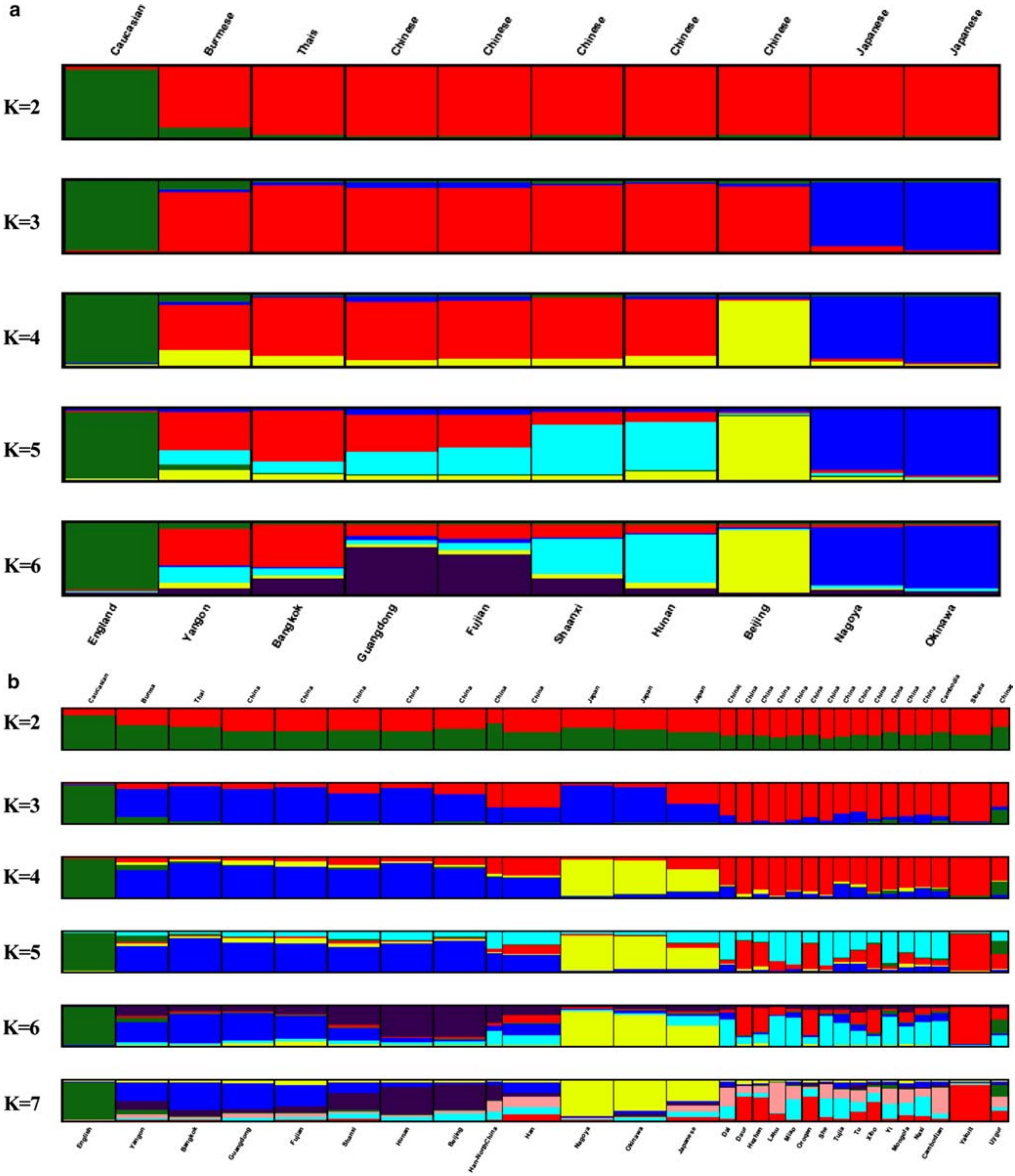


Fig. 4 Population sub-structures of (**a**) the ten populations in this study and of (**b**) the combined 29 populations with Rosenberg et al.'s 19 populations are estimated by genotype data of 105 and 91 tetranucleotide STR markers, respectively, using the DISTRUCT program with the assistance of the STRUCTURE program. Each population is separated by a *vertical line*, which is sectioned into *K* colored segments that represent the proportion of membership of each pre-defined population in *K* clusters. The populations are labeled below the figure, with their regional affiliations above the figure

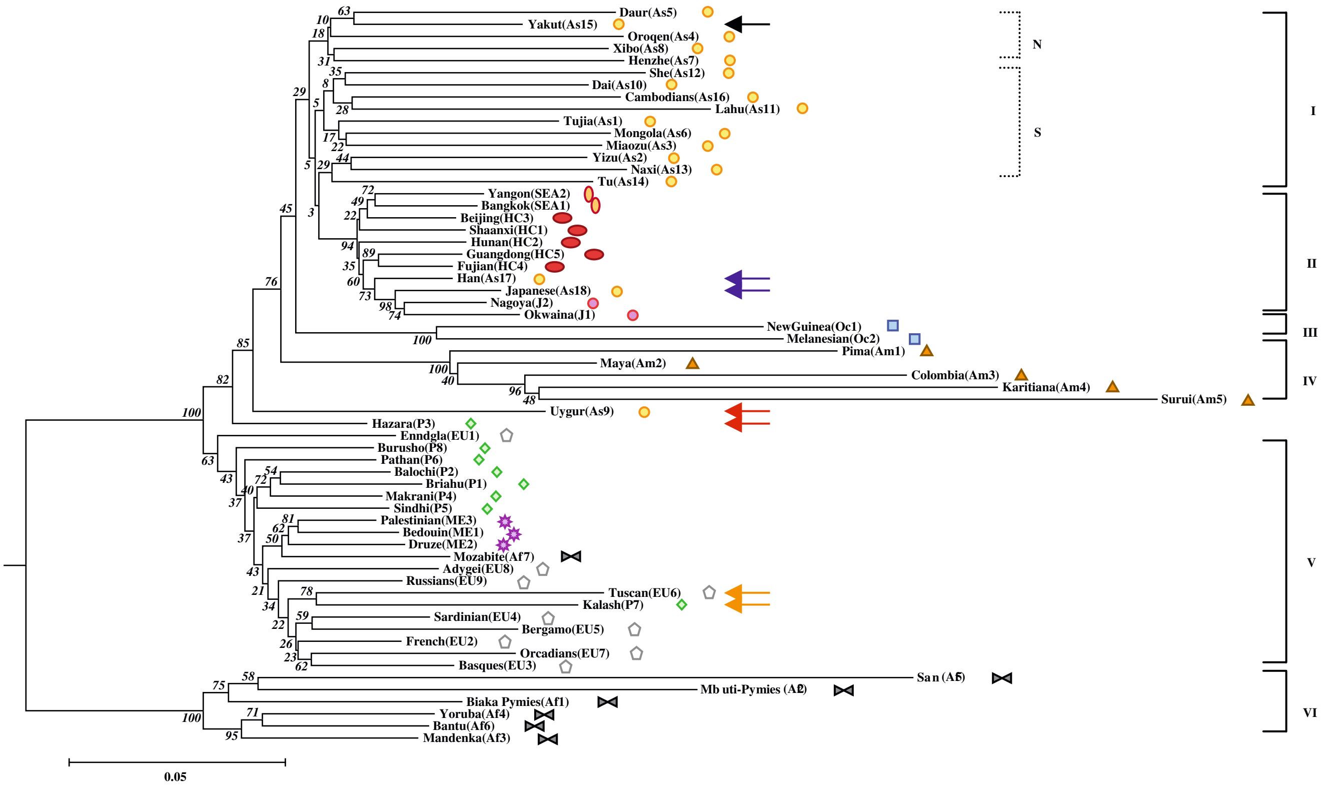


Fig. 5 Neighbor-joining (*NJ*) tree for 61 worldwide human populations including the present study data and the data previously reported (Rosenberg et al. 2002) on the basis of D_A distances calculated from the allele frequencies at 91 tetranucleotide STR loci. The scale for the distance is shown *bottom left*. Bootstrap values are provided at each fork of branches as *italic numbers*. Our own data are the following ten populations: two Japanese populations in pink-colored *circles* (J1: Okinawa, J2: Nagoya), five Han Chinese populations in red-colored *ellipses* (HC1: Shaanxi, HC2: Hunan, HC3: Beijing, HC4: Fujian, HC5: Guangdong), two Southeast Asian populations in orange-colored *ellipses* (SEA1: Bangkok, SEA2: Yangon), and one European population in white-colored *pentagon* (EU1: England). The remaining 51 population data were from Rosenberg et al. (2002): 18 Asian populations in yellow colored *circles* (As1: Tujia, As2: Yizu, As3: Miaozu, As4: Oroqen, As5: Daur, As6: Mongolia, As7: Hezhen, As8: Xibo, As9: Uygur, As10: Dai, As11: Lahu, As12: She, As13: Naxi, As14: Tu, As15: Yakut, As16: Cambodian, As17: Han, As18: Japanese), two Oceania populations in light blue-colored *squares* (Oc1: New Guinea, Oc2: Melanesian), five native American populations in orange-colored *triangles* (Am1: Pima, Am2: Maya, Am3: Colombia, Am4: Karitiana, Am5: Surui), eight Pakistani populations in light green-colored *squres* (P1: Brahui, P2: Balochi, P3: Hazara, P4: Makrani, P5: Sindhi, P6: Pathan, P7: Kalash, P8: Burusho), eight European populations in white-colored *pentagons* (EU2: French, EU3: Basques, EU4: Sardinian, EU5: Bergamo, EU6: Tuscan, EU7: Orcadians, EU8: Adygei, EU9: Russians), three Middle-East populations in purple-colored star-like *shapes* (ME1: Bedouin, ME2: Druze, ME3: Palestinian), and seven African populations in gray colored two *small triangles* (Af1: Biaka·Pygmies, Af2: Mbuti·Pygmies, Af3: Mandeka, Af4: Yoruba, Af5: San, Af 6: Bantu, Af7: Mozabite). *Arrows* in various colors pointed some populations to help the explanations in Discussion

slightly existed only in the Burmese population at all the K values. Furthermore, at $K=6$, the Japanese populations slightly included "population" (purple color) that is dominant in the southern Chinese populations.

Discussion

The phylogenetic relationship of Japanese with other Asian populations

In the present study, the phylogenetic tree (Fig. 2) and MDS scattergram (Fig. 3) showed that the Japanese populations were somewhat closer to the southern Chinese populations than the northern and central Chinese populations. This suggests that both the present Japanese and the present southern Han Chinese share some common features with each other. Therefore, there is a possibility that the southern Chinese contributed more to the present day Japanese population than the northern Chinese.

This pattern is somewhat different from some previous results using classic markers (Saitou et al. 1994; Nei 1995; Omoto and Saitou 1997) and HLA markers (Saitou et al. 1992; Hatta et al. 1999; Bannai et al. 2000), where Japanese were more closely related to North East Asian populations. The reason for this discrepancy is not clear.

Hanihara (1991) proposed the dual structure model on peopling of Japanese. According to this model, Ainu and Okinawa Japanese (Ryukyuan) originated from Jomonese, of which origins have been argued to

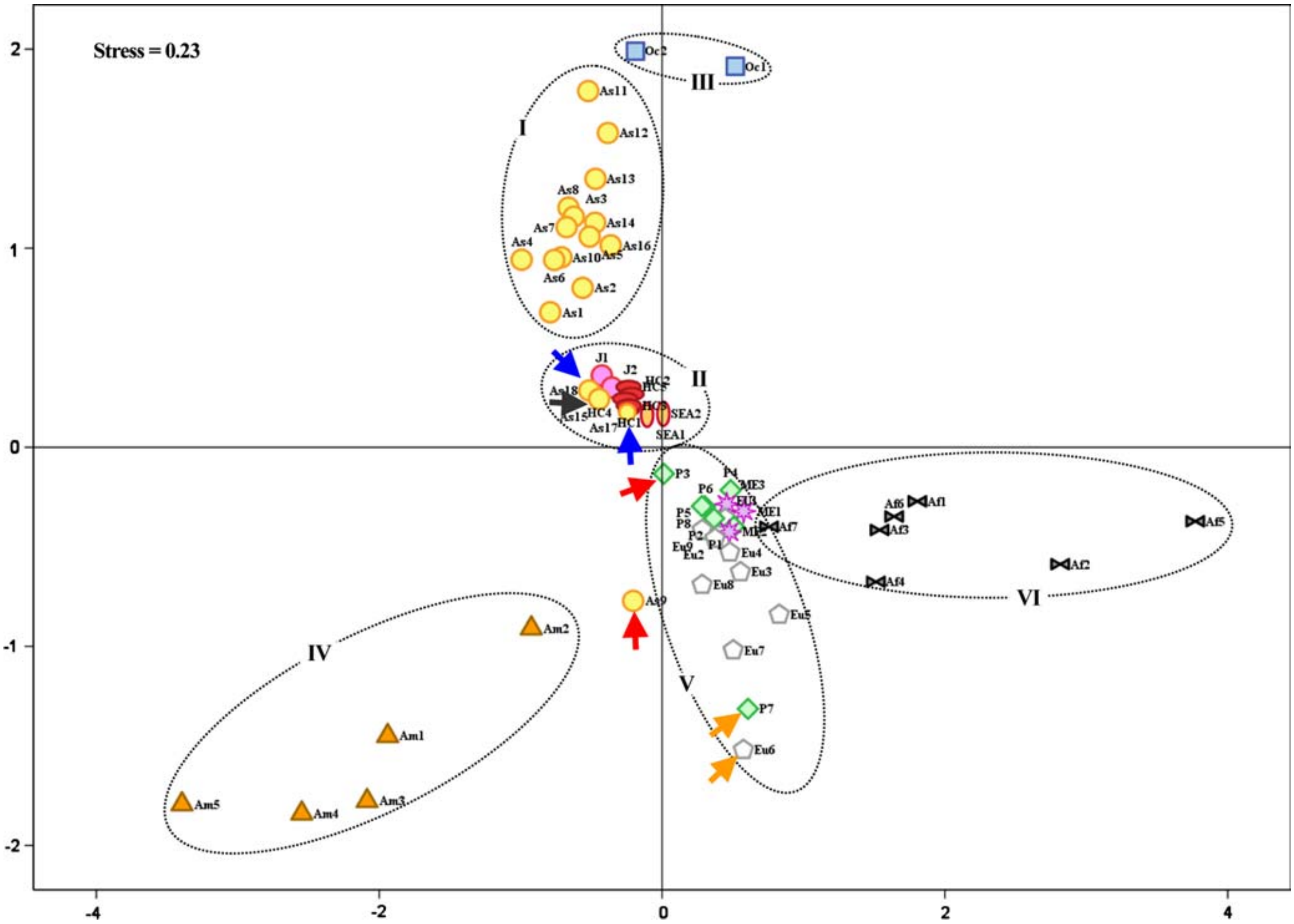


Fig. 6 Multidimensional scaling (*MDS*) plot of 61 worldwide human populations, based on D_A genetic distances calculated from 91 autosomal tetranucleotide STR loci. Population ID's were the same as in Fig. 5. *Arrows* in various colors pointed some populations to help the explanations in Discussion

give rise to southern part of East Asia (Hanihara 1991). Contemporary Okinawa Japanese, however, are genetically much closer to Mainland Japanese than Ainu (Omoto and Saitou 1997; Tajima et al. 2002). In the analysis using six radiation groups from mtDNA phylogenetic network, the frequency patterns were similar between Cantonese and Ryukyu Japanese, and between Korean and mainland Japanese (Oota et al. 1999). Furthermore, the analyses using HLA genes and haplotypes suggested a recent gene flow to Okinawa from south China (Hatta et al. 1999; Bannai et al. 2000; Tokunaga et al. 2001).

STRUCTURE analysis for 18 east Asian populations reported previously (Rosenberg et al. 2002) showed that Japanese shared a greater degree of similarity with small ethnic minority populations of Northern China (Daur, Hezhen, and Oroqen) than with southern Chinese populations, and that Japanese were closer to Han of northern China than Han people migrated to the USA. Accordingly, if membership shared with small ethnic groups from northern China and Japanese in Rosenberg et al. (2002) was the same as membership in blue in the present study, the modern Japanese may consist of people who originated from northern China and blended with people affected by those from southern part of China while migrating through the Korean peninsula, and not affected by those from around Beijing. To obtain more information, STRUCTURE and DISTRUCT analyses were also performed using the genotype data at the 91 STR loci in the total 29 populations by combining the data in our ten populations with those in the 18 East Asia populations and Uygur population out of the Central/South Asia populations (Rosenberg et al. 2002). As shown in Fig. 4b, the characteristic memberships of the three populations in Japan (in yellow) were also observed in the small ethnic minority populations of Northern China (Daur, Hezhen, Oroqen, and Mongolia), same as in the southern Han Chinese population (Guangdong and Fujian) at somewhat higher ratio than the other populations ($K = 4$–7). However, since Ainu Japanese and Korean were not analyzed in this study, this may be only a speculation. Further study is needed to clear up the relationship between Okinawa Japanese and the other Japanese populations including Ainu.

Combined phylogenetic analysis with published data

We also examined the phylogenetic relationship of these nine East and Southeast Asian populations newly examined in this study and the already published worldwide population data (Rosenberg et al. 2002). The NJ tree based on D_A genetic distance and the MDS scattergram were constructed from the allele frequency data at 91 STR loci for the total of 61 worldwide populations, as shown in Figs. 5 and 6, respectively. The reasons why we reanalyzed a part of their data by adding our data are: (1) Rosenberg et al. (2002) did not present any phylogenic tree, (2) Rosenberg et al. (2002) analyzed their data by mixing di-, tri-, and tetranucleotide repeat STR markers with extremely different mutation rates, and (3) The genetic distances D_A and D_C show more correct topology than other distance measures (Rogers' D_R: 1972, Nei's D_S: 1972, Latter's F_{ST}: 1972 and so on) in SMM under such conditions where the number of samples are about 10–30 using about 100 STR markers with about 0.80 of the heterozygosities, whether the bottleneck effect exists or not, according to a simulation study (Takezaki and Nei 1996).

Although a NJ tree construction and MDS analysis based on F_{ST} were performed, in fact, however, since the NJ tree showed very similar topology only with different blanching lengths, and the MDS plots were also very similar with those from 377 loci reported previously (Zhivotovsky et al. 2003) but with lower 0.16 in stress value, only the NJ tree and MDS plots based on D_A distance are shown here.

Six major clusters (I–VI) can be recognized both in the NJ tree (Fig. 5) and in the MDS scattergram (Fig. 6). The nine East and Southeast Asian populations examined in the present study all belong to cluster II. Japanese (As18) and Han (As17) populations studied by Rosenberg et al. (2002) are also included in cluster II (pointed by two blue arrows in Figs. 5 and 6), and the Japanese (As18) located closer to Nagoya (J2) and Okinawa (J1), both in the NJ tree and in the MDS scattergram.

The 15 East Asian small ethnic populations formed cluster I, and divided into two sub-clusters (N and S) in which the former included some small ethnic groups of northern China (Daur, Oroqen, Xibo, and Hezhen) and Yakut in Siberia, while the latter included those of southern China such as She, Dai, Tujia, and Lahu. It is, however, ambiguous whether the cluster of a small ethnic group including Yizu, Naxi, and Tu distributed in the cluster I or II because of its low bootstrap values and long branch length. The lengths of branches in the populations of cluster I were more than twice longer than those of cluster II. This observation suggested that the cluster II populations are very close with each other, and that the cluster I populations are affected by bottleneck effect or depend on the small number of sampling.

Interestingly, even though the Yakut population (As15 marked by a black arrow in Figs. 5 and 6) inhabits near the Lake Baikal and in the basin of the Middle Lena River in Siberia, it belonged to cluster I-N in the NJ tree, but was very close to cluster II in the MDS scattergram. The pattern observed in the MDS scattergram is consistent with Matsumoto (1988) who concluded that the origin of Japanese is near the Lake Baikal because of the very similar allele distribution estimated from immunoglobin phenotypes (Km and Gm) after examining many East Eurasian populations. However, since the Yakut population belonged to cluster I-N in the NJ tree, its origin remains ambiguous.

The other clusters consisted of Oceanian (cluster III), Native American (cluster IV), non-Asian Eurasian (cluster V), and African (cluster VI), respectively. Interestingly, however, the NJ tree shows that Uygur (northwestern China) and Hazara (Pakistan) populations did not belong to the six major clusters, instead located between clusters IV and V. In the MDS scattergram, the Uygur population (As9) located in the border of clusters V and IV, while the Hazara population (P3) located in the Eurasia cluster (V) at the closest position to the Asian cluster (II). These two populations are marked with two red arrows both in Figs. 5 and 6. Positions of the Uygur population in these figures suggests that its ancestral population was shared with that of native American and admixed with Eurasian populations later. Hazara population can be explained as the descendants of Middle-East Asian who slightly admixed with some East Asian populations. The latter possibility was supported by a study on Y-chromosomal DNA variation in Pakistan (Qamar et al. 2002).

More interestingly, Kalash (P7) and Tuscan (Eu6) ethnic groups made a sub-cluster with Italian and French ethnic groups (Bergamo, Sardinian, French, Basques, and Orcadians) in the Eurasian cluster V in the NJ tree, and they are located at the closest position with each other in the MDS plot (pointed by two orange arrows in Figs. 5 and 6). These results may be consistent with Kalash people's oral tradition that they are the descendants of the Alexander the Great's army (Lines 1999).

Since Y-STR haplotypes and mtDNA were transmitted along only male and female lineages, respectively, no test can be performed for deviation in samplings, especially for small number of samples in small population size. However, in the case of autosomal markers, at least tests for HWE exist to confirm the deviation because of their biparental inheritance. In the present study, after the samples collected from each Asian population were confirmed as no deviation from HWE at almost all autosomal STR loci, those reliable and reasonable allele frequency data could be provided for not only the phylogenetic tree analysis, but also the structure analysis as those populations by the STRUCTURE–DISTRUCT program, but not that as individuals by only Structure program (Rosenberg et al. 2002). Accordingly, the results of the distance based methods (NJ tree and MDS plots based on D_A distance as a SMM model) as one of bottom-up procedures could be very similar to one of the model based methods (configuration of the memberships constructed by STRUCTURE–DISTRUCT program) as one of top-down procedures. In short, Japanese (both Hondo- and Okinawa Japanese) are isolated and distinguishable from other east and Southeast Asian population, and slightly more affected by southern part of Chinese than by northern or middle part of Chinese. However, with regard to studies on the peopling of Japanese or the origin of Japanese, further studies are necessary with these methods in the present study using more detailed samples from other ethnic, regional, or national populations within and around Japan such as Ainu, Koreans, Mongolia, Eskimo, and so on.

Acknowledgements We are grateful to Prof. Sirirurg Songsivilai, Dr. Hla Hla Htay, and Dr. Yuri E. Dubrova for providing Thais, Burmese, and Caucasian DNA samples, respectively. We also thank Dr. Naoko Takezaki for helpful technical assistance and comments. This work was in part supported by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (12012204 to TY and 11307008 to YK).

References

Aikens CM, Higuchi T (1982) Prehistory of Japan. Academic Press, New York

Ayub Q, Mansoor A, Ismail M, Khaliq S, Mohyuddin A, Hameed A, Mazhar K, Rehman S, Siddiqi S, Papaioannou M, Piazza A, Cavalli-Sforza LL, Mehdi SQ (2003) Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. Am J Phys Anthropol 122:259–268

Bannai M, Ohashi J, Harihara S, Takahashi Y, Juji T, Omoto K, Tokunaga K (2000) Analysis of HLA genes and haplotypes in Ainu (from Hokkaido, northern Japan) supports the premise that they descent from Upper Paleolithic populations of East Asia. Tissue Antigens 55:128–139

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Brinkmann B, Sajantila A, Goedde HW, Matsumoto H, Nishi K, Wiegand P (1996) Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci. Eur J Hum Genet 4:175–182

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257

Chakraborty R, Fornage M, Gueguen R, Boerwinkle E (1991) Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population. In: Burke T, Dolf G, Jeffreys AJ, Wolff R (eds) DNA fingerprinting: approaches and applications. Birkhauser Verlag, Berlin, pp 127–143

Chu JY, Huang W, Kuang SQ, Wang JM, Xu JJ, Chu ZT, Yang ZQ, Lin KQ, Li P, Wu M, Geng ZC, Tan CC, Du RF, Jin L (1998) Genetic relationship of populations in China. Proc Natl Acad Sci USA 95:11763–11768

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241–253

Felsenstein J (1995) Phylogeny Inference Package (PHYLIP).v.3.57c. Universuty of Washington, Seattle

Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48(2):361–372

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951–962

Hanihara K (1991) Dual structure model for the formation of the Japanese. Jpn Rev 2:1–33

Hatta Y, Ohashi J, Imanishi T, Kamiyama H, Iha M, Simabukuro T, Ogawa A, Tanaka H, Akaza T, Gojobori T, Juji T, Tokunaga K (1999) HLA genes and haplotypes in Ryukyuans suggest recent gene flow to the Okinawa Islands. Hum Biol 71:353–365

Horai S, Murayama K, Hayasaka K, Matsubayashi S, Hattori Y, Fucharoen G, Harihara S, Park K-S, Omoto K, Pan I-H (1996) mtDNA polymorphism in Asian populations, with special reference to the peopling of Japan. Am J Hum Genet 59:579–590

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Lines M (1999) The Kalasha people of north-western Pakistan. Emjay books International, Peshawar

Latter BDH (1972) Selection in finite populations with multiple alleles III Genetic divergence with centripetal selection and mutation. Genetics 70:475–490

Mansoor A, Mazhar K, Khaliq S, Hameed A, Rehman S, Siddiqi S, Papaioannou M, Cavalli-Sforza LL, Mehdi SQ, Ayub Q (2004) Investigation of the Greek ancestry of populations from northern Pakistan. Hum Genet 114:484–490

Matsumoto H (1988) Characteristics of Mongoloid and neighboring populations based on the genetic markers of human immunoglobulins. Hum Genet 80:207–218

Mizutani M, Yamamoto T, Torii K, Kawase H, Yoshimoto T, Uchihi R, Tanaka M, Tamaki K, Katsumata Y (2001) Analysis of 168 short tandem repeat loci in the Japanese population, using a screening set for human genetic mapping. J Hum Genet 46(8):448–455

Nei M (1972) Genetic distance between populations. Amer Nat 106:283–291

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170

Nei M (1995) The origins of human populations: genetic, linguistic, and archeological data. In: Brenner S, Hanihara K (eds) The origin and past of modern humans as viewed from DNA. World Scientific, Singapore, pp 71–91

Omoto K, Saitou N (1997) Genetic origins of the Japanese: a partial support for the dual structure hypothesis. Am J Phys Anthropol 102:437–446

Oota H, Saitou N, Matsushita T, Ueda S (1999) Molecular genetic analysis of remains of a 2,000-year-old human population in China-and Its relevance for the origin of the modern Japanese population. Am J Hum Genet 64:250–258

Oota H, Kitano T, Jin F, Yuasa I, Wang L, Ueda S, Saitou N, Stoneking M (2002) Extreme mtDNA Homogeneity in Continental Asian Populations. Am J Phys Anthropol 118:146–153

Pérez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R, Betranpetit J (1997) Microsatellite variation and the differentiation of modern humans. Hum Genet 99:1–7

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ (2002) Y chromosomal DNA variation in Pakistan. Am J Hum Genet 70:1107–1124

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short term genetic distance. Genetics 105:767–779

Rogers JS (1972) Measures of genetic similarity and genetic distance. In: Wheeler MR (ed) Studies in genetics VII. University of Texas, Austin, TX, pp 145–153

Rosenberg NA, Pritchard JK, Weber JL, Cann HW, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Roubinet F, Despiau S, Calafell F, Jin F, Bertanpetit J, Saitou N, Blancher A (2004) Evolution of the O alleles of the human ABO blood group gene. Immunohematology 44:707–715

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:404–425

Saitou N, Omoto K, Du C, Du R (1994) Population genetic study in Hainan Island, China II Genetic affinity analyses. Anthropol Sci 102:129–147

Saitou N, Tokunaga K, Omoto K (1992) Genetic affinities of human populations. In: Roberts DL, Fujiki N, Torizuka K (eds) Society for the study of human biology symposium series 33: isolation and migration. Cambridge University Press, Cambridge, pp 118–129

Tajima A, Pan IH, Fucharoen G, Fucharoen S, Matsuo M, Tokunaga K, Juji T, Hayami M, Omoto K, Horai S (2002) Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. Hum Genet 110:80–88

Tajima A, Hayami M, Tokunaga K, Juji T, Matsuo M, Marzuki S, Omoto K, Horai S (2004) Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. J Hum Genet 49:187–93

Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144:389–399

Tanaka M, Takeyasu T, Fuku N, Li-Jun G, Kurata M (2004) Mitochondrial genome single nucleotide polymorphisms and their phenotypes in the Japanese. Ann NY Acad Sci 1011:7–20

Tokunaga K, Ohashi J, Bannai M, Juji T (2001) Genetic link between asians and native americans: evidence from HLA genes and haplotypes. Hum Immunol 62:1001–1008

Weir BS (1992) Independence of VNTR alleles defined as fixed bins. Genetics 130:873–887

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186