No.45

In: THE ORIGINS AND PAST OF MODERN HUMANS - TOWARDS RECONCILIATION

GENETIC STUDIES ON THE ORIGIN OF MODERN HUMANS

Masatoshi Nei
Institute of Molecular Evolutionary Genetics
and Department of Biology
The Pennsylvania State University
University Park, PA 16802 USA

ABSTRACT

Although African populations have been shown to be most divergent from any other human populations, it has been difficult to establish the root of the phylogenetic tree of human populations since the rate of evolutionary change may vary from population to population owing to the fluctuation of population size and other factors. However, the root can be determined by using the chimpanzee as an outgroup and by employing proper statistical methods. Using this strategy, we constructed phylogenetic trees of human populations for five different sets of gene frequency data. All these data sets showed that the root is located in the branch connecting African and non-African populations, and in the four data sets the root was established at a statistically significant level. These results indicate that Africans are the first group of people that split from the rest of the human populations. These results together with the studies of the time of divergence between Africans and non-Africans support the out-of-Africa theory of origins of modern humans.

INTRODUCTION

Currently a great controversy is going on over the origin of modern humans (Homo sapiens). One hypothesis termed the evolution by regional continuity or multiregional theory [1] maintains that our ancestral species H. erectus, whose brain size was about half that of H. sapiens, moved out of Africa and spread to various parts of the world more than one million years (MY) ago, and that H. sapiens evolved gradually from H. erectus worldwide with an effect of gene flow. Yet this hypothesis claims that the extent of gene flow was so low that several regional characters such as the shovel-shaped incisors in northeastern Asians and the prominent eyebrow ridge in Australian aborigines have remained unchanged for more than 1 MY from the time of their ancestral H. erectus. The other hypothesis called the out-of-Africa theory [2,3] proposes that H. sapiens originated in Africa about 100,000 – 200,000 years ago and all the present human populations are descendants of the population or the population that moved out of Africa about 100,000 years ago.
The out-of-Africa theory was initially based on a phylogenetic analysis of RFLP data of mitochondrial DNAs (mtDNAs) sampled from different parts of the world [2] and some paleontological data [3]. However, this study was criticized by a number of authors. First, the topology of the mtDNA tree reconstructed was statistically unreliable [4], and it was difficult to prove that the African populations separated first from the rest of the human populations. Second, the estimate of the time of the deepest split of mtDNA lineages had a large standard error, so the time estimate could be as old as 800,000 years depending on the assumptions made. This made it difficult to distinguish between the two hypotheses of human origins. Third, the mitochondrial DNA is inherited as though it is a single gene. Therefore, it is extremely difficult to infer the phylogenetic tree of human populations from mtDNA variation. Vigilant et al. [5] examined this problem by using DNA sequence data for the control region (about 1000 bp) of mtDNA. However, their data were still insufficient to resolve the above problem [6-8].

To study the evolutionary history of human populations, it is important to examine population trees based on many nuclear genes [9,10]. This can be done by using allele frequency data for many different loci. Another important thing is to use data from an outgroup species to root the human population tree. The best species to be used for this purpose is the chimpanzee, because this is now known to be the species closest to humans. We have therefore collected all nuclear allele frequency data that can be used for this purpose and constructed phylogenetic trees of human populations with the root [11]. In this paper I summarize the results obtained and discuss their implications for the origins of human populations.

DATA SETS USED

We collected five different sets of allele frequency data for major human population groups and the chimpanzee. The first data set (microsatellite DNA data set I) came from Bowcock et al. [12], who generated gene frequency data for 30 microsatellite loci for 14 human populations and for 25 loci for the chimpanzee. In this study, we used data for 25 loci that are shared by the human and the chimpanzee. The second set of data (microsatellite DNA data set II) was taken from Deka et al. [13], who published allele frequency data for 8 microsatellite loci from 8 human populations and the chimpanzee. The human populations and most microsatellite loci studied by these authors were different from those of Bowcock et al. [12]. The third data set was restriction fragment length polymorphism (RFLP) data obtained by Mountain and Cavalli-Sforza [14] for 79 anonymous DNA marker loci from 8 human populations and the chimpanzee. The fourth set was protein polymorphism data shared by three major races of man and the chimpanzee [15,16]. There are many polymorphic protein loci that have been examined for human populations and the chimpanzee, but the number of loci shared by the three major races and the chimpanzee we could use was 15. They were ACP1, ADA, AK1, G6PD, GSR, HBB, PGM1, PGD, ORM1, PI, CP, C3, GC, HPA, and TF (see [17] for gene symbols). The fifth set of data was allele frequency data for 4 polymorphic loci for Alu sequence insertion in the human genome [18]. This insertion apparently occurred after the human-chimpanzee divergence, so that they are polymorphic only in humans.
For each of the five different sets of data we computed pairwise genetic distances using Nei et al.'s $D_A$ distance [19]. This distance measure is not proportional to evolutionary time but is efficient for obtaining correct phylogenetic trees (topologies) under various evolutionary conditions whether the mutation pattern follows the infinite-allele model or the stepwise-mutation model [see 20] or whether population size changes or not [19,21,22]. Since we are interested primarily in the topology of the phylogenetic tree of human populations rather than estimating branch lengths, we used this distance measure. For constructing phylogenetic trees from pairwise distances, we used the neighbor-joining method [23], because human populations are known to vary considerably over evolutionary time and thus the evolutionary rate would vary from population to population [24]. The reliability of the phylogenetic trees obtained was examined by Felsenstein's [25] bootstrap test with 500 replications.

PHYLOGENETIC TREES

Microsatellite DNA Data Set I

The phylogenetic tree obtained from this data set is given in Fig. 1. The root of the human phylogenetic tree is located in the branch between Africans and non-Africans, and a bootstrap test indicates that both interior branches connecting the chimpanzee and Africans (97%) and the chimpanzee and non-Africans (99%) are statistically significant. Bowcock et al. [12] gave the root to the same place by using the midpoint method, but this method gives no statistical assurance. They did not use the chimpanzee as an outgroup population.

However, if we consider the branching orders of human populations only, our results are identical with Bowcock et al.'s except for three African populations and three Asian populations, of which the branching order is not statistically significant. They are also similar to those of Nei and Roychoudhury [26] for classical genetic markers, though in their tree Oceanians are slightly closer to eastern Asians than Amerindians are. In none of these studies, however, the branching order of the three groups of populations has been statistically established.

Microsatellite DNA Data Set II

Deka et al. [13] already constructed the neighbor-joining tree for this set of data with $D_A$ distance. However, for the sake of comparison it is reproduced in Fig. 2. This tree is identical with Deka et al.'s, though it is presented in a different form and the bootstrap values are slightly different. In the present tree, the bootstrap value (83%) for the branch separating Africans from non-Africans is lower than 95%. This is probably due to the fact that the number of loci used is small. However, Felsenstein's bootstrap test is known to be conservative [e.g., 27-30]. By contrast, Dopazo's [31] interior branch test by means of bootstrapping generally gives more accurate test results [17]. We therefore applied this method to the branch between Africans and non-Africans and obtained a confidence
probability of 98 percent. This suggests that the root of the human population tree is again located between the African and non-African populations, though some caution is required in this case because the number of loci used is small. Little can be said about the branching order of Oceanians, Europeans, Asians, and Amerindians with this data set.

RFLP Data

Fig. 3 shows the phylogenetic tree obtained from this data set. The topology of this tree is the same as that of the tree in Fig. 1, though the number of populations examined is smaller here. The root of the tree is again placed in the branch connecting Africans and non-Africans, and the interior branch connecting the chimpanzee and non-African human populations is supported at a bootstrap probability of 99 percent, indicating that Africans were the first to split from the rest of the world populations.

Protein Polymorphism Data

Although only three major human populations (Africans, Europeans, and Asians) are used, the phylogenetic tree obtained (Fig. 4) again shows that the root exists between Africans and non-Africans and that the location of the root is statistically supported.

Alu Insertion Polymorphism Data

The final data set is Alu-insertion polymorphism data. Batzer et al. [18] examined the allele frequencies of 16 human populations, including African Americans, two populations from Indonesia, two Amerindian populations in South America, and two pygmy populations. In this study we eliminated African Americans because this population has been admixed with Caucasians [32]. We also combined each of the two Indonesian populations, two Amerindian populations, and two pygmy populations into one group, because the number of individuals examined was small.

The phylogenetic tree based on $D_\lambda$ distance is given in Fig. 5. Since only four loci were used, the reliability of the branching pattern is low. However, it is interesting to note that the root of the tree is still located in the branch between Nigerians and other populations.

DISCUSSION

We have seen that all five sets of gene frequency data indicate that the root of the human populations exists in the branch connecting Africans and non-Africans and that the root is statistically supported in four data sets. In addition, the mtDNA trees obtained for worldwide samples are also consistent with this inference, though statistical support is weak in these trees [2,5,33]. Therefore, it now seems to be safe to conclude that sub-Saharan Africans were the first to split from the rest of the world populations. Previous studies [12,34-36] have suggested that Africans are genetically most divergent from the other populations, but these results were not sufficient to establish the root of the tree of human populations because the evolutionary rate varies considerably with population [8,24].
Although the pattern of the phylogenetic trees obtained by Nei and Takezaki [11] is consistent with the out-of-Africa theory, it is also important to examine the divergence time between Africans and non-Africans to establish the out-of-Africa theory. Fossil records are not decisive on this issue, but supporters of the out-of-Africa theory [e.g., 37] interpret them to imply that modern humans moved out of African about 100,000 years ago.

Nei and Roychoudhury [35] seem to be the first to study this problem using molecular data and obtained an estimate of 117,000 years. This estimate was based on only 33 protein loci, but the estimate hardly changed when they increased the number of loci to 62 (see Table 1). Note these estimates were obtained before the out-of-Africa theory was formally presented. Recently, Goldstein et al. [38] used microsatellite DNA loci for this purpose and obtained an estimate of 152,000 years, whereas Nei [10] obtained 115,000 years for the same data set under a slightly different assumption. In addition, a number of authors used mtDNAs to estimate this divergence time under various assumptions, but most estimates have a large confidence interval (Table 1). The estimate with the smallest confidence interval is Horai et al.'s [39]. Although this estimate is considered to be biased upward because a mtDNA tree is a gene tree rather than a population tree, even the 95 percent upper bound of confidence of their estimate was about 180,000 years. Therefore, molecular estimates are generally consistent with the idea that Africans separated from the rest of the world population 100,000 - 200,000 years ago.

These results support the out-of-Africa theory rather than the multiregional theory. However, one can argue that modern humans (H. sapiens) originated in Eurasia and some of the Eurasian populations moved into Africa (H. Harpending, personal communication). If this is the case, one would expect that some Eurasian populations are more different from the rest of the Eurasians than Africans are, because the differentiation of Eurasian populations should have occurred before one of them moved to Africa. Actual data do not show this pattern of genetic differentiation. One would also wonder why the peopling of the Americas and Australia occurred so late if modern humans originated in Eurasia. Furthermore, recent archeological studies suggest that some modern humans who produced sophisticated tools were already living in sub-Saharan areas about 90,000 years ago [40,41]. If this new finding holds up, an advanced tool-making culture evolved much earlier in Africa than in Europe (about 40,000 years ago), and the African origin of modern humans is supported.

Some authors [e.g., 42] have contended that the first splitting of African populations is also consistent with the multiregional theory, because if the extent of gene flow among different populations has been sufficiently large, the genetic depth of the first split of human populations would look shallow. However, if this is the case, it seems very difficult to maintain regional continuity of certain morphological characters (e.g., shovel-shaped incisors in northeastern Asia and prominent eyebrow ridge in southeastern Asia) for more than 1 MY. To maintain this regional continuity, the extent of gene flow must be very low. If so, the age of the latest common ancestor of human mtDNAs should be much higher than the observed value of about 145,000 years [10,39,43].
There are several other reasons to believe that the multiregional theory or gene flow hypothesis is suspect. First, if the genes from different geographical areas are mixed extensively by migration and the extent of gene flow is the primary factor to determine the distance values between populations, the tree constructed would not reflect the real evolutionary history of the populations and the root located would be unreliable; the location of the root is expected to vary from gene to gene. However, our analysis shows that the five different sets of data give the same root and similar phylogenetic trees. Mitochondrial DNA data also gives essentially the same root [6], though it is not statistically significant. This suggests that our tree is not unreasonable and the effect of gene flow is probably unimportant for our purpose.

Second, if we assume that the effects of mutation, migration, and genetic drift are balanced between two populations with either the infinite-allele or the stepwise-mutation model, the equilibrium value of Nei's genetic distance (D) is approximately given by $D = v/v + m$ when $D$ is small, where $v$ and $m$ are the mutation rate and the migration rate, respectively [9,44]. Using allozyme data, Nei and Roychoudhury [35] obtained $D = 0.01$ between Europeans and northeastern Asians and $D = 0.02$ between Europeans and Africans. Therefore, if we assume $v = 2 \times 10^4$ per locus per generation following Nei [44, p. 194], we obtain $m = (1 - 2) \times 10^4$ per generation. This is a quite small value but still seems to be too high for the migration rate among Europeans, Africans, and Asians, because in the prehistoric age with primitive culture, it should have been very difficult to move around among different continents or subcontinents.

Third, the Australian and Highland New Guinean populations, which have been isolated from Asian populations for about 50,000 years, show smaller genetic distances from Asians and Europeans than sub-Saharan African populations do. Since very little migration seems to have occurred between Australopapuans and other populations for the last 40,000 - 50,000 years, the greater distances between Africans and other populations suggest that Africans have been isolated for a longer period of time. A similar argument can be made with respect to Amerindian populations, which also have been isolated for 12,000 - 40,000 years. Goldstein et al. [38] examined the relationship between their genetic distance $(\hat{D} + \hat{v})^2$ and divergence time using information on several well-dated non-genetic evolutionary events during the last 100,000 years. Their results support the view that African populations have genetically diverged from other populations without much migration.

Fourth, if the evolution of large mammals gives any hint about the evolution of Homo sapiens, it suggests that evolution of a species with worldwide gene flow is rare. Most species of large mammals are distributed within a continent or relatively small part of it, unless they have been introduced into new territories by humans. This suggests that even within one continent various types of territorial barriers exist and these barriers generate many different species in the course of evolution. H. sapiens seem to be the only species which has a worldwide distribution and whose genetic distances among different geographical populations are very small. Since H. sapiens do not necessarily have a high degree of mobility compared with other large mammals such as tigers, lions, and wolves, the close genetic relationships of different human populations suggest that they dispersed in recent years.
Of course, if we consider human populations which are geographically closely located, there is almost always a substantial amount of gene flow. However, we are interested in tracing the evolutionary pathways of human populations by using distantly related populations from the entire world and thus avoiding the effect of local gene flow as much as possible.

Nevertheless, it should be noted that our study does not resolve all the questions concerning the controversy over the origin of *H. sapiens*. For example, the multiregional theory is based on cranial and dental characters of human remains, and it is still unclear whether the out-of-Africa theory is capable of explaining all paleontological and archeological data or not [see 10]. Therefore, it is necessary to conduct a more detailed study from various points of view to resolve this controversy.

ACKNOWLEDGEMENTS

This study was supported by research grants from the National Institutes of Health and the National Science Foundation.

REFERENCES


Table 1. Some estimates of the time of divergence between African and non-African populations (nuclear genes) or of the first split of mtDNA lineages in humans. $K =$ Kilo years.

<table>
<thead>
<tr>
<th>Study (Reference)</th>
<th>Data used</th>
<th>Estimate</th>
<th>95% C.I. (Estimate ± 2 SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nuclear genes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nei &amp; Roychoudhury, 1974 [35]</td>
<td>33 protein loci</td>
<td>117K</td>
<td>?</td>
</tr>
<tr>
<td>Goldstein et al., 1995 [38]</td>
<td>30 microsatellite loci</td>
<td>156K</td>
<td>75 - 287K</td>
</tr>
<tr>
<td>Nei, 1995 [10]</td>
<td>30 microsatellite loci</td>
<td>115K</td>
<td>56 - 213K</td>
</tr>
<tr>
<td><strong>Mitochondrial DNA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cann et al., 1986 [2]</td>
<td>RFLP data</td>
<td>200K</td>
<td>?</td>
</tr>
<tr>
<td>Pesole et al., 1992 [46]</td>
<td>D loop (650 bp)</td>
<td>400K</td>
<td>300 - 800K</td>
</tr>
<tr>
<td>Tamura &amp; Nei, 1993 [33]</td>
<td>D loop (625 bp)</td>
<td>160K</td>
<td>80 - 480K</td>
</tr>
<tr>
<td>Ruvolo et al. (1993) [47]</td>
<td>Coding genes (1,580 bp)</td>
<td>298K</td>
<td>129 - 536K</td>
</tr>
<tr>
<td>Templeton, 1993 [48]</td>
<td>D loop</td>
<td>213K</td>
<td>102 - 554K</td>
</tr>
</tbody>
</table>
Fig. 1. Phylogenetic tree for microsatellite DNA data set I. The tree was constructed by using D∞ distance. The number for each interior branch is the bootstrap value. M: Maya. K: Karitiana. S: Surui. CAR: Central African Republic. Note that the branch lengths are not proportional to evolutionary time. From ref. [11]
Fig. 2. Phylogenetic tree for microsatellite DNA data set II. The number for each interior branch is the bootstrap value. The number in parentheses for the interior branch leading to the non-African population cluster is the probability that the branch is different from zero (confidence probability). From ref. [11]
Fig. 3. Phylogenetic tree for RFLP data. CAR: Central African Republic. From ref. [11].
Figure 4. Phylogenetic tree for three major races of man based on protein polymorphism data. From ref. [11].
Fig. 5. Phylogenetic tree for Alu insertion frequency data. C-V: Chinese and Vietnamese. From ref. [11].