Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 1816–1819, March 1990
Population Biology

Zones of sharp genetic change in Europe are also linguistic boundaries

(human variation/gene frequencies/genetic boundaries)

GUIDO BARBUJANI*† AND ROBERT R. SOKAL†‡

*Dipartimento di Biologia, Universita di Padova, via Trieste 75, I-35121 Padova, Italy; and †Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245

Contributed by Robert R. Sokal, December 15, 1989

ABSTRACT **A newly elaborated method, "Wombling," for detecting regions of abrupt change in biological variables was applied to 63 human allele frequencies in Europe. Of the 33 gene-frequency boundaries discovered in this way, 31 are coincident with linguistic boundaries marking contiguous regions of different language families, languages, or dialects. The remaining two boundaries (through Iceland and Greece) separate descendants of different ethnic or geographical provenance but lack modern linguistic correlates. These findings support a model of genetic differentiation in Europe in which the genetic structure of the population is determined mainly by gene flow and admixture, rather than by adaptation to varying environmental conditions. Of the 33 boundaries, 27 reflect diverse population origins at often distant locations. Language affiliation of European populations plays a major role in maintaining and probably causing genetic differences.**

Genetic differences among populations are caused either by adaptive response to different selective pressures (1) or to chance—i.e., random genetic drift, founder effects, etc. (2, 3). These forces are opposed by gene flow, which tends to drive the gene frequencies of various populations toward a common equilibrium value (4, 5). A very small amount of migration—namely, one individual per generation between all pairs of population units (6)—is sufficient to prevent random divergence of gene frequencies; gene flow also will oppose local differentiation under differential selection (5, 7). Therefore, whatever the origin of genetic diversity, its maintenance depends largely on limited gene flow among populations (8).

Under genetic drift and short-range dispersal of individuals (i.e., under isolation by distance), the genetic similarity between populations decreases exponentially with their distance (9–11). Areas of abrupt change are not expected (12). Deviations from these expectations suggest rejection of the hypothesis of pure isolation by distance in favor of models involving either differential selection or limited gene flow. In humans, evidence for selection differentials exists only for polymorphisms associated with malaria resistance, such as glucose-6-phosphate dehydrogenase deficiency (13) and thalassemias (14). Therefore, whenever the rate of change is consistently increased for other gene frequencies, factors other than distance can be safely assumed to isolate populations. This would be true of equilibrium populations and even more so for nonequilibrium ones, as must be the case with present-day humans. Both geographical and cultural isolating factors can be envisaged.

Previous work (15) has shown that the rate of change in the frequency of some alleles across the boundaries between language families in Europe is higher than across comparable lines drawn at random on the map of Europe. Other studies (16–19) confirmed that genetic and linguistic variations are correlated in the geographical space. Schematically, this may have two explanations. Either (*i*) the processes leading to linguistic differentiation also brought about genetic differentiation, or (*ii*) linguistic differences act as reproductive barriers, leading spatially close populations to diverge also in gene frequencies. However, demonstrating that there is an increased rate of genetic change across language boundaries is not sufficient to correlate the two variables. It could be that other regions of increased genetic change, objectively determined in the gene-frequency surfaces, are not also regions of linguistic change. Therefore, we looked for zones of abrupt genetic change in Europe and associated them with possible causal factors: physical and linguistic barriers and the effects of historical events. If all or most zones of rapid genetic change turn out to be also linguistic boundaries, this would support a major role of language barriers in preventing population admixture. This, in turn, would argue for a causal role of language differences in maintaining genetic variation among populations.

MATERIALS AND METHODS

The data consist of 63 allele frequencies, measured at 19 genetic loci in 3119 European localities. They are described in detail elsewhere (15, 16, 18). The number of sample localities for different allele frequencies ranges from 34 to 870, with a median value of 73. Minimum sample size per locality is 50, but most samples are far larger. Each set of allele frequencies was interpolated by using inverse squared-distance weighting (20) into a quasi-continuous allele-frequency surface mapped onto a regular 124 × 74 lattice covering Europe.

Following a method originally suggested by Womble (21) and elaborated by us (ref. 22; we call it "Wombling"), we computed the partial derivatives of the allele frequencies at each lattice point with respect to the *X*- and *Y*-axes, recording both magnitude and direction of maximum slope. We averaged the magnitudes and slope angles at each lattice point of 60 allele-frequency surfaces. The resulting "systemic function" (21), in which each lattice point is represented by a vector possessing both magnitude and direction, was then plotted (Fig. 1). By analogy to hypothesis testing in statistics, we chose to consider only lattice points whose magnitudes fall in the highest 5% of their distribution. Since we wanted to recognize boundaries rather than scattered single high-magnitude vectors, we imposed a connectedness criterion among high-magnitude lattice points and also required their direction to differ by no more than 30°. Additionally, we analyzed 32 allele-frequency surfaces based on more than 66 localities. We constructed boundaries from these individual surfaces by criteria identical to those applied to the systemic function. The resulting 32 maps were of the kind shown for the systemic function in Fig. 1. All maps were examined for boundaries.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡To whom reprint requests should be addressed.

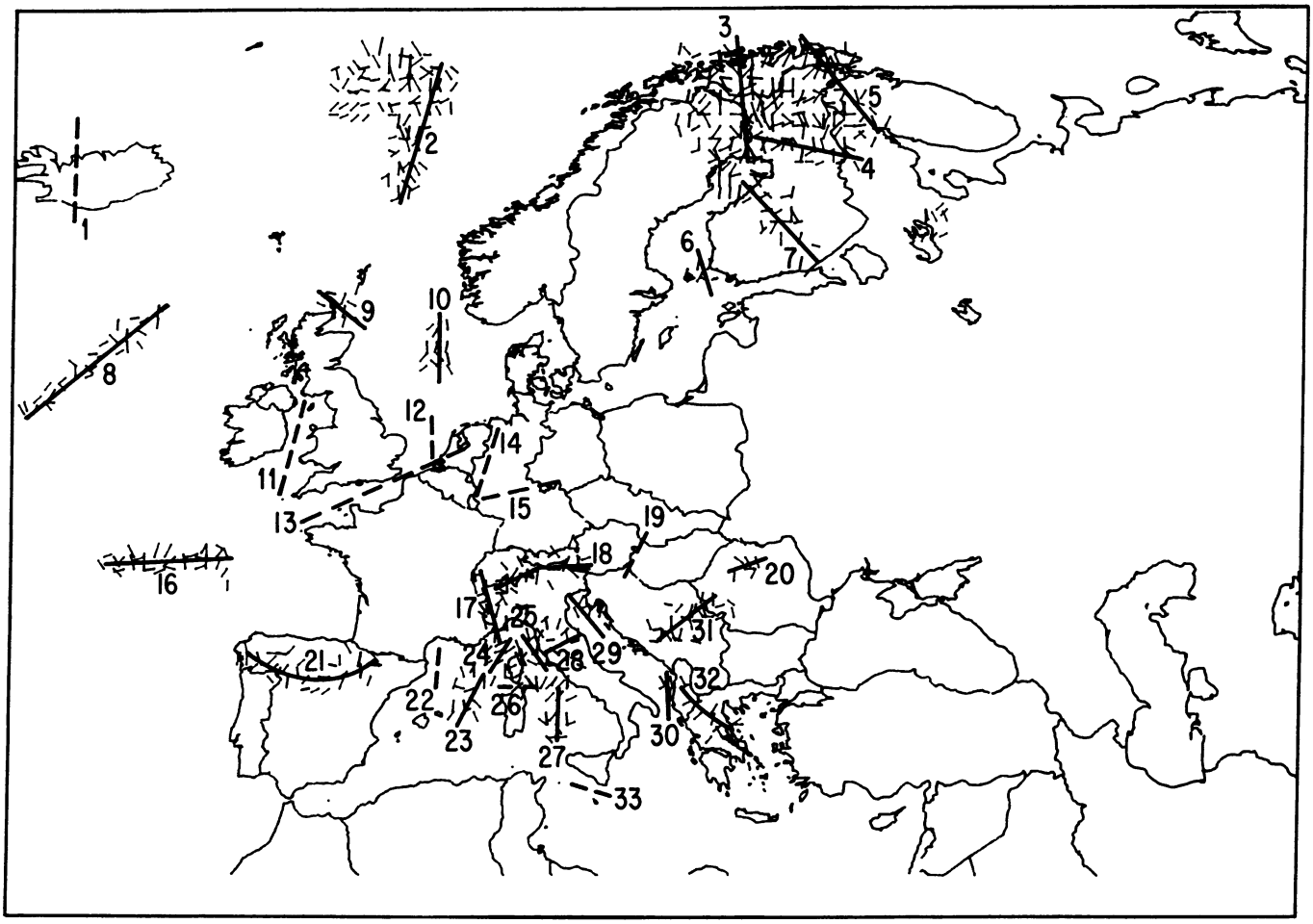

FIG. 1. Plot of the systemic function for 60 human allele frequencies in Europe. The lengths of the rods are proportional to the average magnitude of the gene-frequency change, the directions of the rods are the average directions of the maximum slope across gene-frequency surfaces. We plot only the vectors in the top 5% of the distribution of magnitudes that are connected and roughly parallel (see text) and those vectors in the top 10% connecting such top 5% values for the purpose of forming boundaries. The 33 gene-frequency boundaries recognized by the analysis are shown by thick lines. The 24 boundaries abstracting the systemic function are drawn as solid lines. To be recognized, boundaries had to have a "reasonable" length. The 9 boundaries resulting from the analysis of individual surfaces are indicated as dashed lines. In cases where the abstracted lines do not appear to reflect the exact locations of the vectors, this is because the final positioning of the line was influenced by the results of the individual surfaces. The boundaries, identified by arbitrary code numbers, are listed below. Each boundary is followed by the percentage of individual surfaces in which it is recognized. To be recognized in the analysis of individual surfaces, a given boundary had to be supported for each surface by a minimum number of sampled points to both sides of the boundary; also the stated percentage (over all acceptable surfaces) had to exceed the 95% tolerance limit of a 5% type I error rate. When this did not occur, the percentage is shown enclosed by parentheses. Brackets enclose language-family boundaries (italics) or language boundaries (roman type) that coincide with the genetic boundaries. The 33 gene-frequency boundaries: 1, western vs. eastern Iceland, 69.2%; 2, Norway vs. Iceland, 62.7% [Norwegian–Icelandic]; 3, northern Finland vs. Sweden,* 81.8% [*Finnic–Germanic*]; 4, central vs. northern Finland,* 38.5% [Finnish–Saamic (Lappic)]; 5, Kola Peninsula vs. Finland,* 80.0% [*Slavic–Finnic*]; 6, Finland vs. Sweden, across the Gulf of Bothnia, 40.0% [*Finnic–Germanic*]; 7, southwestern vs. northern and eastern Finland, 26.9%; 8, Ireland vs. Iceland, 48.4% [English–Icelandic]; 9, Scotland vs. Orkney and Shetland Islands (50.0%) [*Celtic–Germanic*]; 10, southern Scandinavia vs. Scotland, 28.0% [Norwegian, Danish–English]; 11, England and Wales vs. Ireland, 30.8% [*Celtic–Germanic*, in part]; 12, England vs. The Netherlands, 23.3% [English–Dutch]; 13, England vs. France and Belgium, across the English Channel, 20.7% [*Germanic–Romance*]; 14, Germany vs. The Netherlands, 19.4% [German–Dutch]; 15, central vs. southern Germany, 20.7%; 16, Iberia vs. Iceland, 22.6% [*Romance–Germanic*]; 17, France vs. Italy, 26.7% [French–Italian]; 18, Switzerland and Austria vs. Italy, 40.6% [*Germanic–Romance*]; 19, Austria vs. Hungary, 24.0% [*Germanic–Ugric*]; 20, through Transylvania, (0.0%) [*Romance–Ugric*]; 21, northwestern Iberia vs. the rest (21.4%) [*Basque–Romance*, in part]; 22, Catalonia vs. Corsica, 36.4% [Catalan–Italian]; 23, Balearic Islands vs. Sardinia, 37.5% [Spanish–Sardinian]; 24, Corsica vs. France, 41.4% [Italian–French]; 25, Corsica vs. Italy, 38.5%; 26, Corsica vs. Sardinia, (15.4%) [Italian–Sardinian]; 27, Sardinia vs. Italy, 32.1% [Sardinian–Italian]; 28, northern vs. southern Italy, 40.0%; 29, Italy vs. Yugoslavia, across the Adriatic, 41.9% [*Romance–Slavic*]; 30, Italy vs. Albania, (15.4%) [*Romance-Albanian*]; 31, northwestern vs. southeastern Yugoslavia, (13.7%); 32, northern vs. central and southern Greece (21.4%); 33, Sicily vs. Malta, 66.7% [*Romance–Semitic*].

*In the systemic function (see Fig. 1), this gene-frequency boundary cannot be recognized as such but is noticed as a region of general differentiation of North Scandinavia from regions to the south and east. Because various individual surfaces recognize distinct boundaries 3, 4, and 5, they are listed separately here.

In the systemic function of Fig. 1, we recognized 24 boundaries, abstracted as superimposed solid lines. The 32 individual maps yielded all but one of these boundaries plus 71 additional ones, which we reduced to 9 shown as broken lines in Fig. 1. They were chosen by criteria described under that figure.

RESULTS

Examination in Fig. 1 of the 33 recognized gene-frequency boundaries (24 from the systemic function, 9 from the individual surfaces) reveals that they are not a haphazard collection of lines across the map of Europe. At least 22 are

distinct physical barriers (19 oceanic and 3 montane). However, even more map into well-known language boundaries. Of the 33, 15 represent all or part of the boundaries between modern language families in Europe (23) (listed beneath Fig. 1). Another 11 gene-frequency boundaries represent boundaries between different languages (23) within a language family (also listed).

Among the remaining seven boundaries, boundary 25 between Corsica and Italy is a dialect or language boundary, depending on the rank assigned to Corsican, a distant member of the Tuscan subdivision of Italian (24). Boundary 7 between southwestern Finland and northern and eastern Finland approximately coincides with the line separating the western from the eastern dialect of Finnish. The former is spoken by the descendants of the Suomaläiset and the Hämäläiset (the Tavastians), both of whom came to Finland via Estonia, crossing the Gulf of Finland. The eastern dialect is spoken by descendants of Karelian tribes that came into modern Finland overland from the southeast (25). Even though the language boundary between Finnish and Karelian currently runs farther to the east and in a north–south direction, this ancient boundary continues to be marked by a sharp gene-frequency change.

Boundary 15 approximates the border between speakers of Low German and those of Middle and High German (24). Boundary 28 cuts across the area of the central Italian dialects, roughly separating speakers of northern Italian dialects from those of southern dialects (24, 26).

Boundary 31 separates northern and western Yugoslavia from the southeastern part of the country. The area northwest of the boundary had been settled by Avaroslavs affiliated with the major group of the Sclavini, coming from Moravia and the upper courses of the Elbe, Oder, and Vistula rivers. The area southeast of the boundary was settled in the 7th century by Slavic tribes belonging to the major group of the Antes, coming from modern Romania.§ These ancient settlement patterns are supported by linguistic differences. The boundary approximates the division between Old and New Štokavian dialects of Serbo-Croat (24).

The remaining two boundaries have no obvious modern linguistic correlates. Boundary 1 between western and eastern Iceland can be substantiated by the ethnic settlement pattern of the island. The west of the island was preferentially settled by persons reaching Iceland via Ireland or Scotland (27), where they intermarried with the native Celtic-speaking population, bringing Irish wives and servants with them. Using the map furnished in ref. 27, we tested the counts of settlers who either (*i*) were Irish or Scots or Vikings who came via Ireland or (*ii*) were those Vikings who came directly from Norway. West of the boundary found by us, the percentage of settlers affiliated with the British Isles is 23.48%; east of the boundary, it is only 10.23%. The difference is highly significant ($G_{adj} = 13.179$; $P < 0.001$). The incidence of Icelandic toponyms of Irish origin is also greater in western Iceland. Thus, differences in origins of the Icelandic populations dating back 1000 years are reflected in modern gene frequencies. Finally, boundary 32 separates northern from central Greece but is somewhat north of the generally accepted limits of the northern Greek dialects (28). However, this line approximately demarcates the southern boundary of an area settled by relocated Greek-speaking populations from Asia Minor during the population exchange with Turkey following World War I (29).

In summary, 31 of the 33 recognized genetic boundaries are also linguistic boundaries. Twenty-two of these are also obvious physical boundaries (4 montane and 18 marine). The remaining 9 genetic boundaries (boundary numbers 3, 4, 5, 7, 14, 15, 19, 28, and 31) are also linguistic boundaries but do not correspond to evident physical barriers. If we examine the history of the 33 boundaries studied, 27 of them mark zones of contact between different ethnic groups that originated elsewhere often at great distances from each other.

DISCUSSION

The robustness of these results is shown by the substantial correspondence between the systemic function of Fig. 1 and the individual analyses (see the percentages listed beneath the figure). This correspondence appears even though the numbers and locations of the samples differ appreciably among the 19 loci. The congruence of the discovered boundaries with linguistic boundaries is far too strong to be due to chance. The position, orientation, and shape of the boundaries are not constrained by the Wombling method. Thus, their coincidence with lines defined by other criteria is striking.

This study shows that the zones of abrupt genetic change in European human populations correspond with only two exceptions to two kinds of obstacles to population admixture: geographical barriers and language boundaries. The association of rapid genetic change with these obstacles, and particularly with linguistic change in 31 of the 33 boundaries, strongly suggests that the genetic variation observed has little to do with adaptation to local environments. More likely, it reflects the diverse origin of populations that often evolved elsewhere and came into contact through various migration processes. Similar phenomena of interaction between random genetic differentiation on the one hand and migration followed by incomplete admixture on the other have been proposed to account for the modes of genetic variation described in other human populations, both in smaller areas, such as Italy (30), and on a continental scale (19, 31). The overall picture emerging is one in which adaptive differences play a minor role, although they have probably determined a few wide clines (32–34) that are not expected to be blurred by the effects of genetic drift (35). Demographic phenomena, such as individual dispersal, population displacement, breaking of isolates, and admixture, contribute to smoothing the gene-frequency distributions. But when they are constrained by physical or cultural factors, or both, then genetic differences persist across time and are large enough to be detected even in a limited subset of the loci of the human genome, such as those considered in the present study.

Both theoretical considerations (36, 37) and simulation studies (38, 39) suggest that the genetic structure of natural populations, as described by large-scale studies based on protein polymorphisms, reflects mainly past patterns of gene flow and other demographic episodes. In humans, the observation that populations differing in language differ also genetically, above and beyond the differences induced by spatial separation (15–18, 40) is in contrast with the expectations based on models of isolation by distance (41). In particular, the existence of regions where several allele frequencies vary abruptly and simultaneously (such as those detected by Wombling) indicates that additional factors isolate population units, leading to patterns and rates of migration that do not simply reflect the geographical distances. In 9 of the 11 genetic boundaries detected by us that are not associated with physical barriers, language barriers may oppose the process of population admixture. At other boundaries language differences may reinforce effects of physical barriers. Given the limited amount of gene flow even in modern times across these boundaries, it would not appear that these differences are likely to disappear in the foreseeable future.

§Vlachovic, P. (1979) *Ethnology of Yugoslavia: Ethnogenesis of the Yugoslav Peoples*, General Part 1 (mimeographed lecture notes in Serbo–Croat, Belgrade).

We thank Geoffrey Jacquez and Barbara Thomson for computational assistance and Donna DiGiovanni, Marie-Josee Fortin, and Chester Wilson for technical help. Prof. L. L. Cavalli-Sforza and Dr. Neal Oden furnished useful comments. This research was supported by Grant GM28262 from the National Institutes of Health to R.R.S. This article is contribution number 742 in ecology and evolution from the State University of New York at Stony Brook.

1. Hedrick, P. W. (1986) *Annu. Rev. Ecol. Syst.* **17,** 535–566.
2. Cavalli-Sforza, L. L. (1966) *Proc. R. Soc. London Ser. B* **164,** 362–379.
3. Wright, S. (1969) *Evolution and the Genetics of Populations: The Theory of Gene Frequencies* (Univ. of Chicago Press, Chicago), Vol. 2.
4. Slatkin, M. (1985) *Annu. Rev. Ecol. Syst.* **16,** 393–430.
5. Slatkin, M. (1987) *Science* **236,** 787–792.
6. Wright, S. (1931) *Genetics* **16,** 97–159.
7. May, R. M., Endler, J. A. & McMurtrie, R. E. (1975) *Am. Nat.* **109,** 659–676.
8. Endler, J. A. (1977) *Geographic Variation, Selection, and Clines* (Princeton Univ. Press, Princeton).
9. Kimura, M. & Weiss, G. H. (1964) *Genetics* **49,** 561–576.
10. Morton, N. E., Yee, S., Harris, D. E. & Law, R. (1971) *Theor. Pop. Biol.* **2,** 507–524.
11. Barbujani, G. (1987) *Genetics* **117,** 777–782.
12. Sokal, R. R. & Wartenberg, D. E. (1983) *Genetics* **105,** 219–237.
13. Livingstone, F. B. (1971) *Annu. Rev. Genet.* **5,** 33–64.
14. Silvestroni, E. & Bianco, I. (1975) *Am. J. Hum. Genet.* **27,** 198–212.
15. Sokal, R. R., Oden, N. L. & Thomson, B. A. (1988) *Am. J. Phys. Anthropol.* **76,** 337–361.
16. Sokal, R. R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 1722–1726.
17. Sokal, R. R., Oden, N. L., Legendre, P., Fortin, M.-J., Kim, J. & Vaudor, A. (1989) *Am. J. Phys. Anthropol.* **79,** 489–502.
18. Sokal, R. R., Oden, N. L., Legendre, P., Fortin, M.-J., Kim, J., Thomson, B. A., Vaudor, A., Harding, R. M. & Barbujani, G. (1990) *Am. Nat.*, in press.
19. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, A. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6002–6006.
20. Dougenik, J. A. & Sheehan, D. E. (1979) *SYMAP User's Reference Manual* (Camera Statistics, Bedford, MA), Version 5.20.
21. Womble, W. H. (1951) *Science* **114,** 315–322.
22. Barbujani, G., Oden, N. L. & Sokal, R. R. (1989) *Syst. Zool.*, in press.
23. Ruhlen, M. (1987) *A Guide to the World's Languages: Classification* (Stanford Univ. Press, Stanford, CA), Vol. 1.
24. Comrie, B., ed. (1987) *The World's Major Languages* (Oxford Univ. Press, New York).
25. Jutkilla, E. (1962) *A History of Finland* (Praeger, New York).
26. Pellegrini, G. D. (1977) *Carta dei Dialetti d'Italia* (Pacini, Pisa, Italy).
27. Pálsson, J. (1976) in *Rassengeschichte der Menschheit*, ed. Schwidetzky, I. (Oldenbourg, Munich), Vol. 4, pp. 147–155.
28. Lejeune, M. & Newton, B. E. (1976) *The New Encyclopaedia Britannica (Macropaedia)* (Encyclopaedia Britannica, Chicago), Vol. 8, pp. 392–398.
29. Paidoussis, M. & Krimbas, C. B. (1980) in *Physical Anthropology in European Populations*, eds. Schwidetsky, I., Chiarelli, B. & Necrasov, O. (Mouton, The Hague, The Netherlands), pp. 145–170.
30. Piazza, A., Cappello, N., Olivetti, E. & Rendine, S. (1988) *Ann. Hum. Genet.* **52,** 203–213.
31. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. L. (1978) *Science* **201,** 786–792.
32. Hedrick, P. W. & Thomson, G. (1983) *Genetics* **104,** 449–456.
33. Hedrick, P. W., Thomson, G. & Klitz, W. (1986) in *Evolutionary Processes and Theories*, eds. Karlin, S. & Nevo, E. (Academic, Orlando, FL), pp. 583–606.
34. Barbujani, G. (1988) *Ann. Hum. Genet.* **52,** 215–225.
35. Slatkin, M. & Maruyama, T. (1975) *Genetics* **81,** 209–222.
36. Barton, N. H. & Hewitt, G. M. (1985) *Annu. Rev. Ecol. Syst.* **16,** 113–148.
37. Slatkin, M. (1989) *Genome* **31,** 196–202.
38. Rendine, S., Piazza, A. & Cavalli-Sforza, L. L. (1986) *Am. Nat.* **128,** 681–706.
39. Sokal, R. R., Jacquez, G. M. & Wooten, M. (1989) *Genetics* **121,** 845–855.
40. Harding, R. M. & Sokal, R. R. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 9370–9372.
41. Sokal, R. R., Harding, R. M. & Oden, N. L. (1989) *Am. J. Phys. Anthropol.* **80,** 267–294.