Hum Genet (2005) 117: 511–519
DOI 10.1007/s00439-005-1334-8

ORIGINAL INVESTIGATION

Jong-Jin Kim · Paul Verdu · Andrew J. Pakstis
William C. Speed · Judith R. Kidd · Kenneth K. Kidd

Use of autosomal loci for clustering individuals and populations of East Asian origin

Received: 13 September 2004 / Accepted: 14 April 2005 / Published online: 19 July 2005

Abstract We studied the genetic relationships among East Asian populations based on allele frequency differences to clarify the relative similarities of East Asian populations with a specific focus on the relationships among the Koreans, the Japanese, and the Chinese populations known to be genetically similar. The goal is to find markers appropriate for differentiating among the specific populations. In this study, no prior data existed for Koreans and the markers were selected to differentiate Chinese and Japanese. We typed, using AB TaqMan assays, single-nucleotide polymorphisms (SNPs) at 43 highly selected mostly independent diallelic sites, on 386 individuals from eight East Asian populations (Han Chinese from San Francisco, Han Chinese from Taiwan, Hakka, Koreans, Japanese, Ami, Atayal, and Cambodians) and one Siberian population (Yakut). We inferred group membership of individuals using a model-based clustering method implemented by the STRUCTURE program and population clustering by using computer programs DISTANCE, NEIGHBOR, LSSEARCH, and DRAWTREE, respectively, calculating genetic distances among populations, calculating neighbor-joining and least-squares trees, and drawing the calculated trees. On average 52% of individuals in the three Chinese groups were assigned into one cluster, and, respectively, 78 and 69% of Koreans and Japanese into a different cluster. Koreans differentiated from the Chinese groups and clustered with the Japanese in the principal component analysis (PCA) and in the best least-squares tree. The majority of Koreans were difficult to distinguish from the Japanese. This study shows that a relatively few highly selected markers can, within limits, differentiate between closely related populations.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00439-005-1334-8

J.-J. Kim
National Institute of Scientific Investigation,
DNA Analysis Division, Seoul, Korea

P. Verdu · A. J. Pakstis · W. C. Speed · J. R. Kidd
K. K. Kidd (✉)
Department of Genetics, Yale University School of Medicine,
333 Cedar Street, New Haven CT 06520, USA
E-mail: Kenneth.Kidd@yale.edu
Tel.: +1-203-7852654
Fax: +1-203-7856568

Introduction

The number of confirmed DNA polymorphisms detectable directly in the DNA and defined in various databases has increased from fewer than 200 in (HGM6 1981) to more than nine million (of which more than four million have been validated) in 2004 (dbSNP and build 121). The majority of these are single-nucleotide polymorphisms (SNPs). The SNPs are clearly the most plentiful genetic variants in the human genome, and a large number of them have high heterozygosities making them very useful DNA markers in researching genetic structure of populations and ethnic origins of individuals in a population (Frudakis et al. 2003; Rosenberg et al. 2003).

Determination of genetic relationships and genetic similarities among populations can be based on allele frequency similarities and differences of SNPs (Osier et al. 2002; Collins-Schramm et al. 2004; Fullerton et al. 2004; Kidd et al. 2004). Many different methods exist to analyze allele frequency data on populations and to represent the resulting relationships. Here we use many different analytic approaches on a highly selected dataset designed to quantify the relative similarities of East Asian populations with a specific focus on the relative similarities of Koreans, Japanese, and Chinese. Korea represents an important region for understanding population structure and origin of East Asians because of its location in Northeast Asia between China and Japan.

There are many arguments for the origin of East Asian populations (Yao et al. 2002). Major issues include the determination of the migration routes of ancestors for modern East Asians (Tajima et al. 2002; Karafet et al. 2001), and the nature of the genetic relationships among Chinese, Koreans, and Japanese (Kim et al. 2000; Rolf et al. 1998). Chu et al. (1998) and Su et al. (1999) examined some of the East Asian populations employing nuclear microsatellites and Y-chromosome haplotypes, respectively. They identified Northern and Southern clustering patterns among the populations they studied with indications of somewhat greater genetic variation in the Southern populations. This evidence supports the model of a northward migration of peoples from Southeast Asia. Kivisild et al. (2002) examined coding and control region variation from complete mtDNA sequence from East Asian samples and found the patterns generally consistent with the Y-chromosome analyses of Su et al. (1999) and supportive of the distinction between Northern and Southern populations, but they also note the complex regional specificities found in Northern groups such as the Koreans and Japanese that indicate other waves of migration that probably occurred in more recent millennia. Jin et al. (2003) studied 11 Y-chromosome markers in males from 11 ethnic groups and interpret their findings as supportive of a dual origin for the modern Korean population—genetic contributions from Northern Asian populations and an expansion of populations from the South.

Here, we explore whether significant clustering can be detected with 43 mostly independent autosomal SNPs, typed in eight East Asian populations and one Siberian population. By selecting markers that show larger than average allele frequency variation among East Asian populations this relatively small number of markers does identify a significant clustering pattern of individuals and populations.

Materials and methods

Samples

We analyzed 386 individuals from eight East Asian populations and one Siberian population. Descriptions of the populations and the specific samples are accessible online in ALFRED (Allele frequency database; http://alfred.med.yale.edu) under their UIDs: (Chinese from San Francisco (UID = SA000009J), Chinese from Taiwan (UID = SA000001B), Hakka from Taiwan (UID = SA000003I), Koreans (UID = SA000936S), Japanese (UID = SA000010B), Ami (UID = SA000002C), and Atayal (UID = SA000021D) from Taiwan, Cambodians(UID = SA000022E), and Siberian Yakut (UID = SA000011C). Sample sizes ranged from 25 individuals in the Cambodian sample to 54 in the Korean sample, with a mean of 43 individuals per sample. The DNA was purified by phenol/chloroform extraction from Epstein-Barr Virus-transformed cell lines as described earlier (Kidd et al. 2000).

SNP markers

We searched the SNP database of Applied Biosystems (http://myscience.appliedbiosystems.com) to find 32 SNP markers with a large difference in allele frequencies (min. $\Delta > 0.1$) between the Chinese and the Japanese—the two population frequencies available in that database. We typed these SNPs on all of our samples: eight East Asian populations and Yakut. We selected a subset of 21 independent markers with the largest differences of allele frequencies among the three Chinese groups and the Korean and Japanese populations. We also searched the ALFRED database to find SNP markers from among the several hundred SNPs typed that have both a large difference of allele frequencies and high F_{st} among the seven of our East Asian populations for which data already existed. In this way we chose an additional 22 SNPs at independent loci and then typed the Korean samples for these markers. Thus, the combined dataset for our analyses consists of data on 43 mostly independent, diallelic loci on individuals in nine populations. These loci are listed in Table 1 along with links to their definitions in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) and ALFRED.

SNP typing

The SNPs were typed by using AB TaqMan assays. All of the typings used 100 ng of genomic DNA and TaqMan probes, following the manufacturer's instruction, in 3-μl reactions. The reactions were analyzed and alleles called using an Applied Biosystems 7900HT Sequence Detection System.

Allele frequencies and statistics

Allele frequencies for the 43 SNPs were determined by gene counting assuming co-dominant inheritance and no silent alleles. For each SNP the value of F_{st} for the nine populations (Wright 1969) was calculated as $\sigma^2/(\bar{p}\bar{q})$ Hardy–Weinberg equilibrium was tested using the FENGEN program (Pakstis, unpublished). No significant departure from equilibrium was observed for any of the 43 markers in any of the nine populations under study. For markers less than 1 Mb apart pairwise linkage disequilibrium values were calculated as Δ^2 (Devlin and Risch 1995).

Cluster analysis

Clustering individuals

We used a model-based clustering method implemented by the program STRUCTURE (Pritchard et al. 2000;

Table 1 Polymorphisms studied and descriptive statistics for nine populations

Locus	TaqMan ID no.	dbSNP rs no.	Cytogenetic map location	Alleles		ALFRED site UID	Ref. allele frequency			Avg. het.	F_{st}[b]	
				Ref[a]			Min.	Max.	Avg. frq.		8p	9p
MSH4	C_2860732_10	rs11161731	1p31	C	T	SI000985W	0.250	0.512	0.371	0.45	0.028	0.033
FN1	C_71462_10	rs3817500	2q34-q36	A	G	SI001141H	0.378	0.815	0.683	0.40	0.077	0.080
CCR5	C_9698604_10	rs1800023	3p21	A	G	SI000999B	0.190	0.849	0.468	0.43	0.074	0.137
SEMA3F	C_15870200_10	rs2072053	3p21	C	T	SI001143J	0.281	0.611	0.451	0.46	0.066	0.061
D3S2465 region	C_8240638_10	rs901134	3p12.1	G	A	SI001164M	0.312	0.667	0.467	0.47	0.051	0.049
ADH7	C_1492617_1_	rs284784	4q23-q24	C	A	SI000878X	0.322	0.784	0.557	0.45	0.077	0.096
CRTL1	C_1142967_10	rs1345420	5q14.3	T	C	SI001144K	0.372	0.717	0.585	0.47	0.036	0.032
MAP3K7IP2	C_934526_10	rs521845	6q25.1-q25.3	G	T	SI001147N	0.300	0.543	0.416	0.47	0.026	0.029
TAS2R38	C_8876467_10	rs713598	7q34	C	G	SI000882S	0.225	0.549	0.373	0.45	0.036	0.048
D8S1024 region	C_1795241_10	rs11782184	8q24.3	G	A	SI001165N	0.256	0.630	0.463	0.48	0.046	0.045
ZFPM2	C_16041659_10	rs2622637	8q23	G	T	SI001148O	0.265	0.679	0.487	0.47	0.043	0.063
TNFSF15 region	C_609154_10	rs2006996	9q32	T	C	SI001166O	0.360	1.000	0.704	0.33	0.155	0.196
RET	C_12009293_10	rs2075914	10q11.2	G	A	SI001145L	0.676	0.952	0.786	0.32	0.043	0.040
D10S94	C_11657228_10	rs10899795	10q11.2	G	T	SI001146M	0.329	0.750	0.591	0.45	0.062	0.059
FADS2	C_2575514_10	rs174592	11q12-q13.1	A	G	SI001149P	0.000	0.717	0.358	0.35	0.238	0.230
CNOT2	C_356630_10	rs2255301	12q13-q14.1	T	A	SI001150H	0.451	0.857	0.733	0.37	0.018	0.065
D12S1635 region	C_470465_1_	rs10506294	12q13.12	T	C	SI001167P	0.274	0.740	0.546	0.46	0.067	0.079
CD4	C_11338582_10	rs11611635	12pter-p12	T	C	SI001069Q	0.250	0.550	0.422	0.47	0.036	0.033
TUBGCP3	C_16085480_10	rs2182268	13q34	G	A	SI001152J	0.500	0.820	0.658	0.43	0.044	0.051
DOCK9	C_11421850_1_	rs1927568	13q32.3	T	C	SI001168Q	0.232	0.830	0.541	0.44	0.131	0.121
FARP1	C_1854890_10	rs3742141	13q32.2-q32.3	C	T	SI001153K	0.262	0.641	0.431	0.47	0.045	0.046
CRIP1	C_51568_10	rs8003942	14q32.33	A	G	SI001154L	0.090	0.660	0.453	0.41	0.166	0.166
RPS6KA5	C_595610_10	rs727258	14q31-q32.1	T	C	SI001155M	0.380	0.663	0.559	0.47	0.041	0.038
PSTPIP1	C_217833_10	n/a	15q24-q25.1	C	A	SI001156N	0.286	0.732	0.509	0.46	0.056	0.075
CBLN1	C_8915676_10	rs893174	16q12.1	G	C	SI001162K	0.312	0.640	0.446	0.47	0.040	0.045
CDC6	C_1123657_1_	rs13706	17q21.3	G	A	SI001158P	0.536	0.722	0.622	0.46	0.017	0.015
RND2	C_3178698_1_	rs2298862	17q21	T	C	SI000952Q	0.440	0.804	0.653	0.43	0.042	0.050
MAPT	C_1016016_1_	rs242557	17q21	A	G	SI001151I	0.362	0.681	0.552	0.48	0.034	0.039
HOXB13	C_2905935_10	rs3110607	17q21	T	C	SI000915P	0.395	0.840	0.689	0.39	0.087	0.093
HOXB13	C_7454215_10	rs890435	17q21	C	T	SI000917R	0.547	0.786	0.640	0.45	0.027	0.024
HOXB2	C_11619715_10	rs1042815	17q21	G	A	SI000921M	0.262	0.500	0.381	0.46	0.033	0.030
SCAP1	C_1570377_10	n/a	17q21.3	C	A	SI001096Q	0.510	0.894	0.783	0.31	0.040	0.086
PHB	C_11620508_1_	rs2233659	17q21	T	G	SI000976W	0.389	0.784	0.628	0.44	0.048	0.056
SYNGR2	C_3068817_10	rs7208422	17qter	A	T	SI001112F	0.409	0.800	0.572	0.47	0.047	0.042
BAIAP2	C_150018_10	rs8079626	17q25	A	G	SI001159Q	0.155	0.532	0.332	0.41	0.069	0.074
CD7	C_11600340_10	rs4789763	17q25.2-q25.3	A	G	SI001157O	0.200	0.568	0.432	0.47	0.053	0.048
TBCD	C_1674429_10	rs733342	17q25.32	A	G	SI000973T	0.000	0.354	0.208	0.31	0.073	0.066
ACAA2	C_2714437_10	rs521861	18q21.1	C	G	SI001163L	0.433	0.643	0.557	0.49	0.019	0.017
RSHL1	C_7830137_10	n/a	19q13.3	C	T	SI000981S	0.560	0.929	0.745	0.35	0.062	0.081
DM1	C_11712219_1_	rs672348	19q13.3	T	G	SI000982T	0.167	0.460	0.315	0.41	0.030	0.040
D20S104 region	C_1274218_10	rs12480506	20p12.1	A	G	SI001169R	0.491	0.776	0.634	0.45	0.029	0.027
TST	C_2478896_10	rs135851	22q13.1	A	G	SI001161J	0.256	0.718	0.495	0.45	0.099	0.094
D22S1170 region	C_2785413_1_	rs738745	22q13.31	A	G	SI001170J	0.478	0.895	0.655	0.42	0.089	0.079

[a] The reference allele, because it is the AB TaqMan allele associated with the FAM dye and is usually designated the X allele by the protocol

[b] Two F_{st} values are shown: *8p* was calculated for the eight East Asian population samples; *9p* was calculated for nine population samples (eight East Asian samples and the Yakut)

Falush et al. 2003) to infer relationships among East Asian populations by assigning individuals to clusters. Previous applications of STRUCTURE were very successful in studies of humans (Rosenberg et al. 2002) and dogs (Parker et al. 2004). We ran STRUCTURE using a model with admixture, separate α for each population, and correlated allele frequencies.

Clustering populations

We used the computer program DISTANCE to calculate the pairwise τ (tau) genetic distances (Cavalli-Sforza and Edwards 1967; Kidd and Cavalli-Sforza 1974; Cavalli-Sforza et al. 1994). Principal components analysis (PCA) was based on the matrix of pairwise tau genetic distances. The relationships among populations were also represented by tree diagrams based upon those pairwise genetic distances. A neighbor-joining tree was computed using the NEIGHBOR program and drawn with the help of the DRAWTREE program; both of these programs are part of the PHYLIP software package (Felsenstein 1989, 1993). The LSSEARCH program (Kidd and Sgaramella-Zonta 1971) was used to calculate an exact least-squares solution for the population tree. Twenty-eight different trees were examined using a heuristic search to generate similar trees. The best tree from the least-squares method has the smallest length and no negative internal segments. For the bootstrap analysis, the PHYLIP SEQBOOT program was applied to generate 1,000 replicate data sets that were then used as input for the GENDIST program in order to compute Reynolds distance matrices. The CONSENSE program summarized the 1,000 neighbor-joining trees.

Results

Table 1 identifies the 43 polymorphisms studied, specifies their chromosomal locations, and supplies the average heterozygosity plus the value range and average allele frequencies for the nine populations. The F_{st} values calculated both for the eight East Asian populations and for the nine populations also appear in Table 1. Sample sizes and allele frequencies for all markers and populations can be found in ALFRED under the UIDs in Table 1. Frequencies of the reference allele in each of the population samples can also be found in Electronic Supplementary Material (ESM) Table 1 for each polymorphism. Expected heterozygosities in each population are given in ESM Table 2 for the 43 markers. ESM Table 3 shows how similar/different each of the 36 unique population pairs are for the 43 markers based on *t*-tests.

Linkage disequilibrium statistics for all markers less than one megabase apart are given in ESM Table 4. Of the 14 such pairs of markers only three marker pairs had consistent significant LD in most populations and the LD was complete ($\Delta^2 = 1$) for only one marker pair in one population. Thus, except for that single instance each marker is contributing some unique information on population relationships and most markers are completely independent.

Marker ascertainment

The method used to select SNPs has created a data set that is clearly biased relative to unselected markers. There is, for example, increased heterozygosity in our data set relative to random SNPs. For our 43 selected markers, 95% of the total number of polymorphisms has an average heterozygosity higher than 30%, among the eight East Asian populations (not including Koreans). If we compare these results with a set of 454 polymorphisms more randomly selected, excluding our 43 marker data set, we find that only 59% have an average heterozygosity higher than 30%, among the same eight East Asian populations.

The average F_{st} value among all 43 markers based on the eight East Asian populations is 0.060 and the median is 0.046 (F_{st} range: 0.017–0.238). Given the special selection procedure, it is not surprising that the mean and median F_{st} are both elevated as compared with a larger unselected set of 370 independent diallelic markers based on seven populations (same East Asian populations as previously introduced excluding the Koreans) and excluding the 43 sites studied here: average $F_{st} = 0.033$ and median = 0.026. Seven of the 43-SNP data set markers (*CCR5*, *ADH7*, *TNFSF15*, *FADS2*, *CRIP1*, *DOCK9*, and *TST*) showed very high F_{st} values among the eight East Asian populations. The allele frequency profiles for these polymorphisms are shown in Fig. 1 for the nine populations.

All of these SNPs represent old global polymorphisms that have high average heterozygosities in East Asian populations. The average heterozygosities in other regions of the world are lower than in East Asia, as expected given the ascertainment bias, but none is fixed in any geographic region, such as Africa or Europe (ESM Fig. 1), although a few are occasionally fixed in a single population. Thus, these are all old SNPs that arose in Africa prior to the expansion of modern humans out of Africa. The simplest explanation for this pattern is that random genetic drift caused the increase in heterozygosity at these SNPs in East Asia and those were the ones ascertained.

Clustering individuals

To examine the genetic structure of the individuals in our dataset, we first applied a model-based clustering method that groups individuals into a specified number (*K*) of clusters, each one of them characterized by a unique set of allele frequencies. In the computer program STRUCTURE that implements this algorithm, *K* is chosen in advance and can vary from run to run of the program. Each individual's genotype can have a proportion of membership in each one of the *K* clusters, summing to one across the *K* clusters. These proportions can be considered as proportions of membership in the different clusters or as relative probabilities of ancestry of the individual derived from the hypothesized clusters. By assigning our sampled individuals to predefined populations, STRUCTURE also prints out the average membership proportions for these predefined populations.

Using our 43-SNP dataset among 339 individuals from eight predefined populations, SF Chinese, TW Chinese, Hakka, Koreans, Japanese, Ami, Atayal, and the Cambodians, the best result using STRUCTURE was obtained assuming four clusters ($K = 4$). Among all

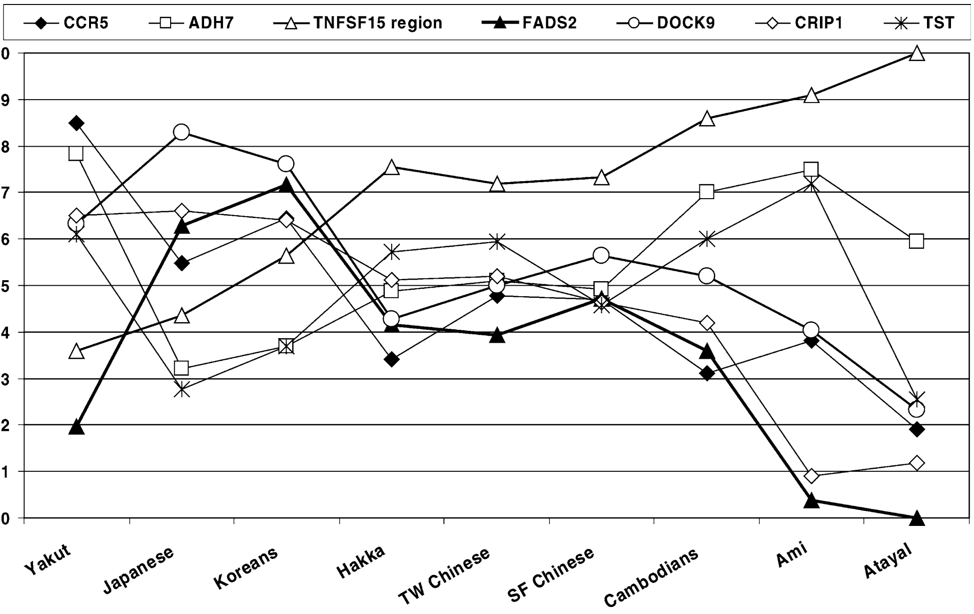


Fig. 1 Allele frequencies for the seven SNPs (from our 43-SNP dataset) with the highest F_{st} values in nine East Asian populations

eight populations, individuals coming from the same predefined population show, most of the time, a similar pattern of membership proportions among the K clusters (Fig. 2a). The Dirichlet parameters (αi) obtained for each cluster [max. ($\alpha 1$, $\alpha 2$, $\alpha 3$, $\alpha 4$) < 0.26, mean (αi) = 0.17] indicate that each individual of the data set has been mainly assigned to one of the K clusters (Pritchard et al. 2000). Among the eight East Asian populations, at $K = 4$ clusters, the variance among individuals of the same cluster is low for all the four groups, and especially low for the first cluster (Fig. 2a). All three Chinese populations present a very similar pattern: individuals in these populations are assigned mainly to the first two clusters for an average of, respectively, 48 and 29% of each individual's genotype. On average, individual Cambodians are assigned 43% to the first cluster and 23% to the fourth one. The second cluster is mainly composed of individuals in two predefined populations: the Japanese and the Koreans, with individuals assigned on average 84 and 76% to this cluster, respectively. The Ami mainly define the third cluster, with 47% of their individuals' genotypes assigned to this cluster. Finally, the fourth cluster is defined primarily by the Atayal individuals; on average, the individuals in this population are assigned 75% to this cluster. Thus, we did identify structure among the East Asian groups ($K > 1$), but we are still unable to differentiate between Japanese and Koreans.

If one considers the allocated percentages for an individual to be proportional to ancestry from each of the hypothesized clusters, it might be possible to include additional populations that would solidify existence of an additional cluster and alter the optimal clustering of the existing individuals. Because we were interested in relationships of Koreans to other populations, it seemed reasonable to incorporate an additional population from Northern Asia into the analysis. Including the Yakut increased the overall sample to 386 individuals and substantially modified clustering: the best result now occurred for $K = 5$ clusters. Among all nine populations, individuals coming from the same predefined population generally show a similar pattern of proportional membership among the five clusters (Fig. 2b). With $K = 5$, the Dirichlet parameters (α i) for each cluster [max. ($\alpha 1$, $\alpha 2$, $\alpha 3$, $\alpha 4$, and $\alpha 5$) < 0.20, mean (αi) = 0.11] indicate that a better job has been done by STRUCTURE to assign each individual mainly to one of the five clusters, than for the previous run with eight populations and four clusters. This analysis also showed low variances among individuals belonging to each one of the five clusters, especially low in the first cluster (Fig. 2b).

In Fig. 3, the three Chinese groups cluster with genotypes of individuals assigned 52%, on average, to the first cluster. Genotypes of the Atayal are assigned 77% to the second cluster, primarily defining this cluster. The Ami genotypes are assigned 24% to the second cluster, 20% to the first cluster but 42% to the fifth cluster, primarily defining this cluster. Genotypes of individuals in the three Chinese groups and the Japanese and Korean groups only share an average of 22% for cluster 3. However, genotypes of individuals in the Korean and Japanese populations are assigned to this third cluster with respective membership proportions of 78 and 69%, on average. The Cambodians present an intermediate pattern among the East Asian populations: 47% of their individuals' genotypes are assigned to the first cluster, 21% to the second, 16% to the third, and 12% to in the fifth cluster. The Yakut essentially define the fourth cluster with genotypes assigned to this cluster an average of 58% with 27% of each individual's genotype assigned to cluster 3, the predominant cluster for Japanese and Koreans.

Even though we improved the clustering by including the Yakut, especially for the Taiwanese populations, we

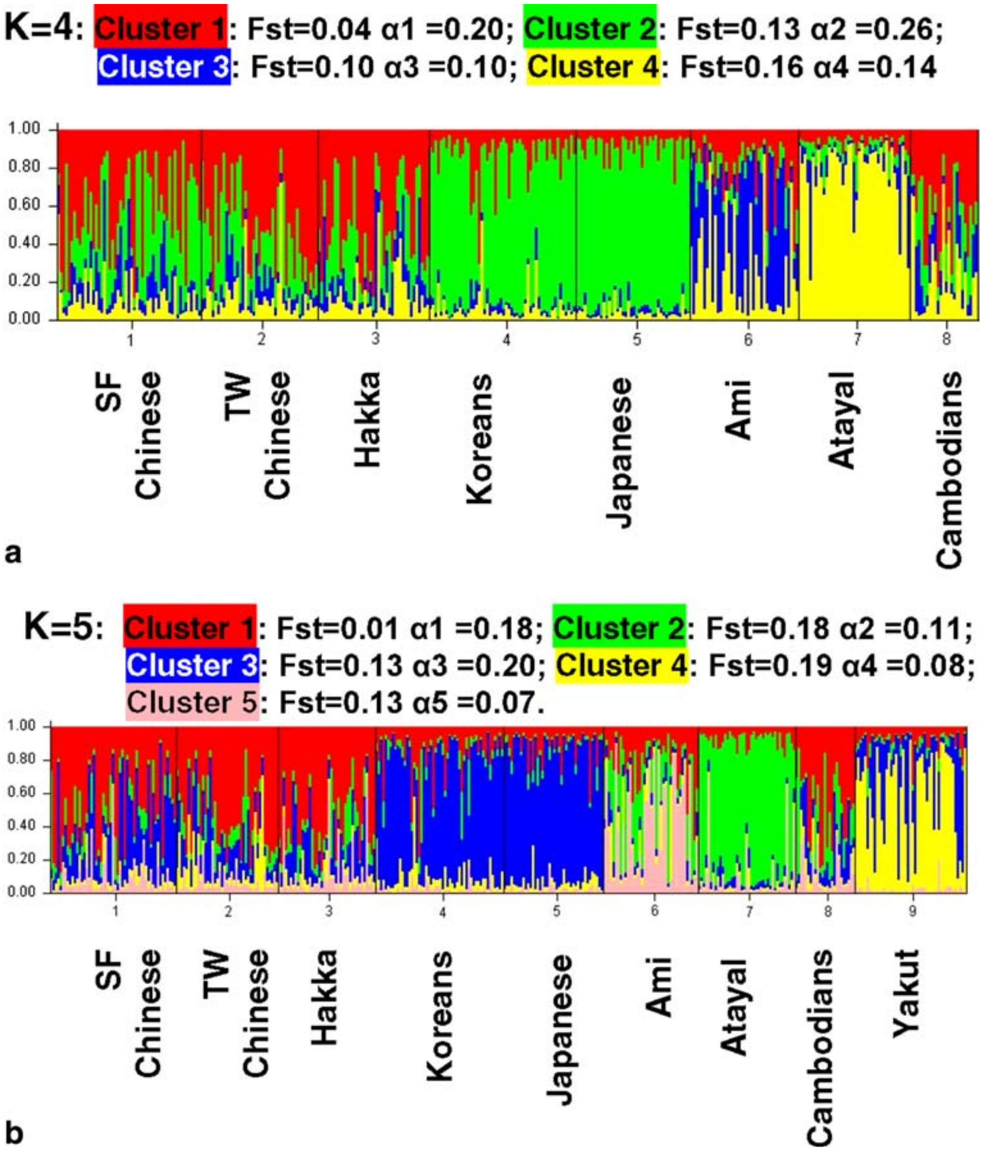


Fig. 2a, b Estimated membership proportions in each of the *K* assumed populations. Each individual is plotted in a *single vertical line*, separated in *K* colored segments representing the proportion of membership in each one of the *K* clusters. *Black lines* separate individuals from two different predefined populations. **a** Forty-three independent diallelic loci, typed in eight East Asian populations; $K=4$ clusters assumed. **b** Forty-three independent diallelic loci, typed in nine East Asian populations; $K=5$ clusters assumed

are still not able to distinguish well between Koreans and Japanese. However, we observe four patterns for East Asia: a Southern (shown by the three Chinese populations, and the Cambodians); a Northern pattern (shown by the Japanese and the Koreans); a Siberian pattern (shown by the Yakut population), and distinct patterns for the two aboriginal Taiwanese populations. We note an increasing average assignment of individuals in the second cluster, going from China (10% on average among individuals from the three populations), south to Cambodia (on average 21%) and Northeast to Taiwan (25% on average for Ami and 77% on average for Atayal individuals) (Fig. 3).

Clustering populations

A PCA based on a tau genetic-distance matrix that includes all nine populations (Fig. 4) shows a clustering pattern consistent with the results shown by the STRUCTURE program. The first PC accounts for more than 81% of the total variation among all nine populations and maximally separates the Atayal from the Japanese. The second PC accounts for less than 14% of the total variation and maximally separates the Yakut and Japanese. The three Chinese populations and the Cambodians are clustered together in the PCA. The first PC separates these four populations from a group composed of the Japanese and Korean populations. We also observe that the distance between the Japanese–Korean grouping and the Yakut is along the axis defined by the second PC with only a small percentage of the total genetic variation accounted for by this component.

These results reflect the same geographical pattern for continental East Asia as the results from STRUCTURE. We also can identify here a Northern and a Southern pattern separated by large genetic distances between Japanese–Koreans and the three Chinese populations.

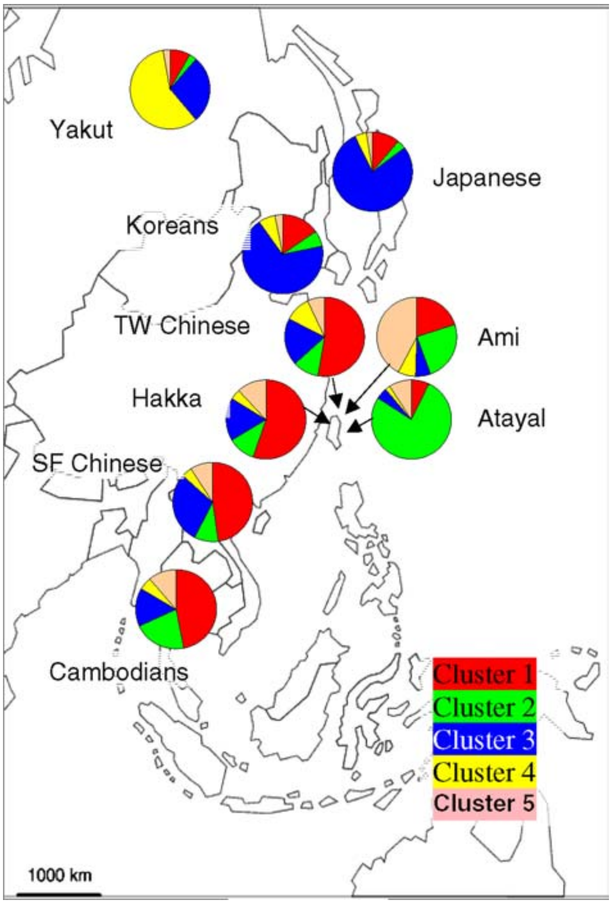


Fig. 3 Average estimated membership proportions among each one of the nine predefined populations plotted in pie charts, for $K = 5$ clusters assumed

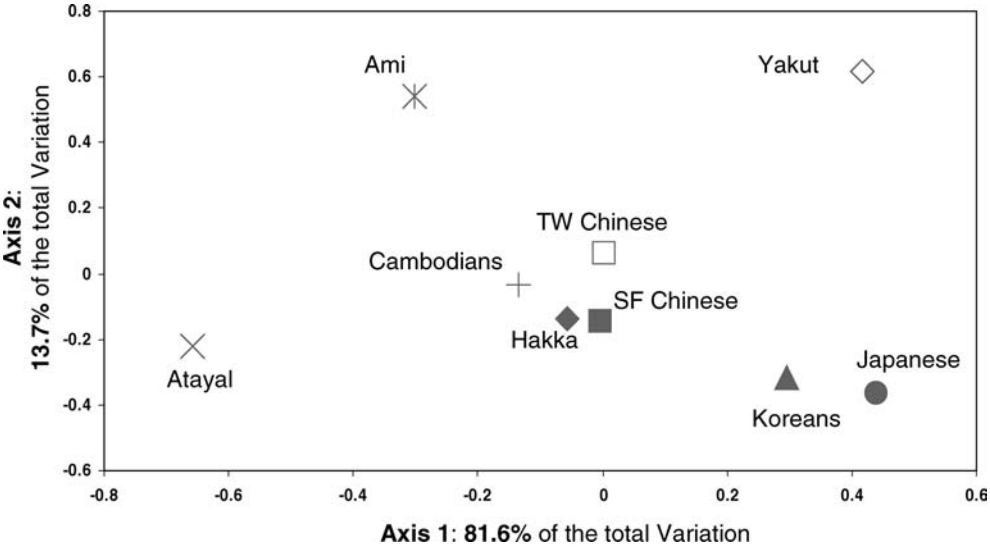
The Southern pattern is consistent with the Taiwanese aboriginal populations (Ami and Atayal) originating from an expansion out of the Southeast Asia into the Pacific (Chu et al. 1998). It is interesting to note that we are still not able to distinguish clearly between Japanese and Koreans in the PCA. However, the Koreans do tend to be more intermediate between the Japanese and the three Chinese populations using this method. (In ESM Table 3, the Koreans are also intermediate relative to the Chinese and Japanese groups in how frequently they are significantly different from one another pairwise across the 43 SNPs.)

We evaluated by exact least-squares analysis 28 of the more than 135,000 different tree structures possible for the nine East Asian populations. The best tree found (Fig. 5) is consistent with the results shown by the PCA and STRUCTURE analyses. The Japanese and Korean populations group together apart from the Chinese and Taiwanese groups, and the three Chinese groups are very close along the same segment. Notice that the greatest distance between two populations along the tree is between the Yakut and the Atayal. The SF Chinese and the TW Chinese populations are very similar, while the PCA differentiated them somewhat more. Except for the branch between the SF and TW Chinese populations, this configuration is strongly supported by very high bootstrap values based on 1,000 replications. The bootstrap value supporting the separation between the Northern cluster (Japanese, Koreans, and Yakut), and the Southern is 99.2%. For randomly selected markers the branch lengths would be proportional to time—in generations divided by twice the effective population size (Kidd and Cavalli-Sforza 1974)—for the correct structure. That cannot be the case for this tree, however, because of the strong bias in selecting markers that were known to have a large allele frequency difference between the Chinese and the Japanese populations.

Discussion

We found that with a small, carefully selected set of SNPs we can identify genetic substructure among our

Fig. 4 The PCA plot based on a tau genetic-distance matrix for 43 independent diallelic markers typed in nine East Asian populations. This map presents the first two axes, accounting for more than 95% of the total genetic variation



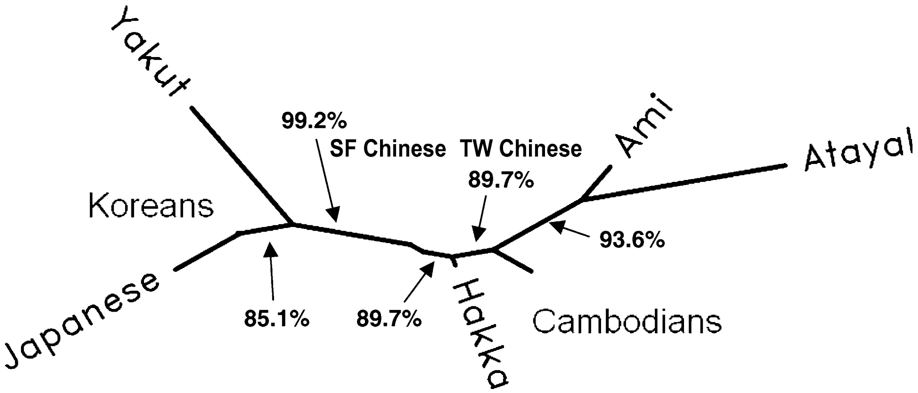


Fig. 5 Least-squares tree for nine East Asian populations and 43 independent diallelic markers, based on a tau genetic-distance matrix. This best tree among those examined was the one with the shortest length and no negative internal segments. The SF Chinese and the TW Chinese cannot be distinguished along the central segment of the tree. Bootstrap values are based on 1,000 replications

existing set of East Asian populations, consistent across three clustering methods: STRUCTURE, PCA, and a least-squares tree search. The northern/southern pattern in East Asia had already been observed using Y-chromosome haplotypes and autosomal microsatellite information (Chu et al. 1998; Su et al. 1999). We now have strong statistical support for the existence of this pattern using autosomal SNPs. Though the markers were selected to emphasize the difference between Chinese and Japanese (see ESM Table 3 for statistical confirmation), the clusterings of the additional populations confirm the North-South pattern. High F_{st} among East Asian populations could be another plausible criterion for selecting a marker set. We empirically observed that this criterion was doing a worse job differentiating East Asian populations than the large allele frequency differences criterion presented here (data not shown).

The SF Chinese population shows a STRUCTURE clustering pattern slightly closer to the Northern populations than to the rest of the Chinese groups. While probably not statistically significant, it may be due to a handful of individuals that, unlike the rest of the sample, originate in the Northern part of China.

It is difficult to choose the appropriate number of clusters (K) for modeling the data with the STRUCTURE program. We presented here the "best K", considering the results of a series of independent runs of STRUCTURE for different values of K, and the other external information we had concerning our samples. Now that we have established that a small set of independent SNPs could significantly differentiate among some very similar populations, it would be interesting to extend our study to other populations (Africans, Europeans, and Amerindians), in order to estimate relative contribution from these populations to the East Asian genetic pattern and reconstruct the history of the peopling of East Asia. This set of markers, with readily available TaqMan assays, is an excellent set for others to study on additional East Asian populations.

This study also demonstrates that one can affect the outcome of population studies by selecting markers that show a specific pattern of allele frequencies. Thus, it is not a surprise that Chinese and Japanese are distinguished by these analyses since the markers were explicitly chosen to show that difference and we did verify the differences by replication in independent samples. The highly non-random markers studied give a different pattern of relationships than a random set of loci. For the eight populations for which prior data exist (this excludes Koreans) we can evaluate the relationships based on the 80 loci with over 600 independent alleles used for the 37 population tree shown in Tishkoff and Kidd (2004). The PCA of these data (ESM Fig. 2) places the Japanese centrally with the two Chinese population samples and the Hakka. The Yakut, Atayal, and Cambodians are the outliers. A least-squares analysis (not shown) likewise has different structure and bootstrap values than the tree in Fig. 5.

Though extensive SNP data on Koreans are not available, F_{st} analysis of 370 SNPs on the other eight populations in this study shows few loci with large frequency differences. Thus, we would expect a STRUCTURE analysis to give little clear discrimination among these populations unless the more variable loci are included.

What is especially relevant is how the other populations, especially the Koreans, relate to the populations used to select the markers, Chinese and Japanese. Unfortunately, we are still unable to distinguish clearly between Japanese and Koreans. The results in ESM Table 3 confirm this, showing that only three of the 43 gene-frequency differences are statistically different ($P \leq 0.010$) when comparing the Japanese and the Koreans. Common ancestry and/or extensive gene flow between these two populations throughout history seem likely and make it very hard to find population-specific alleles that could help differentiate them. By increasing the number of markers used, we might improve the accuracy of the estimates of proportions assigned to clusters (Pritchard et al. 2000), and therefore increase the quality of our clustering with STRUCTURE. Increasing the number of populations studied might also improve our clustering. As larger numbers of SNPs are studied on both Koreans and Japanese it should be possible to find markers that will help differentiate between them.

Acknowledgements We thank the participants who donated DNA samples used in this study. This research was supported by grants from the National Institute of Health (AA09379 and GM 57672) and the Post-doctoral Fellowship Program of the Korean Science and Engineering Foundation.

References

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

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

Collins-Schramm HE, Chima B, Morii T, Wah K, Figueroa Y, Criswell LA, Hanson RL, Knowler WC, Silva G, Belmont JW, Seldin MF (2004) Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet 114:263–271

Devlin B, Risch NA (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure: extensions to linked loci and correlated allele frequencies. Genetics 164:1567–1587

Felsenstein J (1989) PHYLIP—phylogeny inference package, Version 3.2. Cladistics 5:164–166

Felsenstein J (1993) PHYLIP—phylogeny inference package, version 3.5p. (Distributed by the author, who is at the Department of Genetics, University of Washington, Seattle)

Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK (2003) Classifier for the SNP-based inference of ancestry. J Forensic Sci 48:771–782

Fullerton SM, Buchanan AV, Sonpar VA, Taylor SL, Smith JD, Carlson CS, Salomaa V, Stengard JH, Boerwinkle E, Clark AG, Nickerson DA, Weiss KM (2004) The effects of scale: variation in the *APOA1/C3/A4/A5* gene cluster. Hum Genet 115:36–56

HGM6 (Human Gene Mapping 6, Oslo Conference, 1981) 6th International Workshop on Human Gene Mapping, Cytogenetics and Cell Genetics, vol 32, 1982

Jin HJ, Kwak KD, Hammer MF, Nakahori Y, Shinka T, Lee JW, 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

Karafet T, Xu L, Du R, Wang W, Feng S, Wells RS, Redd AJ, Zegura SL, Hammer MF (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet 69:615–628

Kidd KK, Cavalli-Sforza LL (1974) The role of genetic drift in the differentiation of Icelandic and Norwegian cattle. Evolution 28:381–395

Kidd KK, Sgaramella-Zonta LA (1971) Phylogenetic analysis: concepts and methods. Am J Hum Genet 23:235–252

Kidd JR, Pakstis AJ, Zhao H, Lu RB, Okonofua FE, Odunsi A, Grigorenko E, Tamir BB, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH*, in a global representation of populations. Am J Hum Genet 66:1882–1899

Kidd KK, Pakstis AJ, Speed WC, Kidd JR (2004) Understanding human DNA sequence variation. J Hered 95:406–420

Kim W, Shin DJ, Harihara S, Kim YJ (2000) Y chromosomal DNA variation in east Asian populations and its potential for inferring the peopling of Korea. J Hum Genet 45:76–83

Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R (2002) The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol 19:1737–1751

Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK (2002) ALFRED: an allele frequency database for anthropology. Am J Phys Anthro 119:77–83

Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD, Malek TB, Johnson GS, DeFrance HB, Ostrander EA, Kruglyak L (2004) Genetic structure of the purebred dog. Science 304:1160–1164

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

Rolf B, Horst B, Eigel A, Sagansermsri T, Brinkmann B, Horst J (1998) Microsatellite profiles reveal an unexpected genetic relationship between Asian populations. Hum Genet 102:647–652

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

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet 65:1718–1724

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

Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. Nat Genet Suppl 36:21–27

Wright S (1969) Evolution and the genetics of populations. The theory of gene frequencies, vol 2. University of Chicago Press, Chicago, p 511

Yao YG, Kong QP, Bandelt HJ, Kivisild T, ZhangYP (2002) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet 70:635–651