N. Saitou
Laboratory of Evolutionary Genetics, National Institute of Genetics, Mishima411, Japan

A Genetic Affinity Analysis of Human Populations

Genetic affinity of human populations based on allele frequency data was studied from two viewpoints. (1) The effect of the number of polymorphic loci on the reconstruction of a phylogenetic tree of human populations was empirically investigated. Genetic affinity trees were constructed based on data for 1 - 12 polymorphic loci, by using the neighbor-joining method. Geographical clustering of populations gradually appeared when the number of loci was increased. A new classification and terminology of higher order human population clusters is proposed based on these and other studies. (2) A new method of estimating the absolute divergence time of two populations is proposed, which is based on a diffusion equation that describes random genetic drift.

Key words: genetic polymorphism, allele frequency, divergence time, population genetics

Introduction

Evolution is based on the change of genes in gene phylogenies through DNA replication, or "descent with modification" as described by Darwin (1859). Therefore it is obvious that evolution of any organism, including humans, should be studied at the nucleotide or molecular level.

With the advent of molecular techniques, genetic variations in human populations are now extensively studied at the DNA level. In particular, mitochondrial DNA has been the main locus of interest, and extensive DNA polymorphism data have made it possible to construct phylogenetic trees of many alleles (e.g., Cann et al. 1987; Horai et al. 1993). This situation was unthinkable at the time when a two-allele polymorphism was typical. Figure 1 is a frequency distribution of the region V 9-bp deletion allele of mitochondrial DNA among human populations in Asia, Oceania, and America. Under a classical view of genetic polymorphism, this distribution merely suggests some genetic affinity among populations in which the 9-bp deletion allele has been found. When we know the genealogy of the mitochondrial DNA, a more detailed history of this deletion allele appears. For example, Wrischnik et al. (1987) showed that mitochondrial types having this 9-bp deletion were monophyletic, while Torroni et al. (1993a) found the same 9-bp deletion in Yukagir in Siberia on the other mitochondrial lineage.

It should be remembered, however, that data on a single locus are usually not sufficient for delineating the genetic relationship of closely related populations. Even if a population tree (phylogenetic relationship of populations) is estimated based on the detailed information of one locus such as mitochondrial DNA, that tree may be considerably different from one estimated by using data of another locus. This is because the inter-locus stochastic variance produced by random genetic drift is much larger than the intra-locus sampling variance (Nei 1987). Therefore, data for many polymorphic loci are necessary to estimate a population tree. This is theoretically well known and some computer simulations have also been conducted (Nei et al. 1983; Nei and Takezaki 1994). However, the effect of an increase in polymorphic loci on the construction of a population tree has not been properly studied using real allele frequency data.

One reason for difficulty in such empirical studies is that we usually do not know the true tree. This difficulty can, however, be avoided if one accepts the following arguments. Random genetic drift is the main cause of population differentiation, and geographical isolation is the most effective means of attaining the independent changes of allele frequency among different populations. Large-bodied land mammals, including humans, often have great genetic and phenotypic differences among populations or related species living in different continents. Hence, we should expect a human population to have closer genetic affinity with geographically neighboring populations than with populations far away. Therefore, the first subject of the present study is an empirical analysis of the effect of the number of polymorphic loci on the construction of population trees, on the assumption that populations of geographical propinquity are clustered in a reliable population tree.

This first subject is mainly concerned with the branching pattern, or topology, of population trees. At best, only the relative lengths of the branches come into question. In evolutionary studies, however, estimation of the absolute branch lengths, i.e. divergence times, is also quite important. Nei and Roychoudhury (1974, 1982) estimated the divergence times between the three major groups of Man: Negroid, Caucasoid, and Mongoloid. Those divergence time estimates were based on allele frequency data for more than 30 protein loci, including many monomorphic

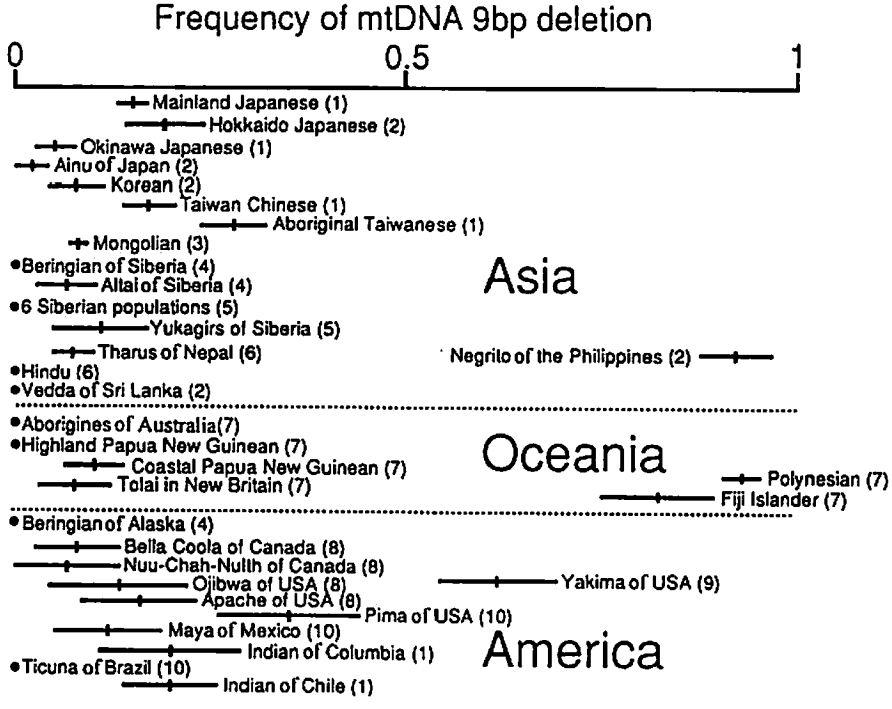

Figure 1 - Allele frequency distribution of mtDNA 9-bp deletion among various human populations in Asia, Oceania, and America. Vertical lines and bars are the mean and its standard errors, respectively (standard error were computed by assuming a binomial variance $p[1-p]/n$, where p is allele frequency and n is sample size). An allele frequency of zero is indicated by a full circle. Numbers in parentheses are references as follows: 1 = Horai et al. (1993) and references therein, 2 = Harihara et al. (1992), 3 = Sambuughin et al. (1991), 4 = Shields et al. (1992), 5 = Torroni et al. (1993a), 6 = Passarino et al. (1993), 7 = Herzberg et al. (1989), 8 = Torroni et al. (1993b), 9 = Shields et al. (1993), 10 = Schurr et al. (1990).

ones. When more closely related populations are concerned, however, the number of loci commonly studied among all the compared populations is limited, and the limited number of loci found to be polymorphic in one population tend to be examined in other populations. Thus, the second subject of this paper is the presentation of a new method for estimating the divergence time between two populations, when only polymorphic loci are examined.

Effect of the number of loci on the construction of human population trees

We have used allele frequency data, compiled by Roychoudhury and Nei (1988), of 50 humans populations, mainly from the circum-Pacific region (Table 1). Forty-eight of the populations are from Oceania, South and North America, and Asia, and Figure 2 shows their approximate geographical locations. The remaining two populations are the English from the United Kingdom (population ID 49) and the Yoruba from Nigeria (population ID 50). The same set of population IDs (1-50) are used in Table 1 and Figures 2 - 7.

Figure 3 is the frequency distribution of the M allele of the MN blood group locus among the 50 human populations. Even these data on a single locus polymorphism show a rough clustering of geographically related populations. All the populations in Oceania (IDs 1-16) have frequencies smaller than 0.45 except population 16 (Samoan), whereas native North and South American populations (IDs 17-32) tend to have high frequencies (above 0.5). Asian populations (IDs 33-48) show intermediate frequencies, ranging 0.4 - 0.7. This rather good clustering of populations, however, does not apply to the African population (ID 50). Although it now seems to be established that the African populations are genetically far removed from non-African populations (Nei and Roychoudhury 1982, 1993; Cavalli-Sforza et al. 1988), this dichotomy does not

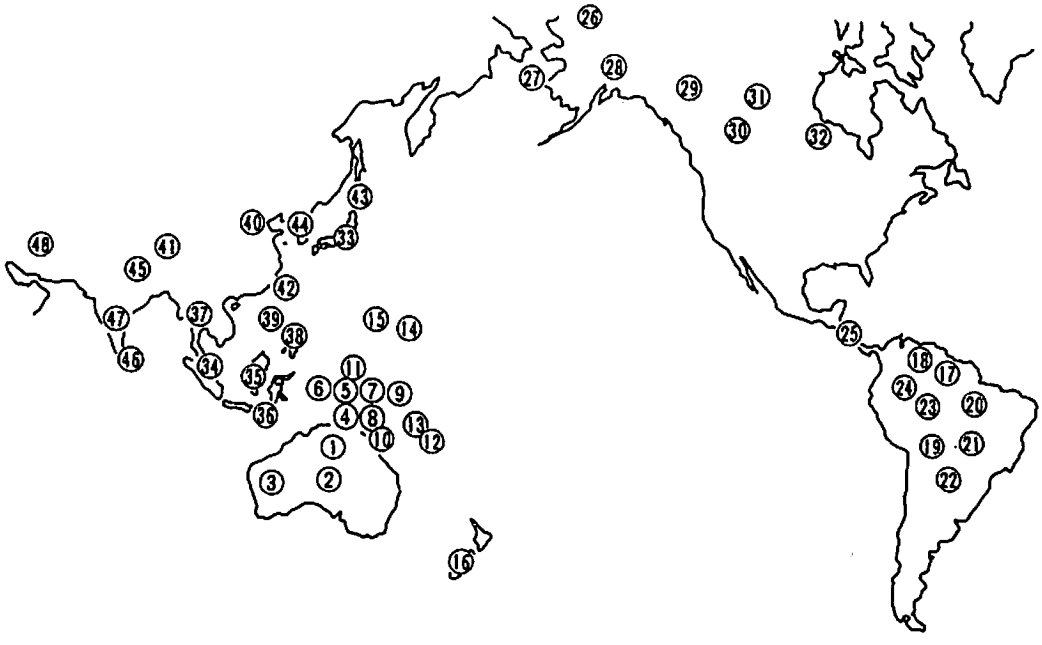

Figure 2 - Approximate geographical locations of 48 human populations compared in the present study. Population ID numbers correspond to those of Table 1.

TABLE 1 - The fifty human populations compared in the present study

Oceania:
1) Australian Aborigines, Northern Territory
2) Australian Aborigines, Central Australia
3) Australian Aborigines, Western Australia
4) Papua New Guinean, South Coastal Plain
5) Papua New Guinean, North Central Highland
6) Papua New Guinean, West Highland
7) Papua New Guinean, East Highland
8) Papua New Guinean, Central Highland
9) Papua New Guinean, Bay Province
10) Melanesian, Karlar Island
11) Melanesian, Manus Island
12) Melanesian, Buka Island
13) Melanesian, Solomon Island
14) Micronesian, East Caroline Island
15) Micronesian, Marshall Island
16) Polynesian, Samoa Island (now living in New Zealand)
South and Middle America:
17) Yanomama
18) Makiritare
19) Aymara
20) Baniwa
21) Cayapo
22) Macushi
23) Wapishana
24) Ticuna
25) Guaymi
North America:
26) Eskimos, North Alaska
27) Eskimos, St. Lawrence Island
28) Athabaskan Indian
29) Eskimos, Canada
30) Ojibwa Indian
31) Cree Indian
32) Dogrib Indian
Asia
33) Japanese
34) Malays
35) Batak
36) Balinese
37) Thais
38) Filipino
39) Negritos
40) Han, Northern China
41) Tibetan
42) Taiwan Aborigines, Toroko
43) Ainu
44) Korean
45) Nepali
46) Shinhalese (Sri Lanka)
47) South Indian (India)
48) Iranian (Iran)
Other geographical areas:
49) English (United Kingdom)
50) Yoruba (Nigeria)

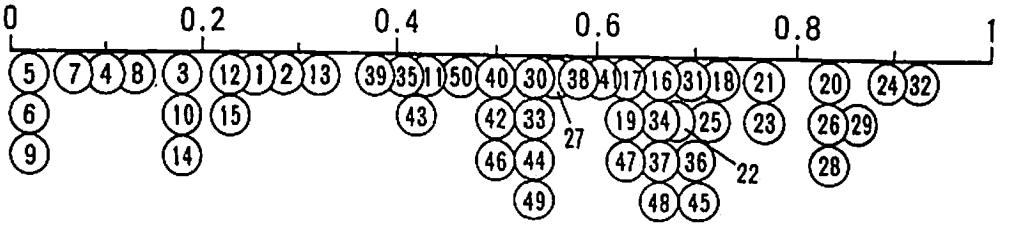

Figure 3 - Allele frequency distribution of the M allele of the MN blood group locus among 50 human populations. Population ID numbers correspond to those of Table 1.

appear in the M allele frequency distribution (Figure 3). Even if the overall amount of genetic differentiation is large between a pair of populations, it is possible to have similar frequencies at a particular allele of a locus.

One-dimensional space was enough to describe the allele frequency distribution of the MN blood group locus, since there were only two alleles at that locus. In general, (n-1)-dimensional space is necessary to describe the allele frequency distribution of a locus with n alleles. This geometrical representation is, however, not appropriate for two reasons. (1) As the number of alleles increases, visually feasible patterns can not be attained because of a high number of dimensions. (2) Allele frequency difference between two populations is not proportional to the expected divergence time of these populations. Therefore, we have computed Nei's (1972) standard genetic distances for all pairs of populations. It should be pointed out that the genetic distance was originally developed for multi-locus data, and the use of the genetic distance measure for single-locus data is only for comparison with the results obtained from a multi-locus analysis.

The neighbor-joining method (Saitou and Nei 1987) was applied to the obtained genetic distance matrices. The principle of minimum evolution (Cavalli-Sforza and Edwards 1967) is used for constructing unrooted trees in the neighbor-joining method, and various computer simu-

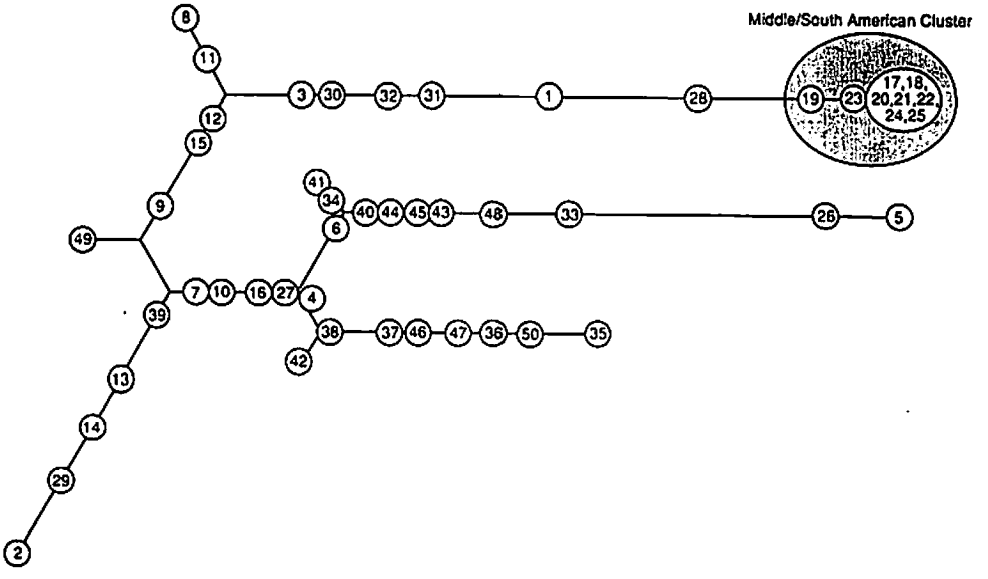

Figure 4 - A genetic affinity tree of 50 human populations based on the allele frequency data for the ABO blood group locus (4 alleles). Nei's standard genetic distances were used. Population ID numbers correspond to those of Table 1. Branch lengths are proportional to genetic distances.

lation studies have shown its high efficiency in recovering true trees (see Saitou 1991). Saitou et al. (1992, 1994), Imanishi et al. (1992), Nei and Roychoudhury (1993), and Bowcock et al. (1994) have used the neighbor-joining method to estimate genetic relationships of human populations.

Figure 4 is a neighbor-joining tree for a single-locus genetic distance matrix based on the ABO blood group polymorphism. We used data for the four allele (A1, A2, B, and O) system. When anti-A2 was not used in some populations, the A allele frequency was considered to be that of A1, and the frequency of A2 was assumed to be zero. The clustering pattern of populations is substantially different from that of Figure 3. Middle and South American populations (IDs 17-25) are monophyletic at the upper tail of the tree (the shaded oval in Figure 4), while most of the North American populations are intermingled with Oceanian ones. A tight clustering of South and Middle American populations is apparently because of a rather high frequency of the O allele among them. Asian populations (IDs 33-48) are more or less clustered, except for the Negritos (ID 39), though this group also contains two Papua New Guinean Highlander populations (IDs 4-6), North Alaskan Eskimo (ID 26), and Yoruba of Africa (ID 50). In any case, single locus polymorphism data are clearly not sufficient to estimate the genetic relationships of populations.

Therefore, we increased the number of polymorphic loci to three. Figure 5 is a neighbor-joining tree for a three-locus genetic distance matrix based on the ABO, MN, and Rh blood group polymorphisms. Data for an eight-allele system were used for the Rh blood group locus. Clustering of populations approaches that of geographical proximity; we can see the three clusters, "American", "Oceanian", and "Western". All the American populations except Yanomama (ID 17) and Guaymi (ID 25) fall into the "American" cluster (the shaded oval in Figure 5), which does not include non-American populations. The "Oceanian" cluster (the shaded triangle in Figure 5) consists of exclusively Oceanian populations, though Manus Islanders (ID 11) and

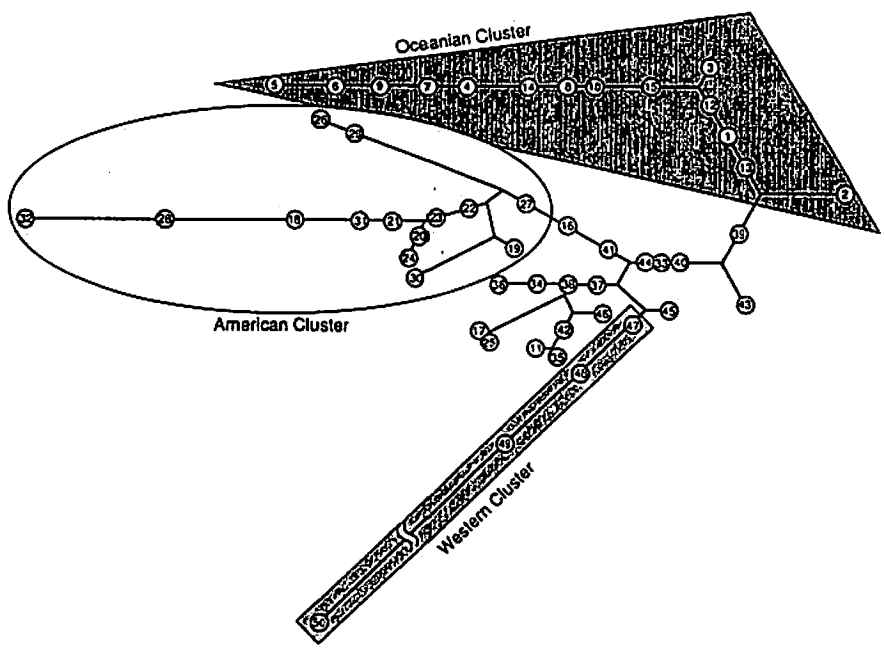

Figure 5 - A genetic affinity tree of 50 human populations based on the allele frequency data for MN, ABO, and Rh blood group loci. Nei's standard genetic distances were used. Population ID numbers correspond to those of Table 1. Branch lengths are proportional to genetic distances. Because the branch going to population 50 (Yoruba) was considerably longer than the remaining ones, it was truncated.

Samoa Islanders (ID 16) were not included in this cluster. The clear clustering of the M allele frequency distribution (Figure 3) definitely contributed to the formation of these two clusters. Yoruba of Africa (ID 50) is now located at the tip of the "Western" cluster, (the shaded rectangle in Figure 5) which consists of populations that are relatively closer to Africa than other populations (IDs 47-50), whereas Yoruba is genetically quite distant from the other populations.

The number of polymorphic loci was further increased to 6. The P blood group locus and two red blood cell enzyme loci (ACP1 and ESD) were added to the three loci used above. Unfortunately, however, those six polymorphic loci were not studied in all 50 populations listed in Table 1. Hence, we had to decrease the number of populations to 39; the IDs of the 11 populations not included in this comparison are 2, 6, 8, 11, 12, 13, 15, 30, 31, 34, and 42. Figure 6 is a neighbor-joining tree of the 39 populations. South American and North American populations both form monophyletic clusters, as do the Oceanian and Western ones (shaded ovals in Figure 6). Samoa Islanders (ID 16) are again not included in this Oceanian cluster, as in the case of Figures 3 and 5. Although East and Southeast Asian populations did not form a cluster in the strict sense, they do form a closely related group, with Samoa Islanders as its satellite; we call this the "East Eurasian" group (see Figure 6). It is clear that geographical clustering becomes apparent at this 6-locus stage.

We also computed D_A distances (Nei et al. 1983) for the same allele frequency data set and constructed a neighbor-joining tree. D_A is closely related to the chord distance (D_C) of Cavalli-Sforza and Edwards (1967), since $D_A = (\pi D_C/2)^2/2$. The obtained tree (not shown) was more or less similar to that of Figure 6 in which Nei's standard genetic distances were used; the four clusters and one group were formed as in the tree of Figure 6. The main difference in the tree using D_A distances is that the North American and South American clusters further form an American supercluster.

Finally we increased the number of polymorphic loci to 12. Allele frequency data for the following six polymorphic loci were added; Duffy, AK1, PGD, PGM1, Gc, and Hp. As in the case of Figure 6, a further nine populations were dropped from the comparison because of missing data. Their IDs are 3, 4, 9, 10, 25, 27, 35, 41, and 46. The remaining 30 populations cover Oceania, Asia, North and South America, Europe, and Africa. As in the case of the 6-locus data, we computed both Nei's standard genetic distances and D_A distances, and constructed neighbor-joining trees based on those two distance matrices. Because the resultant trees were more or less similar, the tree based on D_A distances is shown in Figure 7.

We observe a good correspondence between the genetic clustering and the geographical clustering of the 30 populations. North American, South American, Oceanian, and Western populations form monophyletic clusters as in Figure 6, and the North and South American clusters further form the American supercluster (However, this supercluster was not formed when Nei's standard genetic distance was used.) Within the "Western" cluster, Indian (ID 47), Iranian (ID 48), and English (ID 49) populations are located between Yoruba (ID 50) and the remaining circum-Pacific populations.

Alhough the East Eurasian populations again do not form a cluster, they are genetically closely related and are designated as the "East Eurasian" group in Figure 7. Within this group, four populations in East Asia (Japanese [ID 33], Han Chinese [ID 40], Ainu [ID 43], and Korean [ID 44]) form a cluster. By comparison, four Southeast Asian populations (Balinese [ID 36], Thais [ID 37], Filipino [ID 38], and Negritos [ID 39]) are scattered within the group. Saitou et al. (1994) also found a relatively close relationship among East Asian populations in contrast to the Southeast Asian populations. Interestingly, the Nepal population (ID 45) is closest to the South Indian population (ID 47) of the Western cluster. Samoa Islanders (ID 16), a Polynesian population, is again close to the "East Eurasian" group as in Figure 6. A relatively recent migration of

Polynesians from East Eurasia to the Pacific has been postulated based on linguistic and archeological studies (e.g., Bellwood 1977).

Although allele frequency data for only 12 polymorphic loci were used to construct the tree of Figure 7, five clusters or groups are clearly observed. A similar pattern was also observed by Nei and Roychoudhury (1993) who compared 26 human populations based on 29 polymorphic loci, and by Bowcock et al. (1994) who compared 14 human populations with 30 polymorphic microsatellite loci. Therefore, allele frequency data for only ten or so polymorphic loci seem to be sufficient for the allocation of a population to a continent.

A new method for estimation of the divergence time between two populations

Let us assume that a population splits into two at a certain time, and the resulting daughter populations start to differentiate without migration. For simplicity, the effective population size (N) remains the same for the two daughter populations, X and Y. Let p be the frequency of a particular allele of a particular locus at the time of the population split. This initial frequency is the same for populations X and Y, but it will gradually change due to random genetic drift. Let the present frequencies of this allele be x and y for populations X and Y, respectively, after t generations (see Figure 8). We are interested in the relationship among x, y, and t.

We assume that there is no mutation at this locus. When t is relatively small as in the case of human population differentiation over the past 100,000 years, we can ignore the effect of mutation for a typical protein locus, where genetic polymorphisms have been detected by electrophoresis. It is also assumed that the changes of allele frequencies are solely due to random genetic drift, i.e. all alleles are selectively neutral or equivalent. Under this simple model, the probability density

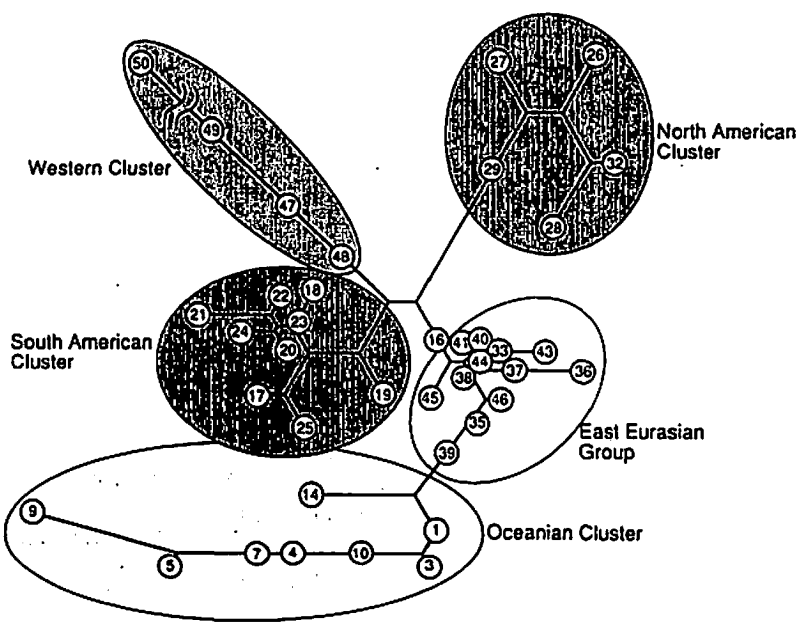

Figure 6 - A genetic affinity tree of 39 human populations based on the allele frequency data for 6 polymorphic loci (MN, ABO, Rh, P, ACP1, and ESD). Nei's standard genetic distances were used. Population ID numbers correspond to those of Table 1. Branch lengths are proportional to genetic distances. Because the branch going to population 50 (Yoruba) was considerably longer than the remaining ones, it was truncated.

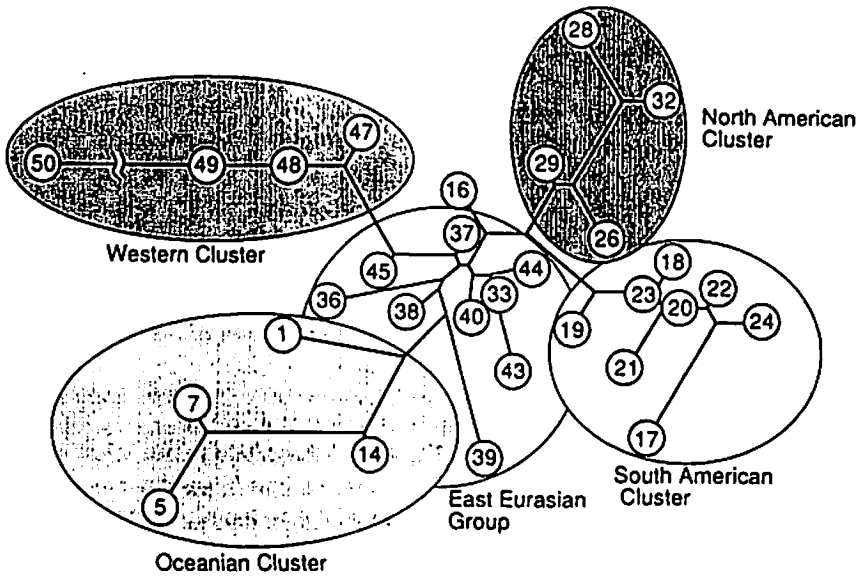

Figure 7 - A genetic affinity tree of 30 human populations based on the allele frequency data for 12 polymorphic loci (MN, ABO, Rh, P, ACP1, ESD, Fy, AK1, PGD, PGM1, Gc, and Hp). D_A distances were used. Population ID numbers correspond to those of Table 1. Branch lengths are proportional to genetic distances. Because the branch going to population 50 (Yoruba) was considerably longer than the remaining ones, it was truncated.

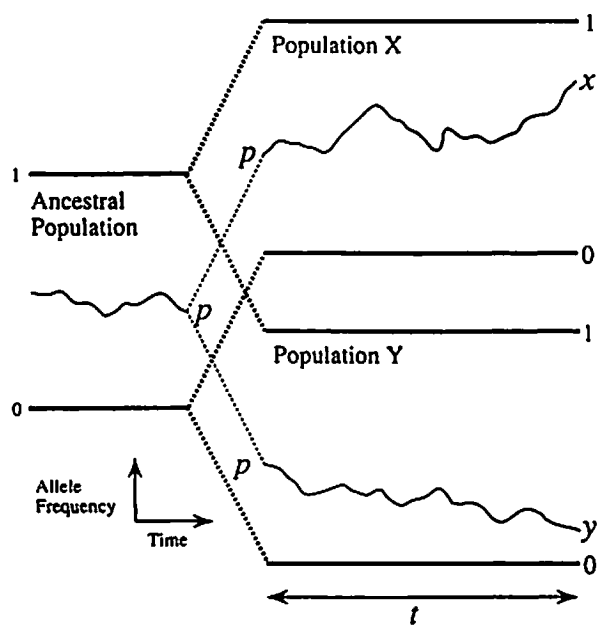

Figure 8 - A schematic relationship of a population split and allele frequency changes. An ancestral population splits into populations X and Y at a certain evolutionary time, when the frequency of an allele was p. At present, allele frequencies are x and y for populations X and Y, respectively.

$\phi(p, x; t)$ that the allele frequency lies between x and $x+dx$ at the t-th generation, given that it was p at the 0-th generation, is given by the following equation (Kimura 1955):

$$\phi(p, x; t) = \sum_{i=1}^{\infty} p(1-p)i(i+1)(2i+1)F(1-i,i+2,2,p) \times F(1-i,i+2,2,x)e^{-i(i+1)t/4N} \quad (1)$$

where $F(a,b,c,z)$ is a special form of hypergeometric function, and can be expressed in the following equation (Abramowitz and Stegun 1964):

$$F(a,b,c,z) = \sum_{n=0}^{-a} [(a)_n (b)_n z^n] / [(c)_n n!]. \quad (2)$$

Let us define the joint probability density $J(x, y; p, t)$ that allele frequencies of populations X and Y lie between x and $x+dx$ and y and $y+dy$, respectively, after t generations of divergence starting from the initial allele frequency p. Since populations X and Y change their allele frequency independently under the present assumption, we can compute $J(x, y; p, t)$ by simply multiplying the ϕ's as follows:

$$J(x, y; p, t) = \phi(p, x; t) \times \phi(p, y; t). \quad (3)$$

Because x and y are observable values, this joint probability density can be considered as a likelihood function of p and t. In the following discussion, we take this likelihood approach. A program for computing values of $J(x, y; p, t)$ was written and some numerical results are given below. Since t/N is involved in equation (1), we use T (= t/N) which measures the divergence time with the unit of N (population size).

Our first question is, "which allele frequency p will give the highest likelihood of $J(x, y; p, T)$ when the divergence time T is set to be constant?" Figure 9 shows a schematic representation

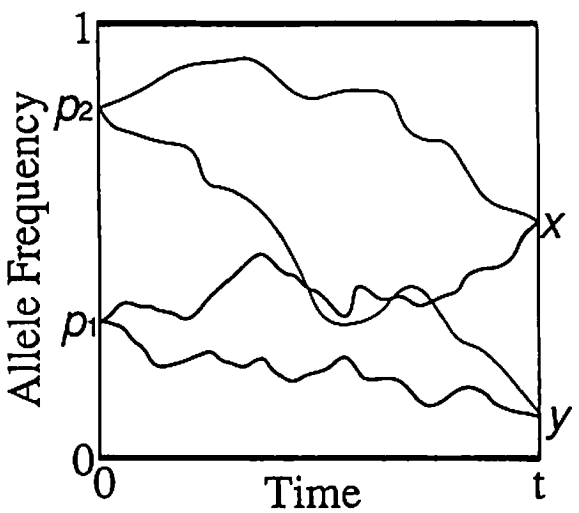

Figure 9 - Two possibilities of allele frequency changes when T (divergence time) is identical. p_1 and p_2 are two possible initial allele frequencies, and x and y are the present allele frequencies for populations X and Y, respectively.

of two possibilities of allele frequency changes for populations X and Y that started to diverge T generations ago, with the initial allele frequency of either p_1 or p_2. Because the expected allele frequency does not change under pure random genetic drift, a p that is close to the average of x and y should give a higher likelihood than other p's. According to this conjecture, p_1 will give a higher likelihood value than p_2 in Figure 9. Figure 10 shows likelihood distributions for $x = 0.2$ and $y = 0.6$, under the five different values of divergence time T. As expected, the highest likelihood is obtained when $p = 0.4$, the average of x and y, for all the five T values.

We now ask the second question; which divergence time T will give the highest likelihood of $J(x, y; p, T)$ when the allele frequency p is set to be constant? Figure 11 illustrates two possibilities of allele frequency changes for two populations with the same p value, and likelihood distributions for $x = 0.2$ and $y = 0.6$ under four different p values are presented in Figure 12. In general, likelihood curves initially go up quickly then go down gradually, but the T value (Tmax) that gives the highest likelihood depends on the p value. As p approaches to the average (0.4) of x and y, Tmax becomes smaller, and Tmax is around 0.34, or 0.34N generations, when p = 0.4.

Generally speaking, we do not know the initial allele frequency (p) of the ancestral population, and any frequency value should have the same probability as the initial frequency. Thus we define the mean likelihood:

$$S(x, y; T) = \sum_{p=1/m}^{1} 1/m \times J(x, y; p, T), \qquad (4)$$

where m is the number of divisions for $0<p<1$ [Note that $J(x, y; 1, T) = 0$ by definition]. We arbitrarily set $m = 100$ in the numerical computation.

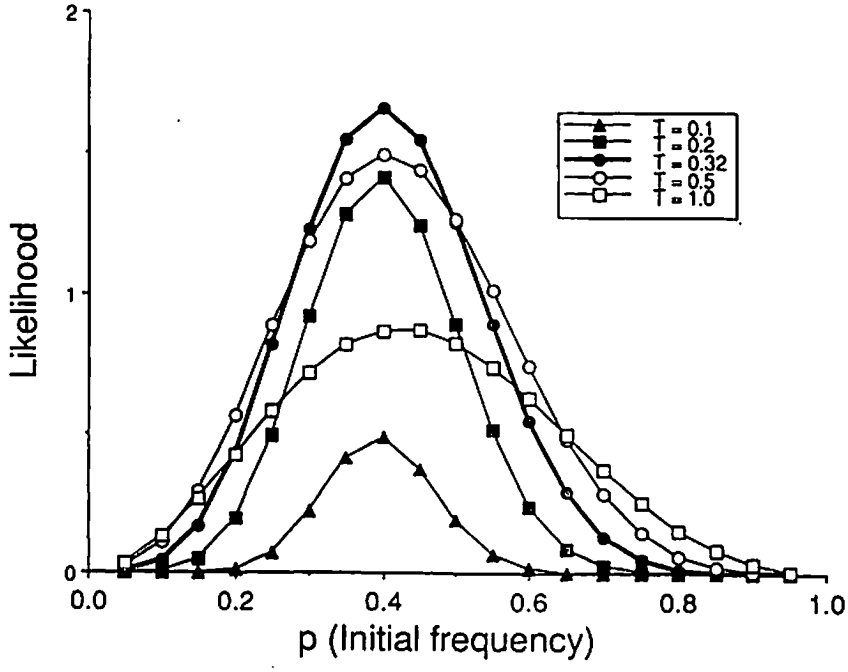

Figure 10 - Likelihood distributions when $x = 0.2$ and $y = 0.6$, for the five different values of T (divergence time).

Figure 13 shows the likelihood distributions when $x = 0.2$ and $y = 0.6$, for the two conditions of p. Closed circles are for $S(x, y; T)$, and open circles are for $J(x, y; p, T)$ when p = 0.4 (mean of x and y). Tmax values for the former and latter cases are 0.32N and 0.46N, respectively. Tmax for $S(x, y; T)$ is about 40% larger than that for $J(x, y; [x+y]/2, T)$. Figure 14 is the likelihood distribution for these two conditions when x = 0.01 and y = 0.99. This is rather an extreme case in which different alleles are about to be fixed in each population. Tmax values for $J(x, y; [x+y]/2, T)$ and for $S(x, y; T)$ are 1.19 and 1.25, respectively, which are much larger than those in Figure 13.

We applied this new method to the estimation of the divergence time between Japanese and the Chukchi of Siberia and between Japanese and St. Lawrence Island Eskimos. Allele frequency data for the three polymorphic loci in which allele frequencies were most different in the two populations were selected from the data collected by Roychoudhury and Nei (1988) (see Table 2). From these data, Tmax was computed both for $J(x, y; [x+y]/2, T)$ and for $S(x, y; T)$. Tmax values for J(x, y; [x+y]/2, T) for Japanese and Chukchi were 0.27N, 0.14N, and 0.12N for Acid phosphatase 1, Kidd, and P loci, respectively (average = 0.18N). Those for the comparison of Japanese and Eskimos were 0.29N, 0.21N, and 0.25N for Esterase D, Glyoxalase 1, and Orosomucoid 1 loci, respectively (average = 0.25N). Thus the divergence time between Japanese and Eskimos seems to be slightly greater than that between Japanese and Chukchi.

The same conclusion is obtained when we used Tmax values for S(x, y; T). If N (effective population size) = 5,000 and one generation is 25 years, 0.18N and 0.25N generations correspond to 22,500 years and 31,250 years, respectively.

The new method proposed in the present study is based on simplified assumptions, and may not reflect the real situations. Therefore, the divergence time estimates obtained by using this method are only tentative ones. Population size is the most vulnerable point in this method, because it is difficult to estimate, may vary from time to time, and may differ in each population. Furthermore, these time estimates rely on the Tmax, or the time that gives the highest likelihood values. It would be desirable to give confidence limits to those maximum likelihood estimates when this method is polished in the future.

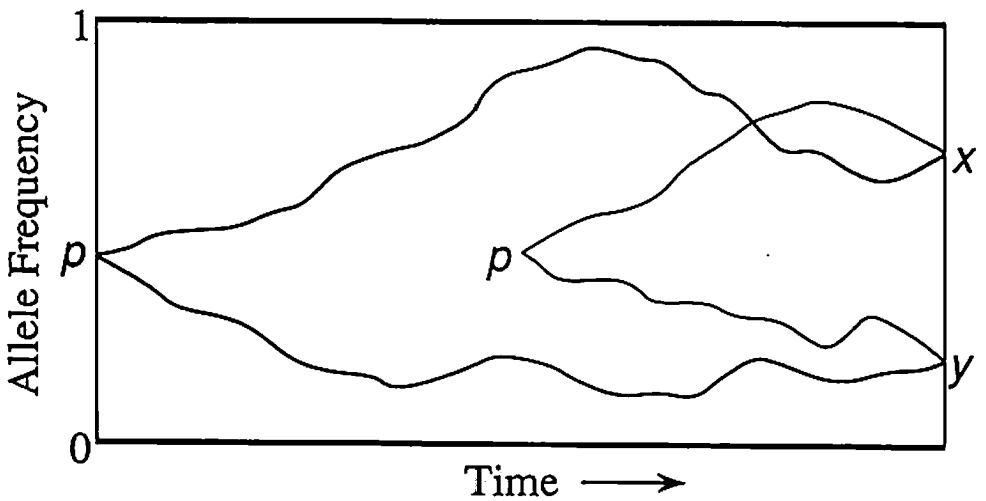

Figure 11 - Two possibilities of allele frequency changes when p (the initial allele frequency) is identical. x and y are the present allele frequencies for populations X and Y, respectively.

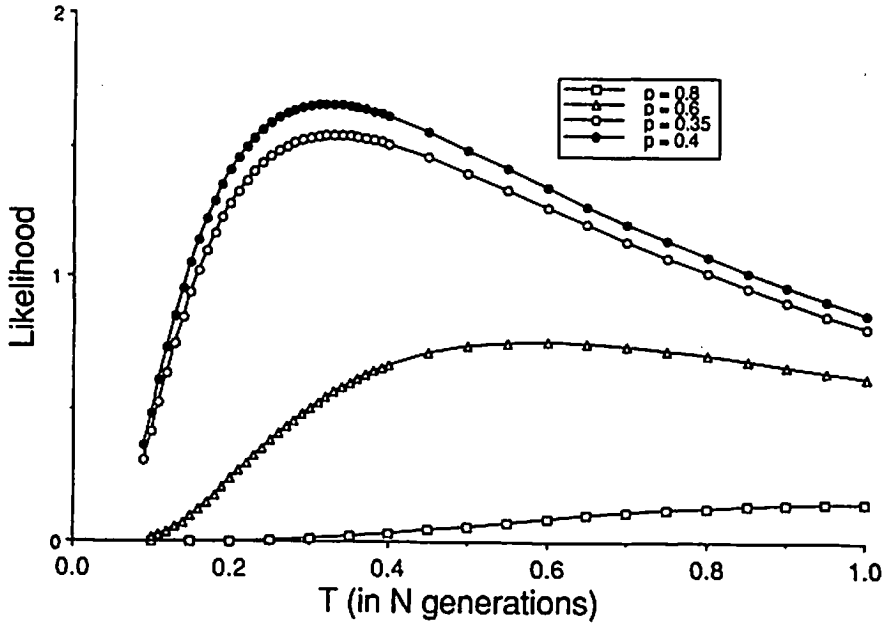

Figure 12. Likelihood distributions when $x = 0.2$ and $y = 0.6$, for four different initial allele frequencies, p.

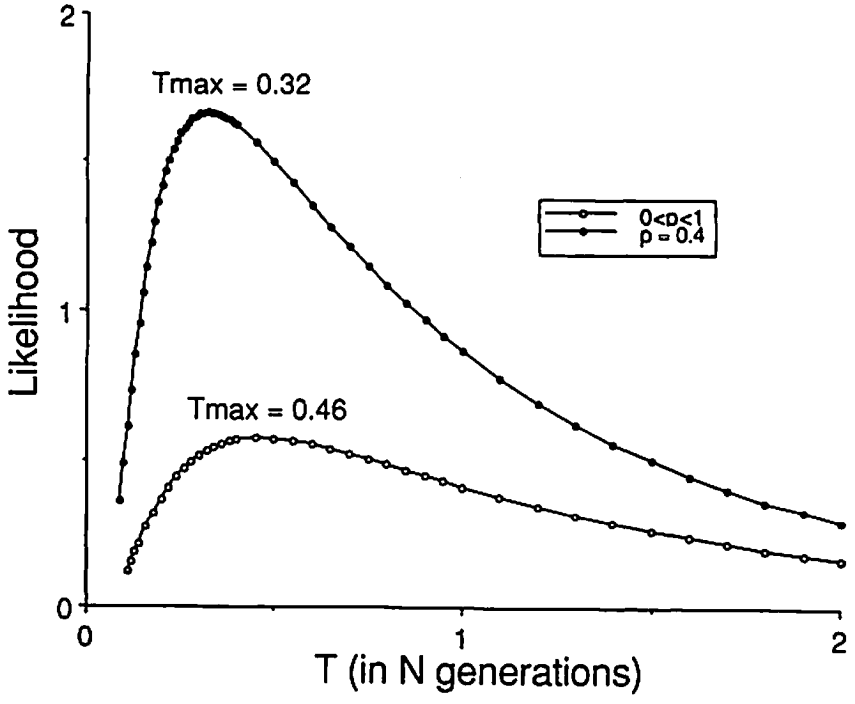

Figure 13 - Likelihood distributions when $x = 0.2$ and $y = 0.6$, under the two values of the initial allele frequency, p. Closed circles are for $S(x, y; T)$, and open circles are for $p = 0.4$.

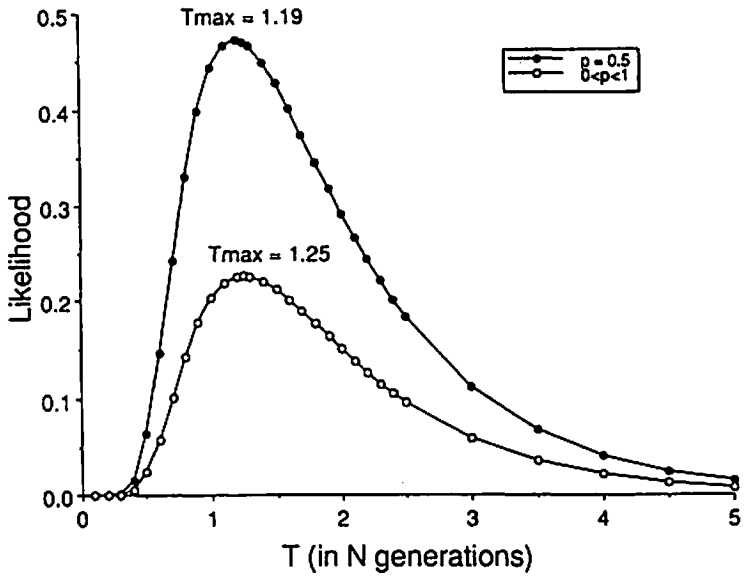

Figure 14 - Likelihood distributions when $x = 0.01$ and $y = 0.99$, for the two values of the initial allele frequency, p. Closed circles are for $S(x, y; T)$, and open circles are for $p = 0.5$.

Conclusion

Many studies have been conducted to reconstruct phylogenetic relationships of human populations. The differentiation of human populations, however, does not necessarily follow a simple model of population divergence that produces phylogenetic relationships of populations. This is because relatively high rates of recent gene flow often blur the effect of isolation in the past. This high rate of migration across a long distance seems to be specific to humans. In contrast, a large-bodied mammalian species is expected to have great genetic differences among populations living in different continents because of low gene flow. Of course, this unique feature of humans is due to our technological advancement for movement, such as the invention of the boat. We can easily predict that the human population as a whole is on the path toward a panmictic population in the future if the current high rate of gene flow between populations continues.

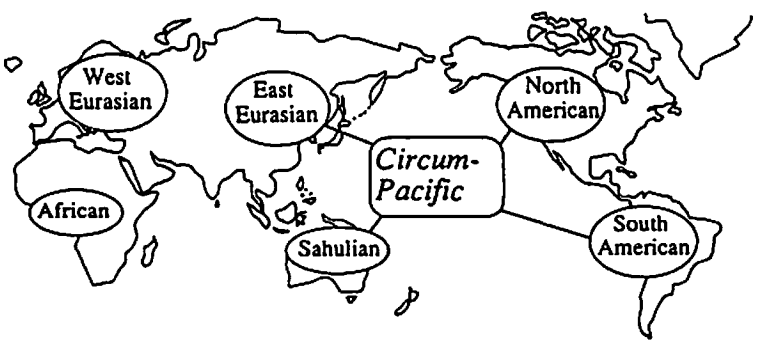

Figure 15 - Classification of *Homo sapiens* at around 10,000 years ago, at the start of the Holocene. The groupings correspond to the six geographical regions in which the genetically clustered populations lived.

TABLE 2 - Estimation of the divergence time between the Japanese and two Northern populations

(A) Japanese (population X) and the Chukchi in Chukotka and Kamuchatka, a Siberian population (population Y)

Locus 1) Acid phosphatase 1, allele a: x = 0.210, y = 0.570
Tmax = 0.27N for J(x, y; [x+y]/2, T)
Tmax = 0.39N for S(x, y; T)
Locus 2) Kidd blood group, allele JkSUP4(a): x = 0.472, y = 0.736
Tmax = 0.14N for J(x, y; [x+y]/2, T)
Tmax = 0.14N for S(x, y; T)
Locus 3) P blood group, allele PSUP4(1): x = 0.198, y = 0.422
Tmax = 0.12N for J(x, y; [x+y]/2, T)
Tmax = 0.20N for S(x, y; T)

(B) Japanese (population X) and Eskimos (Inuit) from St. Lawrence Island (population Y)

Locus 1) Esterase D, allele 2: x = 0.388, y = 0.065
Tmax = 0.29N for J(x, y; [x+y]/2, T)
Tmax = 0.40N for S(x, y; T)
Locus 2) Glyoxalase 1, allele 1: x = 0.088, y = 0.360
Tmax = 0.21N for J(x, y; [x+y]/2, T)
Tmax = 0.32N for S(x, y; T)
Locus 3) Orosomucoid 1, allele 1: x = 0.221, y = 0.573
Tmax = 0.25N for J(x, y; [x+y]/2, T)
Tmax = 0.38N for S(x, y; T)

Therefore, I would like to propose a new viewpoint for the study of the genetic relationships of human populations. Although the present human populations are genetically heterogeneous, it is inappropriate to reconstruct rooted phylogenetic trees of many populations based solely on the present-day data. We should instead produce unrooted trees of human populations only to show the genetic affinity among them. However, the global movement of ancient human populations until the end of the Pleistocene can be traced even if we use the data from present-day populations, because human movements on the globe seem to have accelerated only after the Neolithic Revolution started, namely in last 10,000 years. The traces of ancient human movements and the subsequent genetic differentiation are clearly observed in the unrooted tree of Figure 7.

These topological patterns of the genetic affinity tree of Figure 7 closely resemble the geographical constellation of the six geographical areas: Africa, West Eurasia, East Eurasia, North America, South America, and Sahul land. Sahul land, or the Sahul shelf, existed until the last glacier ended about 10,000 years ago, and later was separated into Australia and Papua New Guinea (White & O'Connell 1979; Ballard 1993). I would like to propose a new classification of human populations based on this genetic affinity tree. Figure 15 shows six human population clusters at the end of the Pleistocene: African, West Eurasian, East Eurasian, Sahulian, North American, and South American. The great movement of Polynesian people occurred much later, and the center of the Pacific was not yet populated at that time. Thus the four clusters surrounding the Pacific (East Eurasian, Sahulian, North American, and South American) can be further grouped to form a "Circum-Pacific"supercluster. This supercluster corresponds to the "Pan-Mongoloid" cluster of Saitou et al. (1992).

Ever since J.F. Blumenbach proposed a racial classification based on physical characters

(see Bendyshe 1865), terms such as Mongoloid, Caucasoid, or Negroid have been frequently used. If we are interested in the differentiation of human populations from the same viewpoint as that for large bodied mammals, we should use population names based on their area of habitation, as suggested in the present study. Finally, let us cease to construct rooted phylogenetic trees of human populations living after the Holocene started, since the substantial amount of gene flow has obscured their phylogenetic histories.

ACKNOWLEDGMENTS—This study was partially supported by Grants-in-Aid for Scientific Research on Priority Areas (Prehistoric Mongoloid Dispersals) of the Ministry of Education, Science and Culture, Japan. A part of this paper was presented at the First International Conference on the Prehistoric Mongoloid Dispersal held November 16-21 1992 at Tokyo. I thank Dr. Takeru Akazawa for giving me the opportunity to proceed with the present study. I also thank Dr. Michiko Intoh for giving me the information on the Sahul land. Finally, I appreciate Dr. C. Loring Brace for his comments on the population terminology at that Conference. I came up with the new classification of human populations given in the Conclusion only after considering his comments seriously.

References

Abramowitz M. & Stegun I.A., 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards, Washington, D.C.

Ballard C, 1993. *Stimulating minds to fantasy? A critical etymology for Sahul.* In (Smith M.A. Spriggs M., & Frankhansen B. eds), Occasional Papers in Prehistory, No. 24, Sahul in Review, pp. 17-23. Department of Prehistory, Australian National University, Sydney.

Bellwood P.S., 1977. *Man's Conquest of the Pacific.* Collins, London and Auckland.

Bendyshe T., ed., 1985. *Anthropological Treatises of Johann Friedrich Blumenbach.* Longman, London.

Bowcock A.M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J.R., & Cavalli-Sforza L.L., 1994. *High resolution of human evolutionary trees with polymorphic microsatellites.* Nature, 368: 455-457.

Cann R.L., Stoneking M., & Wilson A. C., 1987. *Mitochondrial DNA and human evolution.* Nature, 325: 31-36.

Cavalli-Sforza L.L. & Edwards A.W.F., 1967. *Phylogenetic analysis: models and estimation procedures.* American Journal of Human Genetics, 19: 233-257.

Cavalli-Sforza L.L., Piazza A., Menozzi P., & Mountain J., 1988. *Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data.* Proceedings of National Academy of Sciences USA, 85: 6002-6006.

Darwin C., 1859. *The Origin of Species by Means of Natural Selection.* John Murray, London.

Harihara S., Hirai M., Suutou Y., Shimizu K., & Omoto K., 1992. *Frequency of a 9-bp deletion in the mitochondrial DNA among Asian populations.* Human Biology, 64: 161-166.

Hertzberg M., Mickleson K.N.P., Serjeantson S.W., Prior J.F., & Trent R.J., 1989. *An Asian-specific 9-bp deletion of mitochondrial DNA is frequently found in Polynesians.* American Journal of Human Genetics, 44: 504-510.

Horai S., Kondo R., Nakagawa-Hattori Y., Hayashi S., Sonoda S., & Tajima K., 1993. *Peopling of the Americas, founded by four major lineages of mitochondrial DNA.* Molecular Biology and Evolution, 10: 23-47.

Imanishi T., Wakisaka A., & Gojobori T., 1992. *Genetic relationships among various human populations indicated by MHC polymorphisms.* In (Tsuji, K., Aizawa, M., & Sasazuki, T. eds.). HLA1991 Vol. 1, pp. 627-632. Oxford Univ. Press, Oxford.

Kimura M., 1955. *Solution of a process of random genetic drift with a continuous model.* Proceedings of National Academy of Sciences USA, 41: 144-150.

Nei M., 1972. *Genetic distance between populations.* American Naturalist, 106: 283-292.

Nei M., 1987. *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Nei M. & Roychoudhury A. K., 1974. *Genetic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids.* American Journal of Human Genetics, 26: 421-443.

Nei M. & Roychoudhury A.K., 1982. *Genetic relationship and evolution of human races.* Evolutionary Biology, 14: 1-59.

Nei M. & Roychoudhury A.K., 1993. *Evolutionary relationships of human populations on a global scale.* Molecular Biology and Evolution, 10: 927-943.

Nei M., Tajima F., & Tateno Y., 1983. *Accuracy of estimated phylogenetic trees from molecular data.* II. Gene frequency data. Journal of Molecular Evolution, 19: 153-170.

Nei M. & Takezaki N., 1994. *Estimation of genetic distances and phylogenetic trees from data analysis.* Proceedings of the 5th World Congress on Genetics Applied to Liverstok Production, 21: 405-412.

Passarino G., Semino O., Modiano G., & Santachiara-Benerecetti A.S., 1993. *COII/tRNALys Intergenic 9-bp deletion and other mtDNA markers clearly reveal that the Tharus (Southern Nepal) have Oriental affinities.* American Journal of Human Genetics, 53: 609-618.

Roychoudhury A.K. & Nei M., 1988. *Human Polymorphic Genes: A World Distribution.* Oxford University Press, Oxford.

Saitou N., 1991. *Statistical methods for phylogenetic tree reconstruction.* In (Rao C.R. & Chakraborty R. eds.), Handbook of Statistics, Volume 8: Statistical Methods for Biological and Medical Sciences, pp. 317-346. Elsevier Science Publishers B.V., Amsterdam.

Saitou N. & Nei M., 1987. *The neighbor-joining method: a new method for constructing phylogenetic trees.* Molecular Biology and Evolution, 4: 406-425.

Saitou N., Omoto K., Du C., & Du R., 1994. *Population genetic study in Hainan Island, China. II. Genetic affinity analyses.* Anthropological Science, 102: 129-147.

Saitou N., Tokunaga K., & Omoto K., 1992. *Genetic affinities of human populations.* In (Roberts, D.F., Fujiki, N., & Torizuka, K. eds.). Society for Study of Human Biology Symposium 33: Isolation, Migration, and Health, pp. 118-129. Cambridge University Press, Cambridge.

Sambuughin N., Petrishchev V.N., & Rychkov Yu. G., 1991. *DNA polymorphism in Mongolian population: analysis of restriction endonuclease polymorphism of mitochondrial DNA.* Genetika, 27: 2143-2151 (in Russian).

Schurr T.G., Ballinger S.W., Gan Y.-Y., Jodge J.A., Merriwether D.A., Lawrence D.N., Knowler W.C., Weiss K.M., & Wallace D.C., 1990. *Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages.* American Journal of Human Genetics, 46: 613-623.

Shields G.F., Hecker K., Voevoda M. I., & Reed J. K., 1992. *Absence of the Asian-specific region V mitochondrial marker in native Beringians.* American Journal of Human Genetics, 50: 758-765.

Shields G.F., Schmiechen A.M., Frazier B.L., Reed A., Voevoda M.I., Reed J.K., & Ward R.H., 1993. *mtDNA sequences suggest a recent evolutionary divergence for Beringian and northern American populations.* American Journal of Human Genetics, 53: 549-562.

Torroni A., Sukernik R.I., Schurr T.G., Starikovskaya Y.B., Cabell M.F., Crawford M.H., Comuzzie A.G., & Wallace D.C., 1993a. *mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with native Americans.* American Journal of Human Genetics, 53: 591-608.

Torroni A., Schurr T.G., Cabell M.F., Brown M.D., Neel J.V., Larsen M., Smith D.G., Vullo C.M., & Wallace D.C., 1993b. *Asian affinities and continental radiation of the four founding native American mtDNAs.* American Journal of Human Genetics, 53: 563-590.

White J.P. & O'Connell J.F., 1979. *Australian prehistory: new aspects of antiquity.* Science, 203: 21-28.

Wrischnik L.A., Higuchi R.G., Stoneking M., Erlich H.A., Arnheim N., & Wilson A. C., 1987. *Length mutations in human mitochondrial DNA; direct sequencing of enzymatically amplified DNA.* Nucleic Acids Research, 15: 529-542.

Received 27 July 1994 *Accepted 22 September 1994*