Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 5620–5624, June 1992
Evolution

Coevolution of genes and languages revisited

L. L. CAVALLI-SFORZA, ERIC MINCH, AND J. L. MOUNTAIN

Department of Genetics, Stanford University, Stanford, CA 94305

Contributed by L. L. Cavalli-Sforza, March 6, 1992

ABSTRACT **In an earlier paper it was shown that linguistic families of languages spoken by a set of 38 populations associate rather strongly with an evolutionary tree of the same populations derived from genetic data. While the correlation was clearly high, there was no evaluation of statistical significance; no such test was available at the time. This gap has now been filled by adapting to this aim a procedure based on the consistency index, and the level of significance is found to be much stronger than 10^{-3}. Possible reasons for coevolution of strictly genetic characters and the strictly cultural linguistic system are discussed briefly. Results of this global analysis are compared with those obtained in independent local analyses.**

In chapter 14 of *Origin of Species*, Darwin predicted that if one could reconstruct the tree of human evolution one would have the best classification of human languages. One and a half centuries after that prophecy, we still do not have an evolutionary tree for languages. In fact, linguists do not agree upon whether there was a single origin for human languages. Honoring a tradition that goes back to a formal decision against research on the origin of language, made by the Linguistic Society of Paris in 1866, many linguists avoid the issue. The alleged difficulty is that the rate of linguistic change is so high and the origin of human language so remote that very little, if anything, is still shared by languages that diverged long ago. Recently, however, some linguists (ref. 1; M. Ruhlen, personal communication) have found commonalities among many living languages that indicate possible avenues of attack on the problem of a single origin of languages.

In an earlier paper (3) it was shown that an evolutionary tree of a set of human populations representing the whole world, reconstructed entirely from genetic data, was sufficiently similar to a linguistic classification to suggest coevolution of languages and genes. There is a rationale for expecting coevolution: events responsible for genetic differentiation are very likely to determine linguistic diversification as well. For instance, the physical separation of two or more populations occurring after fission of an initially single population, with migration of one or more splinters to remote geographic areas, is likely to reduce or eliminate further contacts among them. Because such reduced contact contributes to both genetic and linguistic divergence between the splinter populations, both types of divergence tend to increase with the passing of time. To the extent that an evolutionary tree reflects the history of separations in time, the linguistic and genetic evolutionary trees must be parallel. Complete separation is unnecessary for linguistic and genetic differentiation. It is well known that geographic distance between populations is a good predictor of decreasing intermigration and, therefore, also of increasing genetic diversity (4–8). The same is largely true of linguistic diversity (9, 10).

The considerations above predict that a correlation between genetic and linguistic evolution should exist, but its strength will depend on the nature of mechanisms of transmission from generation to generation in the two phenomena. Genes are clearly transmitted from parents to offspring, without exception. The transmission of languages is less clear cut, but in many societies parents, and especially the mother, play a dominant role. Languages, however, can be learned at all ages, even if learning may be imperfect at later ages, especially after puberty. Therefore, transmitters other than parents—from age peers to school teachers and, in very recent times, mass media—also have an influence. These types of cultural transmission, called *horizontal*, complement the *vertical* one (from parents to offspring) (11). They play a major role in modern society, whereas in traditional tribal societies they are likely to be of lesser importance (12). Horizontal transmission of language may even determine total or near total replacement of a language, usually by that of foreign conquerors. History records several examples of language replacements. In Europe, the replacement of Celtic languages by Latin in the territories of the Roman Empire and of Latin by Anglo-Saxon in England are well documented. Colonialism led to the spread of four European languages (English, French, Spanish, and Portuguese) to all or almost all continents.

Language replacement can thus reduce and in certain areas completely eliminate linguistic and genetic correlation. Gene replacement may have the same effect, especially when gene flow into a population from a neighboring one continues for many generations, without replacement of language. An imperfect correlation between genetics and language is therefore expected. Observations to be discussed in this paper show that even if the observed correlations are imperfect, they are frequently high.

In the case of the correlation observed in the earlier paper, no significance test was available, and none was given: the results seemed to speak for themselves (3). Nevertheless, it is preferable, and customary when possible, to test the statistical significance of conclusions. The aim of this paper is to test, using a Monte Carlo approach, whether the observed correlation between genetic and linguistic evolutionary history is significantly different from zero.

Comparison of Linguistic Families and Genetic Tree of Human Evolution

A genetic tree based on 42 populations was published earlier (3), and a similar one is reproduced here in Fig. 1 with trivial modifications. In both the original tree and the present one, only 38 populations are listed: all 5 European populations used for the original 42 populations distance matrix were so genetically similar, at the scale of distances used, that they were pooled into one.

As mentioned previously, there is no comprehensive linguistic tree. There are, however, various groupings into linguistic families or phyla, which number 17 in the most recent and complete classification (13, 14). One of these linguistic families, Amerind, was recently proposed by

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.

Abbreviation: CI, consistency index.

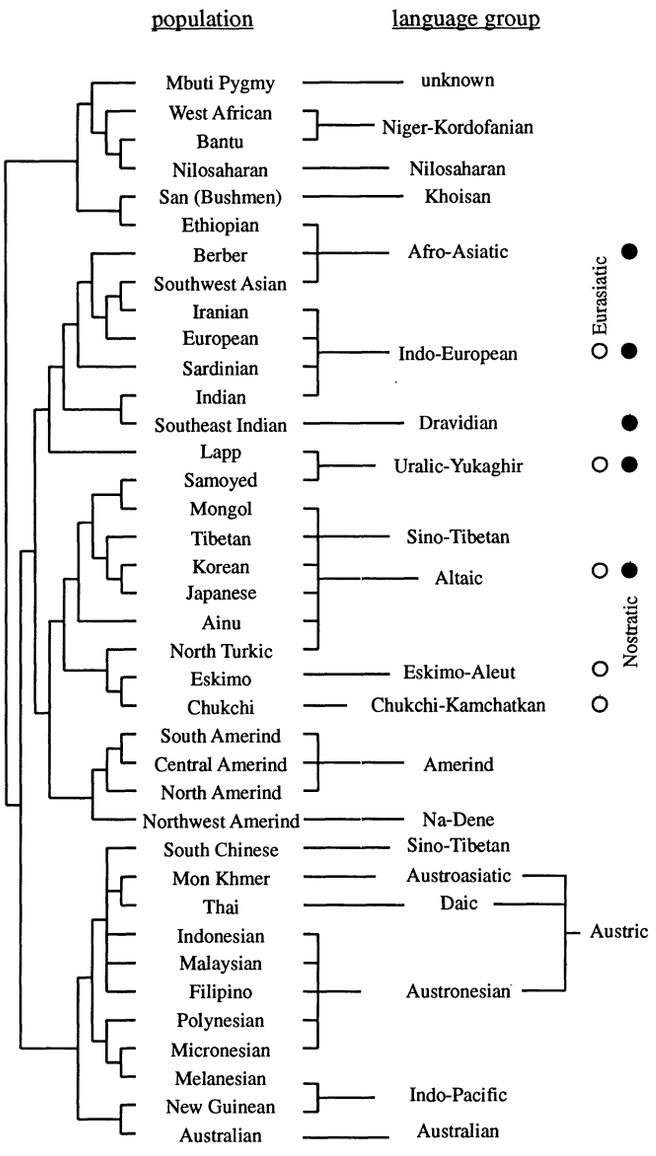

FIG. 1. A simplified version of the tree shown in ref. 1; only the topology of the genetic tree is retained (i.e., the branch lengths do not signify time since separation). The Eurasiatic and Nostratic family groupings are indicated, respectively, by open and filled circles.

Greenberg (1), who used a procedure known as multilateral comparison. His conclusions have been disputed by a group of American linguists who strongly object to clustering all but Eskimo and Na-Dene American languages into a small number of families (1, 14). They follow an entirely different approach, usually comparing no more than two languages at a time and deciding whether the two languages are related or unrelated, without indicating the degree of relationship. This approach is probably the reason (1) why they are unable to classify American Indian languages into fewer than 58 or 60 families (15), a very large number in comparison with the rest of the world. Further comments and references to this controversy can be found in ref. 14.

One of Ruhlen's 17 families, Caucasian, is now considered to consist of two distinct taxa: (North) Caucasian and Kartvelian (14). Neither group is sufficiently represented by genetic data to be included in Fig. 1. There are thus 16 linguistic families in Fig. 1 of this paper. In the original figure, there were also two overlapping groupings, which cluster a number of these 16 families, providing the starting point for a tree of linguistic evolution. These groupings are ignored in the first step of the present analysis, in which we limit our attention to the association observed between the 17 linguistic groups (16 families and an African Pygmy population that is believed to speak a borrowed language, their original one having been replaced) and the genetic classification of the 38 populations suggested by the tree. The linguistic families spoken by the 38 populations are indicated in Fig. 1. It is clear from inspection of Fig. 1 that populations speaking languages from the same family tend to be genetically related, suggesting that there is a strong correlation. There are, however, some exceptions; Berbers and the great majority of Ethiopians, for example, speak languages from the same family (Afro-Asiatic), but associate with two different genetic clusters (Berbers with Caucasoids and Ethiopians with Africans). A list of exceptions was given in the earlier paper (3). The question of whether a specific exception is due to language or gene replacement (or both) can sometimes be addressed (see examples in ref. 16), but, in general, historical information is necessary for a satisfactory solution.

One can assess the degree of an association between linguistic and genetic evolution by comparing the linguistic classification of the 38 populations with their genetic tree using the consistency index (CI). In the present application, the index estimates the number of changes that must be postulated during the evolutionary process, assuming the genetic tree is entirely correct, to obtain the observed distribution of 16 linguistic families among the 37 populations. The CI was proposed by Kluge and Farris (17) and is defined as the ratio between the number of states of a character and the number of changes of state of that character in the tree being examined.

Use of the CI as a test of coevolution, however, requires statistical support. A study by Archie (18), with the goal of determining "if sets of data used for phylogenetic analysis contain phylogenetically non-random information," indicates important weaknesses. Archie analyzed 28 published data sets on 5–44 taxa with 9–92 quantitative morphological or qualitative multistate characters. For each data set he found the minimum length tree (i.e., the one requiring the smallest number of character changes) and used it to calculate the CI. These values were then compared with estimates of the corresponding chance values derived from a large number of Monte Carlo permutation experiments. Archie found that the CI is very sensitive to the number of populations, decreasing on average with increasing numbers of taxa (see also refs. 19 and 20). The index alone, therefore, is a poor test of a phylogenetic hypothesis.

The CI was first applied to our data by O'Grady *et al.* (21) (see also ref. 22). They concluded that "only 48% [the CI value they calculated] of the race–language association supports a conclusion of development and retention of a language within a racial lineage." Since, however, the CI is poorly suited as a direct estimate of the degree of association, we suggested that it might at least be employed in a significance test of the association (23).

The same permutation procedure that Archie used to obtain the expected value of the CI can be extended to generate the entire random distribution of CI values. One can then test whether a particular observed CI value is greater than it would be by chance alone, in the absence of a correlation between a particular multistate character (the linguistic family) and a given evolutionary tree of a group of populations. We have applied the procedure to 10 subsets of our data. The first (data set I) consists of 38 populations or taxa and 17 character states (16 linguistic families and an African Pygmy population that does not speak its original language), derived from the tree in ref. 1 and also shown in figure 2 of ref. 22. The second (data set II) consists of 30 populations and 9 linguistic families, in which the autapo-

morphic (those characterized by a unique linguistic group) populations of data set I were omitted. The other eight data sets are discussed in the following section. The MIX program of the PHYLIP package (24) determined the length of each tree (i.e., the minimum number of character state changes). This value served as the denominator of the CI ratio; the numerator was the number of language groups (17 or 9). The resulting CI values were 0.594 for data set I and 0.474 for data set II.

We ran 10,000 replications of these analyses, using the original genetic tree but randomly permuting the character states (linguistic families) to estimate the chance level CI. The mean CI of these random replications was 0.454 for data set I and 0.309 for data set II, with standard deviations of 0.0191 and 0.0180, respectively. The observed genetic–linguistic CI values of 0.594 and 0.474 are thus 7.33 and 9.15 standard deviations, respectively, above the chance level means of 0.454 and 0.309, indicating a consistency that is greater than chance in both data sets at a very high significance level. If the distributions of random CI values were normal, the probabilities corresponding to these deviations would be of order 10^{-13} and 10^{-20}. Although these significance estimates are probably unacceptable because the distributions are unlikely to be normal, as suggested by Fig. 2, in which the two distributions are shown, none of the 10,000 random permutations gave a CI as high as the observed indices: the probability of a random value as high as or higher than the observed one is most likely to be $<10^{-4}$ for both data sets. We must conclude therefore that there is a very highly significant correlation between genetic evolution and linguistic evolution, based on the previously published tree.

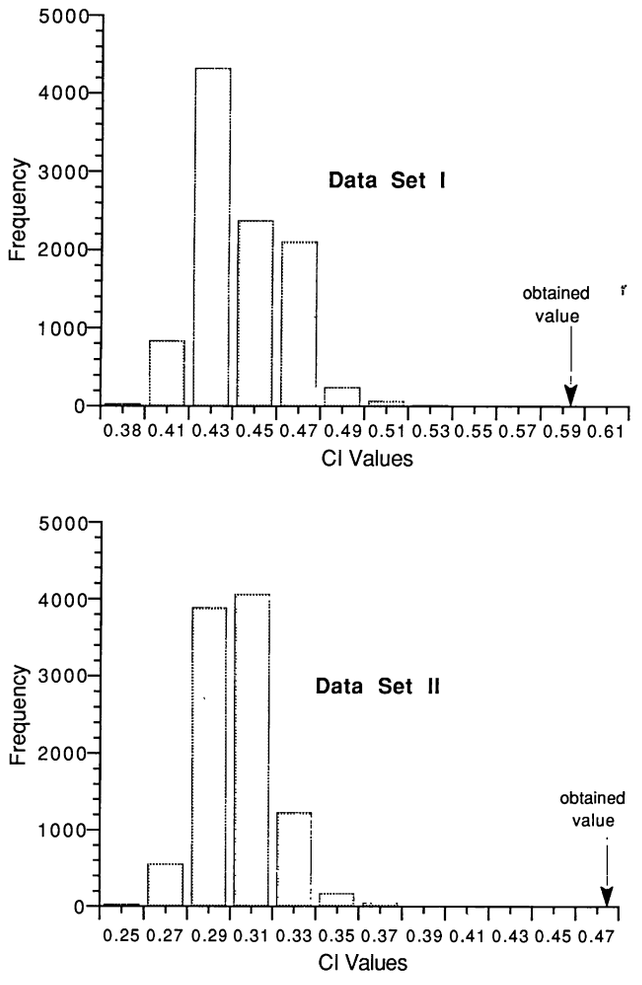

FIG. 2. Distribution of 10,000 random values of CI measuring the congruence of the linguistic families of 38 populations and their genetic tree (data set I) and the same populations after eliminating the 8 autapomorphic ones (data set II). In both cases, the observed CI is well above the range of random CIs.

Introducing Linguistic Family Groupings in the Evolutionary Analysis

Recent attempts at recognizing relationships between families of languages spoken in Europe, Asia, and Africa have led to the generation of linguistic family groupings. Two in particular are noteworthy: the Eurasiatic (ref. 13, p. 260) and the Nostratic groupings (ref. 13, p. 259). They are shown in Fig. 1, extreme right and share a common core of families: Indo-European, Uralic, and Altaic. Other families are added to the common core in each of the two groupings, undoubtedly because of differences in the criteria employed: multilateral comparison in the first case and the comparison of "protolanguages" (hypothetical ancestral languages reconstructed from modern ones) in the second. Protolanguages have been worked out for some linguistic families but not for others, limiting the comparability of the two family groupings. This may be one reason for the discrepancy between the two classifications. It is also likely, however, that Eurasiatic includes families with a slightly smaller degree of divergence than Nostratic (ref. 13, p. 259).

The disagreement between the two classifications is thus not substantial, so it is reasonable to consider pooling them in a larger Eurasiatic–Nostratic family grouping, which includes all the families found in either. We have extended the analysis by testing the significance of the observed CI value in four situations, labeled E, N, EN, and ENA in Table 1. These correspond to tests of the genetic tree versus linguistic classifications in which the Eurasiatic family grouping, the Nostratic family grouping, the union of Eurasiatic and Nostratic, and the union of Eurasiatic, Nostratic, and Amerind replace the individual families forming them. In each of the four situations we have tested two data sets, including or excluding autapomorphic populations, as in data sets I and II above, where all linguistic families were tested. One thousand permutations were tested in each case. The descriptions and results of the eight data sets are given in Table 1. It is clear that the correlation of linguistic and genetic data is very highly significant in all cases. Thus the partial evolutionary trees of languages formed by these family groupings are also in very good agreement with the genetic tree.

Global and Local Studies of Genetic and Linguistic Coevolution

Our global study demonstrates that the association between linguistic families and the genetic history of humans is far from random, and the significance test introduced here confirms our previous conclusions. In the introduction we discussed why such association is to be expected. One may also compare this result with those from regional comparisons of languages and populations.

Results of these local studies vary considerably. Most are based on relatively few languages and populations within restricted regions. In a survey of 15 studies (which did not include ours), 6 were found to be significantly in favor of an association (table 1 in ref. 25; we count as significant the case of Sardinia, not so given in the table). We can also include two comparisons for South America (16) and China (2) briefly described in the last part of this section; the first of these is negative and the second is positive and highly significant. The total number of studies is thus 17, 7 of which are significantly in favor of the correlation. This is a positive verdict, since less than one study ($0.05 \times 17 = 0.85$) would be expected to be

Table 1. Deviation between observed CI among genetic tree and linguistic families, or phyla, and the expected CI on the hypothesis of no correlation between them

Data set	P	G	CI		SD	SDs diff.
			Observed	Expected		
E1	38	13	0.520	0.3739	0.0230	6.35
E2	32	7	0.412	0.2534	0.0229	6.93
N1	38	13	0.591	0.3865	0.0313	6.53
N2	31	6	0.462	0.2364	0.0284	7.94
EN1	38	11	0.579	0.3433	0.0383	6.15
EN2	33	6	0.462	0.2372	0.0331	6.79
ENA1	38	10	0.556	0.3367	0.0424	5.17
ENA2	33	5	0.455	0.2368	0.0534	4.09

The following eight comparisons were made. In E1, the Indo-European, Uralic, Altaic, Eskimo-Aleut, and Chukchi-Kamchatkan families have been replaced by the Eurasiatic family grouping. In E2, Eurasiatic is as above, and six autapomorphic populations are removed (Mbuti, Nilosaharan, San, Southeast Indian, Northwest Amerind, and Australian). In N1, the Afro-Asiatic, Indo-European, Dravidian, Uralic, and Altaic have been replaced by the Nostratic family grouping. In N2, Nostratic is as above, and seven autapomorphic populations are removed (Mbuti, Nilosaharan, San, Eskimo, Chukchi, Northwest Amerind, and Australian). In EN1, all families found in the Eurasiatic and in the Nostratic group appear as a single family grouping. In EN2, five autapomorphic populations are removed (Mbuti, Nilosaharan, San, Northwest Amerind, and Australian). In ENA1, the Amerind language group is joined to the EN family grouping. In ENA2, five autapomorphic populations are removed (Mbuti, Nilosaharan, San, Northwest Amerind, and Australian). One thousand random permutations of the linguistic families were generated for each analysis, and the CI was calculated for each permutation. The first two numerical columns show the number of populations (P) in the tree and the number of linguistic groups (G) in the classification to which it is compared.The third column gives the observed CI; the fourth, the expected CI in the absence of correlation (the mean of the CIs of the 1000 randomly permuted classifications); the fifth, their standard deviation; and the last column, the deviation between observed and expected CI expressed as a multiple of the standard deviation (SDs diff.).

significant by chance out of 17 at a significance level of 5%. Seven significant studies, when less than one is expected, is extremely unlikely (3×10^{-5}) to be a chance result. There are, however, some sources of uncertainty in the majority of these studies, as discussed below.

It is unfortunate that most investigations of coevolution have employed linear correlation coefficients between some form of genetic distance and some form of linguistic distance. The particular form of genetic distance employed is not important. Measurements of genetic distance are numerous, but results are highly correlated even if their formulas are superficially very different (26). Two basic types of linguistic distance have been used: the frequency of shared cognates (its logarithm, with the sign changed, would be preferable) and the separation in a classification tree as measured by the number of tiers between the two languages and their common ancestor. The former method is reasonably accurate, especially if based on the standard Swadesh list of words, but the second is unlikely to give a precise distance or a time of separation. The rate of branching in an evolutionary tree of languages depends on their local birth and death rate, as new ones are born and old ones become extinct. Even if this rate were constant, the number of branchings would be subject to considerable stochastic fluctuations. But the rate of branching is very unlikely to be constant. As an extreme example, in the lineage eventually generating Bantu languages in the Niger-Kordofanian family, there are 17 nodes from the root to the most distant Bantu language, while all other branches are much shorter. The length of the Bantu lineage in terms of number of branchings reflects the success of the agricultural developments preceding the Bantu expansion and terminating with it and is a poor indicator of linguistic distance from the top to the bottom of the branch leading to Bantu languages. Tests listed in ref. 25 using as linguistic distance the number of separations in the linguistic evolutionary tree gave correlation coefficients not significantly different from zero. In part, this was also due to the fact that these investigations used very few languages, and correlation coefficients with a small number of observation pairs are unlikely to be significant given that the threshold for significance is very high.

There are other, more subtle drawbacks to the use of a linear correlation between a genetic distance and a linguistic distance. There is a serious problem in testing the significance of such correlations, because an observation from each of the n locations generates $n - 1$ pairs of comparisons (where n is the number of locations) with all the other locations. This generates an autocorrelation between pairs, which makes it unwise to use standard significance tests of these correlation coefficients. A solution is through Mantel tests (27), which were applied in only one of the cases listed in ref. 25. A possible further consideration is the need to correct for the effect of geographic distance, which is usually correlated with both genetics and language. It is unclear, however, whether this is truly necessary, as geography is part of the mechanism. Moreover, geography is not always strictly correlated with both genetics and languages, especially if tribes move around; in at least two cases, which are among the more complete studies (Chibchan languages in Central America and Han languages in China), geography has only a second-order effect.

At least two studies show very clearly that linear correlations of genetic and linguistic distances are misleading. Spuhler found a negative linear correlation between genetic and linguistic distance in North America, but was able to show the existence of coevolution by more subtle statistical techniques: (*i*) the variance of genetic distances among tribes belonging to the same linguistic family was significantly smaller than the overall variance, and (*ii*) discriminant analysis of genetic similarities between tribes allowed classification of a substantial proportion of tribes correctly in the same linguistic family (28). In another study Minch, Ruhlen, and Cavalli-Sforza (unpublished results) found a very weak positive correlation ($r = 0.191$) between genetic distance and geographic distance in South America and similarly weak negative correlations (-0.139 and -0.212) between linguistic and both genetic and geographic distance. South America has the highest intracontinental genetic variation, undoubtedly because of extreme drift (16). South America also has an exceptional geographic and ecological structure: the western part of the continent is formed by the Andes, and the eastern part is flat and shaped by a few independent, major river basins. The Andean region is relatively homogeneous genetically and quite different from the eastern region. Although the greatest geographic distances are between the northern and southern Andes, these have slightly less genetic and linguistic divergence than one finds between the eastern and western regions, which may explain the negative correlations.

These regional studies of genetic–linguistic correlation strengthen the case for coevolution of genes and languages, despite their use of inappropriate measures of association. An alternative to the linear correlation would be the direct comparison of trees, if possible. The solution we have introduced here may be useful in other applications when, as in our examples, one of the two descriptions is a multistate character and the other is a tree. We have compared the trees of Chinese languages and Chinese surnames reconstructed in the same 10 Chinese provinces (2). Surnames are highly correlated with genes here as elsewhere (10, 29, 30) and one can consider the tree of Han surnames as a good approximation of the genetic tree. Although the surname and lin-

guistic trees of Chinese Han were not identical, their topologies were so similar that they could be made equivalent by exchanging only two adjacent nodes out of nine, a result that would be extremely unlikely by chance unless the two trees were highly congruent.

We wish to express our gratitude to W. S. Wang for stimulating us to write this paper. Many thanks also to M. Feldman and M. Ruhlen for their comments and critical review. This work was funded in part by National Institutes of Health Grant GM 20467.

1. Greenberg, J. H. (1987) *Language in the Americas* (Stanford Univ. Press, Stanford, CA).
2. Mountain, J. L., Wang, W. S.-Y., Du, R., Yuan, Y. & Cavalli-Sforza, L. L. (1992) *J. Chinese Ling.*, in press.
3. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6002–6006.
4. Wright, S. (1943) *Genetics* **28,** 114–138.
5. Wright, S. (1946) *Genetics* **31,** 39–59.
6. Malecot, G. (1950) *Ann. Univ. Lyon Sci. Sect. A* **13,** 37–60.
7. Kimura, M. & Weiss, G. H. (1964) *Genetics* **49,** 561–576.
8. Morton, N. E. (1982) *Outline of Genetic Epidemiology* (Karger, Basel).
9. Cavalli-Sforza, L. L. & Wang, W. S. (1986) *Language* **62,** 38–55.
10. Du, R., Yuan, Y., Hwang, J., Mountain, J. & Cavalli-Sforza, L. L. (1992) *J. Chinese Ling.*, in press.
11. Cavalli-Sforza, L. L. & Feldman, M. W. (1981) *Cultural Transmission and Evolution* (Princeton Univ. Press, Princeton, NJ).
12. Hewlett, B. S. & Cavalli-Sforza, L. L. (1986) *Am. Anthropol.* **88,** 922–934.
13. Ruhlen, M. (1987) *A Guide to the World's Languages, Vol. 1: Classification* (Stanford Univ. Press, Stanford, CA).
14. Ruhlen, M. (1991) *A Guide to the World's Languages, Vol. 1: Classification* (Stanford Univ. Press, Stanford, CA).
15. Campbell, L. & Mithun, M., eds. (1979) *The Languages of Native America* (Univ. Texas Press, Austin, TX).
16. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1992) *History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ).
17. Kluge, A. G. & Farris, J. S. (1969) *Syst. Zool.* **18,** 1–32.
18. Archie, J. W. (1989) *Syst. Zool.* **38,** 239–252.
19. Sanderson, M. J. & Donoghue, M. J. (1989) *Evolution* **43,** 1781–1795.
20. Archie, J. W. (1989) *Syst. Zool.* **38,** 253–269.
21. O'Grady, R. T., Goddard, I., Bateman, R. M., DiMichele, W. A., Funk, V. A., Kress, W. J., Mooi, R. & Cannell, P. F. (1989) *Science* **243,** 1651.
22. Bateman, R. M., Goddard, I., O'Grady, R. T., Funk, V. A., Mooi, R., Kress, W. J. & Cannell, P. F. (1990) *Curr. Anthropol.* **31,** 1–13.
23. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1990) *Curr. Anthropol.* **31,** 16–18.
24. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
25. Barbujani, G. (1991) *Trends Ecol. Evol.* **6,** 151–156.
26. Karlin, S., Kenett, R. & Bonne-Tamir, B. (1979) *Am. J. Hum. Genet.* **31,** 341–365.
27. Mantel, N. (1967) *Cancer Res.* **27,** 209–220.
28. Spuhler, J. N. (1979) in *The First Americans*, eds. Laughlin, W. S. & Harper, A. B. (Gustav Fischer, New York), pp. 135–183.
29. Zei, G., Guglielmino, C. R., Siri, E., Moroni, A. & Cavalli-Sforza, L. L. (1983) *Human Biol.* **55,** 357–365.
30. Piazza, A., Griffo, R., Cappello, N., Grassini, M., Olivetti, E., Rendine, S. & Zei, G. (1992) in *Language Change and Biological Evolution*, eds. Piazza, A. & Cavalli-Sforza, L. L. (Stanford Univ. Press, Stanford, CA), in press.