AMERICAN JOURNAL OF HUMAN BIOLOGY 3:83–89 (1991)

Phylogenies and the Forces of Evolution

FRANK B. LIVINGSTONE
Department of Anthropology, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT The construction of phylogenetic trees from gene frequency data assumes that a history of binary fissioning of populations has been the major cause of genetic variation. However, in many areas of the world human populations have been relatively stable with local gene flow. This population history is closer to an isolation by distance model. It was modelled by a simulation of gene frequency changes in a linear sequence of 50 stable populations with gene flow among neighboring populations. Phylogenetic trees were constructed from the gene frequencies after the simulation was run for 500 generations. Using only a few loci there is little correlation between genetic and geographic distance, but with 40 or more loci, there was a perfect correlation with geographic distance. A different population model can thus result in a phylogenetic tree comparable to those assumed to be produced by binary fission.

In the last few years there has been a continuing and literally explosive increase in knowledge of genetic variation among human populations. As would be expected, the understanding or explanation of this variation has lagged behind the pace of discovery. One reason is that the analysis of such variation in terms of the forces of evolution has not been a major concern; instead most efforts are attempts to use the genetic variation to reconstruct human demographic history, primarily by the construction of phylogenetic trees. The tree for a single locus or closely linked loci frequently results in a phylogeny that is obviously wrong, but the current view is that if enough loci are used, such errors will even out and so produce a "true" tree. Cavalli-Sforza et al. (1988) is the most comprehensive recent attempt that uses most known genetic variation to construct a phylogenetic tree for the human species. Commenting on this work, Gould (1989: pp. 22–23) has asserted, "The best way to work past these difficulties lies in a 'brute force' approach: the greater the quantity of measured differences, the greater the likelihood of a primary correlation between time and overall distance." Gould's brute force approach does have a theoretical basis (Nei et al., 1983), but other models of population history in addition to phylogenetic trees can result in a similar pattern of genetic variation.

The basic assumption of this approach is that genetic change is primarily due to neutral evolution. All the differences in gene frequencies that are used to construct the phylogeny are assumed to be due to genetic drift and possibly gene flow. Nei et al. (1983) have shown that one or a few loci will give an inaccurate estimate, and the error decreases gradually, so that with about 50 loci the estimate approaches the actual populational phylogeny with minimum error. More loci do not increase the accuracy very much. The same problem exists for the construction of species trees (Pamilo and Nei, 1988). The human data used by Cavalli-Sforza et al. (1988) with 42 loci and 120 alleles are approaching this necessary amount, and it is interesting to note that the original phylogenetic tree constructed by Cavalli-Sforza et al. (1964) on much less data was very different, with Europeans more closely linked to Africans than to Asians. It is also noteworthy that neither study used the great amount of data known for the hemoglobin and G6PD loci because of the known selection operating at these loci. However, the A1A2BO, FY, HLAA, HLAB, and El loci are also known to have some selection operating on them and are used in their most recent tree construction. Other recent studies using hemoglobin restriction site polymorphisms to construct phylogenetic trees (Wainscoat et al., 1986; Long et al., 1990) make the same assumption of neutrality.

In addition to neutral evolution, phyloge-

Received May 1, 1990, revision received November 14, 1990; accepted November 19, 1990.

netic tree construction also assumes that the differences are caused by population fission. There have been many reports that find strong correlations between linguistic, geographic, and genetic differences; the detailed analysis of Barrantes et al. (1990) of several central American tribes is the latest on a small region, while Sokal (1988) has reported on correlations throughout Europe. These correlations presumably result from the same causes of differentiation of language and genes. Cavall-Sforza et al. (1990: p. 18) have stated, "... the two evolutions follow in principal the same history, namely sequence of fissions. Two populations that have separated begin a process of differentiation of both genes and language." Nevertheless, other processes can result in the differentiation of populations. Phylogenetic trees are only one way of describing differences among a set of populations or entities, and geographic variation can result from many other forces and very different demographic histories. The purpose of this paper is to show that other processes or a very different demographic history of these populations will result in phylogenetic trees that have an almost perfect correlation with geography and thus seem to be equally plausible explanations of genetic variation among humans.

Clinal variation can be due to other forces, especially natural selection, and will produce a "good" phylogeny that seems to accord with other data such as linguistic or geographical variation providing the cline is monotonically increasing or decreasing. Variation in hemoglobin S and blood group A genes in Liberia is shown in Figure 1. The phylogeny based on the hemoglobin S cline is shown in Figure 2. I have not found any cognate data to measure linguistic differences, but there appears to be some clustering of closely related languages having similar gene frequencies. The correlation with geographical distance is .55. The most reasonable explanation for this cline is the wave of advance of an advantageous gene and demonstrates that clines that are due to other forces do produce good phylogenies. If the blood group A data are added, the resulting tree is a poorer fit with language and geography (Fig. 3), and adding five more blood group alleles does not improve the tree very much. A combination of many loci would increase the probability of including other loci with variation approximating a monotonic cline, and these would seem to account for the convergence to a perfect tree.

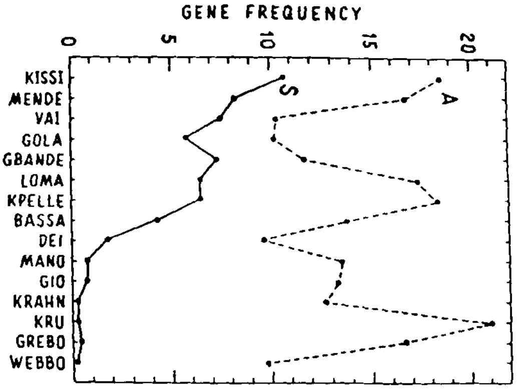


Fig. 1. The clines for the hemoglobin S and blood group A genes in Liberia.

The analysis of most clines whether they are stable or advancing assumes that the populations are reasonably stable and population displacement or expansion are not occurring, although a cline could also be due to population expansion. If we assume that the human species has occupied much of the area of the Old World for a long time and most gene flow has been among neighboring populations, then phylogenetic trees can be constructed even in the absence of population fission. This is essentially the isolation by distance model (Cavalli-Sforza and Bodmer, 1971), where most of the differentiation of the populations is due to gene drift and restricted by local gene flow. To model this, a linear series of 50 populations were all begun with the same gene frequencies of .5 for two alleles. After 500 or 1,000 generations the variation or cline was recorded. Using the same initial frequencies in all populations is comparable to assuming that the region was occupied by a rapid population expansion in a relatively short time and then has remained relatively stable. Thus, the genetic variation has been produced by isolation by distance. Migration rates of .05 with adjacent populations, .01 with the two populations adjacent to the adjacent ones, and .01 with a population randomly drawn from five on either side were assumed to be a reasonable approximation to early humans. The rate of .05 had been used by others (Weiss and Maruyama, 1976; Rouhani, 1989), but I have found that occasional long-distance mi-

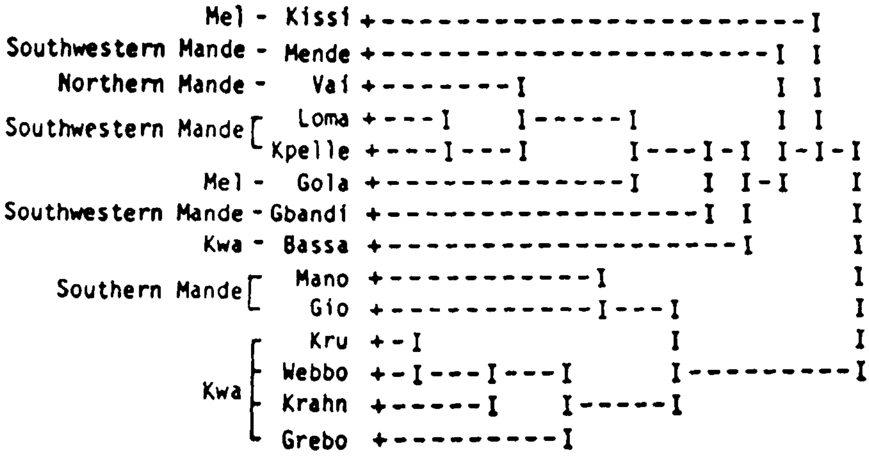


Fig. 2. The phylogenetic tree constructed from the hemoglobin S gene frequencies.

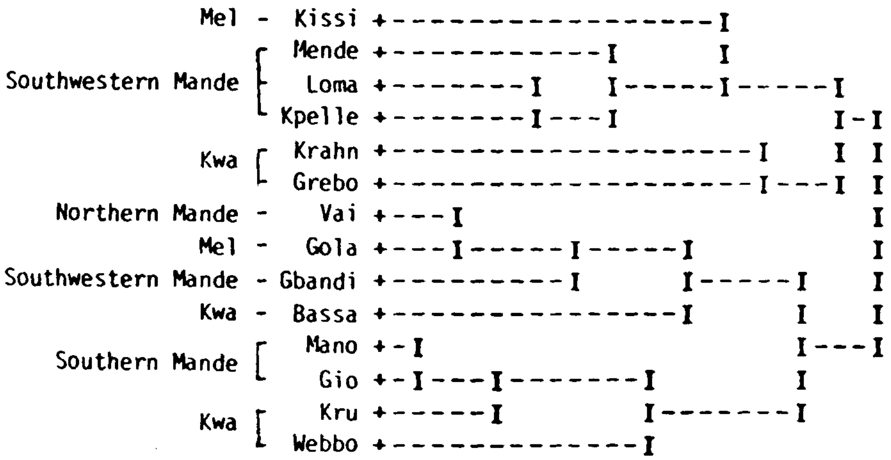
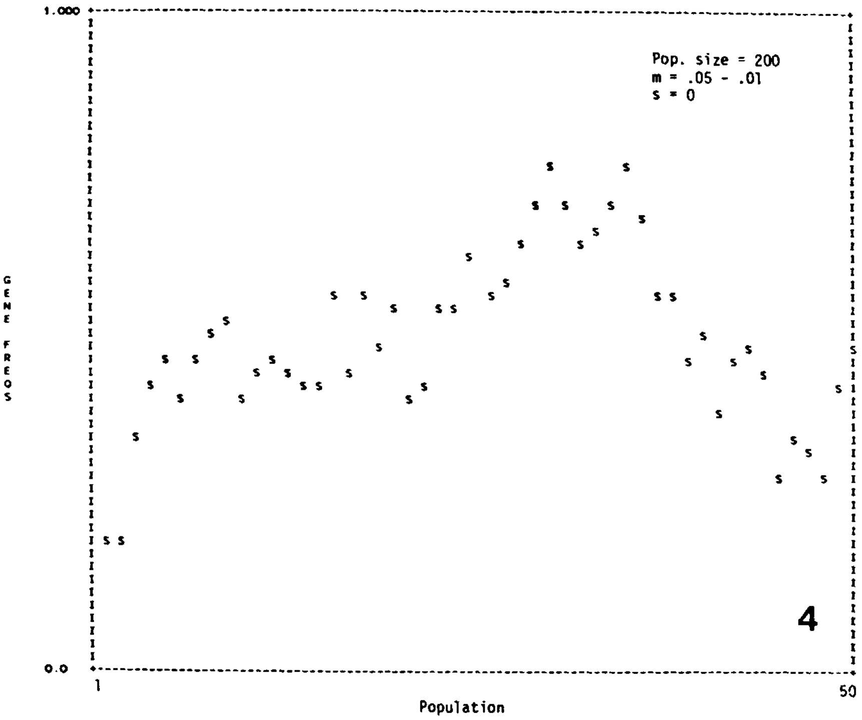


Fig. 3. The tree constructed from the hemoglobin S and blood groups A frequencies.

grants can have an important effect on gene diffusion (Livingstone, 1989). Each population was programmed to have a size of 200 individuals or 400 genes in every generation, although the small amount of selection programmed on some runs reduced the population each generation and other population sizes were used, Figures 4 and 5 are two examples of the clines that form after 500 generations with no selection or mutation, so that the variation is all due to gene drift that is opposed by gene flow. Fixation of one of the alleles never occurred in any population in any run, and after 1,000 generations the variability in gene frequencies was not greater than after 500 generations. Obviously any one run, which simulates one locus, does not give a tree correlated with geographic distance, and even using ten loci many populations are "out of place" as shown in Figure 6. However, with 50 loci a perfect fit with geographic distance is obtained (Fig. 7). The trees shown are with every third population of the linear sequence of 50 populations being clustered. But with every fifth or every other population clustered, the results are the same, and in all cases a perfect tree is produced by 50 loci. The cluster anal-



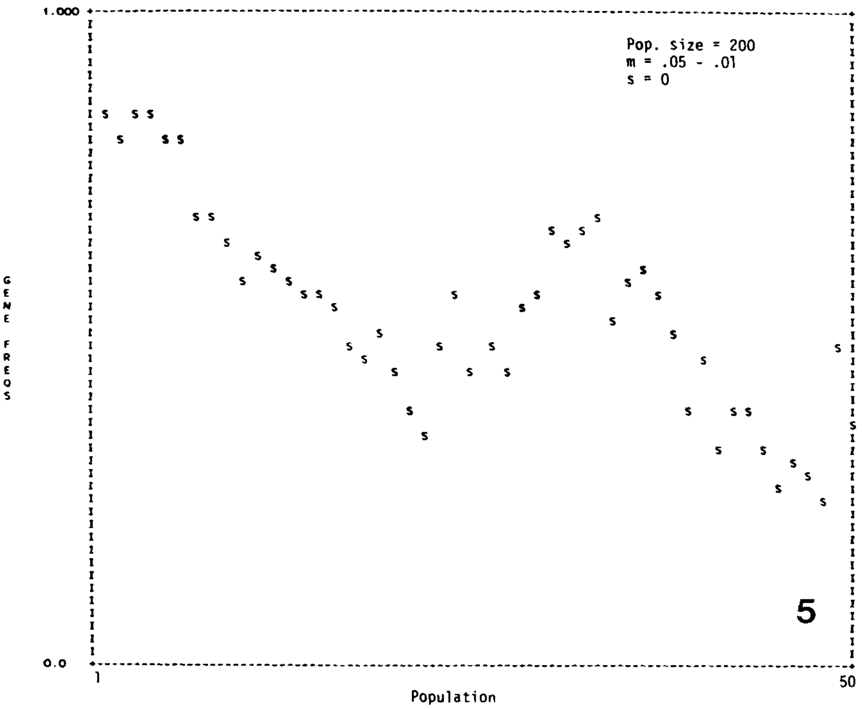
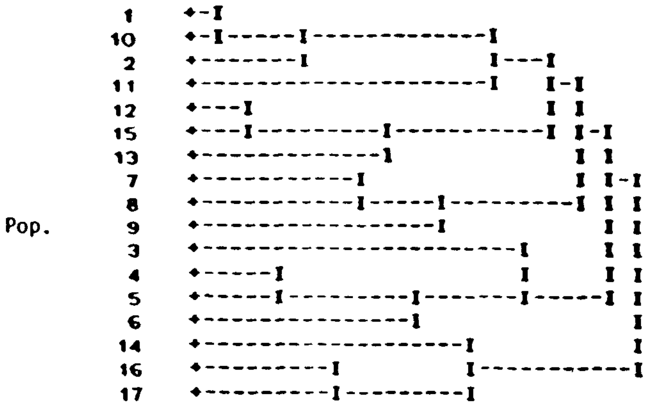
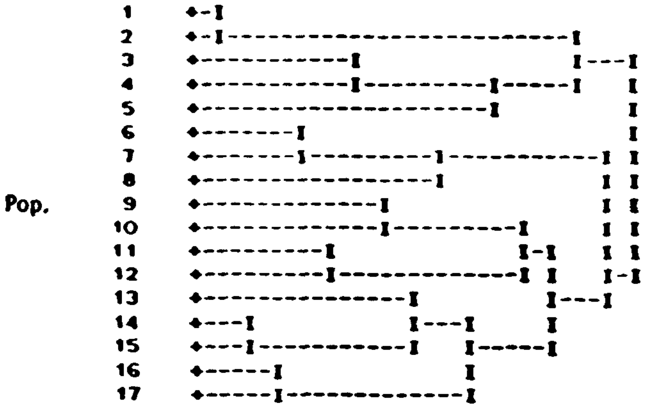


Fig. 4–5. Clines generated after 500 generations with .05 gene flow between adjacent populations, .01 with those two removed, and .01 gene flow randomly assigned to a population within 5 on either side.



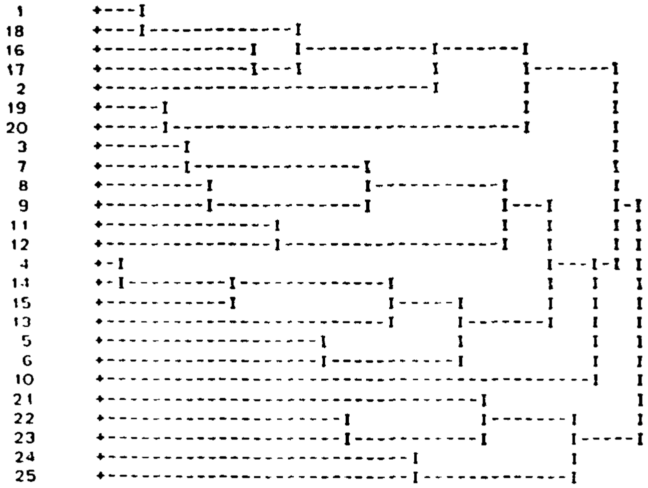
No. loci = 10
m = .05 - .01
s = 0
Pop. size = 200
No. Pops. = 50
Gen. = 500

6



No. loci = 50
m = .05 - .01
s = 0
Pop. size = 200
No. Pops. = 50
Gen. = 500

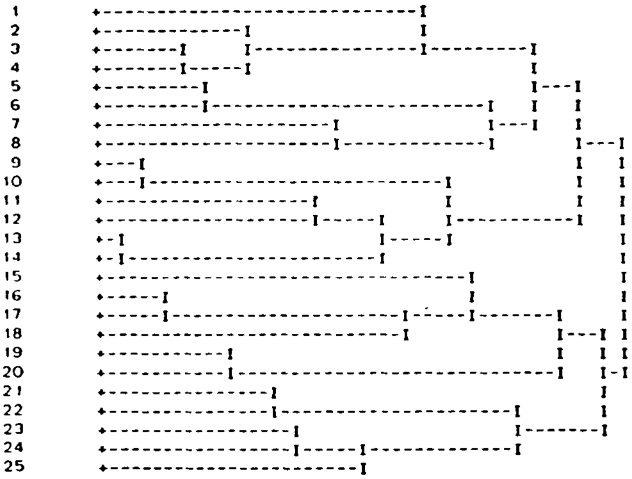
7



No. loci = 10
No. alleles = 24
m = .05 - .01
s = 0
mut = .00001
Pop. size = 200
No. Pops. = 50
Gen. = 500

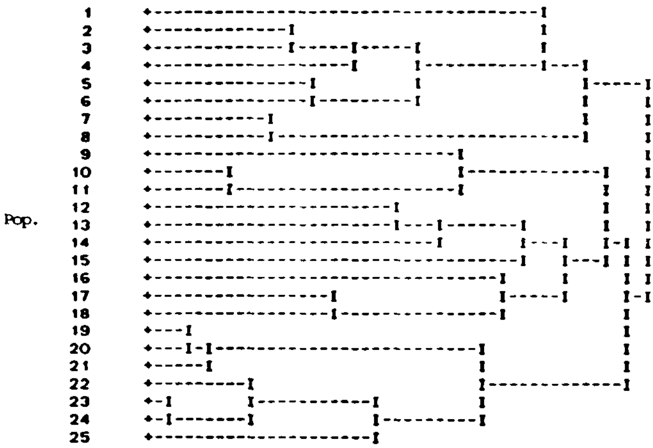
8

Fig. 6–8. See overleaf for legend.



No. loci = 60
No. alleles = 153
m = .05 - .01
s = 0
mut = .00001
Pop. size = 200
No. Pops. = 50
Gen. = 500

9



No. loci = 30
No. alleles = 169
m = .05 - .01
s = .05
mut = .00001
No. Pops. = 50
Pop. size = 200
Gen. = 500

10

Fig. 6–10. The phylogenetic trees constructed from the simulation of a linear sequence of 50 populations with various values of the parameters as indicated.

ysis used was the UPGMA options of the MIDAS program developed at the University of Michigan.

Using a smaller population size does not effect the result. If selection is programmed to act identically in all populations, then as little as .05 selection against both homozygotes seems to increase at times the number of loci needed to produce a perfect tree.

In order to determine the effects of mutation and the generation of new genetic variation, an infinite alleles model was programmed among an identical set of 50 populations with the same amount of migration. Figure 8 shows that with ten loci there will be many populations out of place, but with 60 loci there will be a perfect fit (Fig. 9). On other runs a perfect fit was obtained with as few as 40 loci. In these runs with no selection, there was an average of 3 to 4 alleles at each locus after beginning with two alleles at .5 each and a mutation rate of .00001. This model was also run for 1,000 generations and there did not seem to be any increase in the number of alleles present, and again a perfect tree was produced. Selection of .05 against homozygotes was added, but it did not change the results as shown in

Figure 10. With this model, selection of even .1 did not prevent a perfect tree but did increase the numbers of alleles present, which seems plausible since selection against homozygotes would tend to favor rare alleles.

These results are not unexpected. Similar problems and models were discussed by Felsenstein (1982), Harpending and Ward (1982), and many others. The results show that genetic variation that has a significant correlation with geography can be explained by a number of models other than binary fission. Life is not entirely binary fission. Harding and Sokal (1988) have also concluded that short-range interdemic gene flow is a major cause of the genetic and linguistic variation in Europe. Although Barrantes et al. (1990) interpret their data as due to phyletic fissioning, they also point out that the populations of Chibcha Amerindians have probably inhabited the same region for almost 10,000 years and that adjacent populations tend to resemble each other. I do not think they have excluded the model used in this paper of local migration among relatively stable populations as the cause of genetic variation and of the correlation of genetic and geographic distance.

It has seemed to me for years that the understanding of genetic variation is best approached one locus at a time, and that instead of a "brute force" and ignorance approach, an analysis of how the clines for a specific locus can be produced by the forces of evolution could lead to a better understanding of genetic variation. Among the rather small isolates in the "underdeveloped" world, the genetic variation will be due primarily to a balance of gene flow and gene drift, and a consideration of several loci would lead to a better estimate of these forces. However, the distributions of most polymorphic loci among the world's populations must surely be due in part to other forces.

LITERATURE CITED

Barrantes R, Smouse PE, Mohrenweiser HW, Gershowitz H, Azofeifa J, Arias TD, Neel JV (1990) Microevolution in Lower Central America: Genetic characterization of the Chibcha-speaking groups of Costa Rica and Panama, and a consensus taxonomy based on genetic and linguistic affinity. Am. J. Hum. Genet. *46*:63–84.

Cavalli-Sforza LL, Barrai I, Edwards AWF (1964) Analysis of human evolution under random genetic drift. Cold Spring Harbor Symp. Quant. Biol. *29*:9–20.

Cavalli-Sforza LL, Bodmer WF (1971) The Genetics of Human Populations. San Francisco: W.H. Freeman.

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. U.S.A. *85*:6002–6006.

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1990) Comment on: The feasibility of reconciling human phylogeny and the history of language. Current Anthropol. *31*:16–18.

Felsenstein J (1982) How can we infer geography and history from gene frequencies? J. Theor. Biol. *96*:9–20.

Gould SJ (1989) Grimm's greatest tale. Natural History, *2*:20–27 (Feb.).

Harding, RM, Sokal RR (1988) Classification of the European language families by genetic distance. Proc. Natl. Acad. Sci. U.S.A. *85*:9370–9372.

Harpending HC, Ward RH (1982) Chemical systematics and human populations. In Nitecki MH (ed.): Biochemical Aspects of Evolutionary Biology. Chicago: University of Chicago Press, pp. 213–256.

Livingstone FB (1989) Simulation of the diffusion of the β-globin variants in the Old World. Hum. Biol. *61*:297–310.

Long JC, Chakravarti A, Boehm CD, Antonarakis S, Kazazian HH (1990) Phylogeny of human β-globin haplotypes and its implications for recent human evolution. Am. J. Phys. Anthropol. *81*:113–130.

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

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol. Biol. Evol. *5*:568–583.

Rouhani S (1989) Molecular genetics and the pattern of human evolution: Plausible and implausible models. In Mellars P, Stringer C (eds.): The Human Revolution, Edinburgh, Edinburgh University Press, pp. 47–61.

Sokal RR (1988) Genetic, geographic, and linguistic distances in Europe. Proc. Natl. Acad. Sci. U.S.A. *85*:1722–1726.

Wainscoat JS, Hill AVS, Boyce AL, Flint J, Hernandez M, Thein SL, Olds JM, Lynch JR, Falusi AG, Weatherall DJ, Clegg JB (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature *319*:491–493.

Weiss KM, Maruyama T (1976) Archaeology, population genetics and studies of human racial ancestry. Am. J. Phys. Anthropol. *44*:31–50.