Regional Linguistic and Genetic Differences among Yanomama Indians

The comparison of linguistic and biological differentiation sheds light on both.

Richard S. Spielman, Ernest C. Migliazza, James V. Neel

The processes of biological and cultural evolution, often viewed as entirely dissimilar, both result in the divergence of populations descended from the same ancestral group. We have had an unusual opportunity to determine the tempo of biological divergence and the degree to which it parallels one aspect of cultural divergence in a relatively isolated group, the Yanomama Indians of southern Venezuela and northern Brazil. We shall consider divergence in gene frequencies as a representative, indeed fundamental, indicator of biological evolution. No single feature is so obviously basic to cultural divergence, but we shall take language differentiation as representative (a case might be made for calling it fundamental, too). If there are common principles governing the processes of linguistic and genetic differentiation, one might expect to find a significant parallel between the two kinds of divergence; thus villages that are most similar genetically would be most similar linguistically, and vice versa. On the other hand, the absence of such correspondence might result from the operation in one or the other set of data of some extraneous influence whose identity could be of considerable importance. Moreover, data on linguistic divergence, unlike anthropometric and dermatoglyphic data (*1*, *2*), appear to provide a totally nonbiological means of assessing (and thereby refining our impression of) the unusual degree of isolation of the Yanomama.

The permanent contacts of the outside world with the Yanomama, who are still one of the least acculturated, relatively large tribal groups of South America, began only some 20 years ago. Their 150 or more villages occupy an area of roughly 500 by 300 kilometers, bordered by the equator and latitude 5° North, and by longitudes 61° and 66°30′ West. Our studies and those of others (*1*–*4*) have demonstrated that the Yanomama are distinct from other South American tribes in certain anthropometric and dermatoglyphic traits and in the presence or absence of some genetic polymorphisms. There is also differentiation within the Yanomama tribe; in all traits studied so far that are genetically (or partly genetically) determined, the Yanomama villages show marked differences from each other. The development of small genetic differences among related groups is called microdifferentiation, or microevolution, to distinguish the process from the subject of classical evolutionary biology, while recalling the basic similarities. The Yanomama differ culturally from surrounding tribes, although intratribally they are relatively homogeneous. They differ from neighboring Carib and Arawak groups in design and construction of the village house and in diet, of which the cooking banana (plantain) is the staple in most Yanomama areas. Unlike their neighbors, they lack dugout canoes and fermented beverages. Characteristic practices and beliefs include cremating the dead and drinking the ashes of their bones, nasal insufflation of hallucinogenic drugs, an ever-present cud of tobacco, and organized, duel-like fighting. A chanted, formal variety of the primary language is used as a lingua franca for intervillage communication. Finally, their language, which cannot be readily identified with any of the principal language families of South America, has set them apart from their neighbors.

The distinctiveness of all these features is best explained by the hypothesis that this tribe represents the descendants of a relatively small founding group, which, having reached this area some centuries ago, has differentiated in relative isolation ever since. When in the course of fieldwork it developed that there were clearly defined differences in dialect among Yanomama regions, we undertook to investigate the proposition that, under these circumstances, linguistic and genetic differentiation have proceeded in a parallel fashion.

"Natural" experiments are seldom as well controlled as those done in the laboratory, and this one is no exception. There is ethnographic and genetic evidence that within the past century the Yanomama have had contact with tribes to the north. The resulting genetic (and linguistic?) influences have been of two types: (i) those involving only a few individuals, either in peaceful exchanges or exchanges based on the abduction of women (*5*) and (ii) absorption of remnants of other tribes (*4*). From the outset, we recognized the possibility that such disturbances might obscure the similarities we wish to demonstrate, since linguistic and genetic variables might not be affected similarly.

We begin by sketching an analogy that restates the process of language change in the terms of population genetics and elaborating on some recent parallel developments in population genetics and linguistics. Next, we develop a measure of divergence for language, compatible with standard measures used in population genetics, and carry out a comparison of the two kinds of differentiation. Finally, the Yanomama language data are used to suggest a time depth for language divergence, which is related to the degree of genetic microdifferentiation.

A Population Genetics Analogy for Linguistic Divergence

We are by no means the first to urge comparisons of linguistic and biological evolution. The parallel was recognized even before Darwinian explanations were accepted by most biologists. The 19th century linguist August Schleicher entitled an 1863 monograph *Die Darwinsche Theorie und die Sprachwissen-*

R. S. Spielman is research associate in the Department of Human Genetics and J. V. Neel is Lee R. Dice University Professor of Human Genetics, University of Michigan Medical School, Ann Arbor 48104. E. C. Migliazza is assistant professor of anthropology and linguistics, University of Maryland, College Park 20742.

schaft (*6*) and explicitly incorporated evolutionary theory into historical linguistics. Gerard, Kluckhohn, and Rapoport (*7*) and Cavalli-Sforza (*8*) give modern illustrations of the analogy, and Cavalli-Sforza and Feldman (*9*) have proposed a mathematical model for transmission of cultural traits (especially language) to succeeding generations. We note in passing that there has been strong feeling among contemporary linguists, not necessarily shared by geneticists, that much of the child's knowledge of language structure is "innate" (*10*), hence presumably genetic [but see also (*11*, pp. 378, 394)].

The first step in biological microdifferentiation occurs when a population splits, giving rise to reproductively isolated subgroups that formerly interbred (*12*). Reproductive isolation makes possible divergence of allele proportions, conventionally ascribed to so-called forces of evolution; the most important such forces in the present case are differential reproduction (selection, which may be determined by the biological or the social environment) and sampling variation (random differences resulting from sampling a small, finite gene pool). In principle, a change in allele proportions may also be the result of accumulation of new mutations, but in practice this effect may be presumed to be negligible. Migration is also a force of evolution if the alleles introduced differ in proportions from alleles in the recipient population. In the case of the Yanomama Indians, migration from neighboring villages will, in general, counteract the effects of isolation and may therefore be expected to retard microevolution.

With this basic description of biological microevolution, we may speculate on an analogy with linguistic differentiation. Again it seems likely that isolation is a prerequisite for divergence. The forces of evolution, however, appear in somewhat different guise. Cultural transmission is Lamarckian (*13*)—that is, it permits the inheritance of acquired characteristics such as features of language. For this reason, sociological variables, especially differences in status, might be expected to predominate in determining changes in language. Like genetic traits, language variants in individuals or groups of some special status (for example, headman, shaman) might be differentially adopted by succeeding generations and perhaps lead in a particular village to rapid change, as has been shown in some regions of the United States (see *14*).

The same analogy applies, mutatis mutandis, to sampling variation and migration. Differences may be introduced by chance alone when a village fissions, possibly leading to some differentiation in language. The linguistic process analogous to migration in genetics is language contact, which may, but need not, involve actual geographical movement or incorporation of one group by another. Language contact between diverging groups must retard differentiation. In language, as in genetics, the ultimate source of variation must be the introduction of new elements. We are unable to say whether this kind of "mutational" process plays an important part in language differentiation, since not much is known about the genesis of new linguistic forms.

In fact, the occurrence and significance of variants is the focus of intense effort in both genetics and linguistics today. Biochemical techniques developed in the last two decades have revealed an unexpected degree of genetic variability in all species examined, including man. The realization that these observations do not fit simply into the body of population genetic theory developed since 1920 has led to critical reevaluation in both theory and observation. Summarized in Medawar's phrase, the major problem in population genetics is "the lack of a fully worked out theory of variation, that is, of the candidature for evolution . . ." (*15*, p. 1328). It is extraordinary that, concurrently, linguistics is undergoing a paradigm shift that also involves the appreciation of previously neglected variation. Historical linguistics has always acknowledged the succession of language forms. Until recently, however, each successive form was conceived of as uniform throughout the speech community. As a consequence, there was no theoretically viable explanation of the transition between stages (*16*). This predicament in the theory has been resolved by the demonstration that variation in structures and rules is normal and that "heterogeneity is not only common, it is the natural result of basic linguistic factors" (*17*, p. 42). This development in linguistics parallels precisely geneticists' recognition that a high proportion of genetic loci are probably heterogeneous (polymorphic). Thus the variability necessary for evolution is available in both a speech community and a biological population, and what eventually needs to be explained in both genetics and

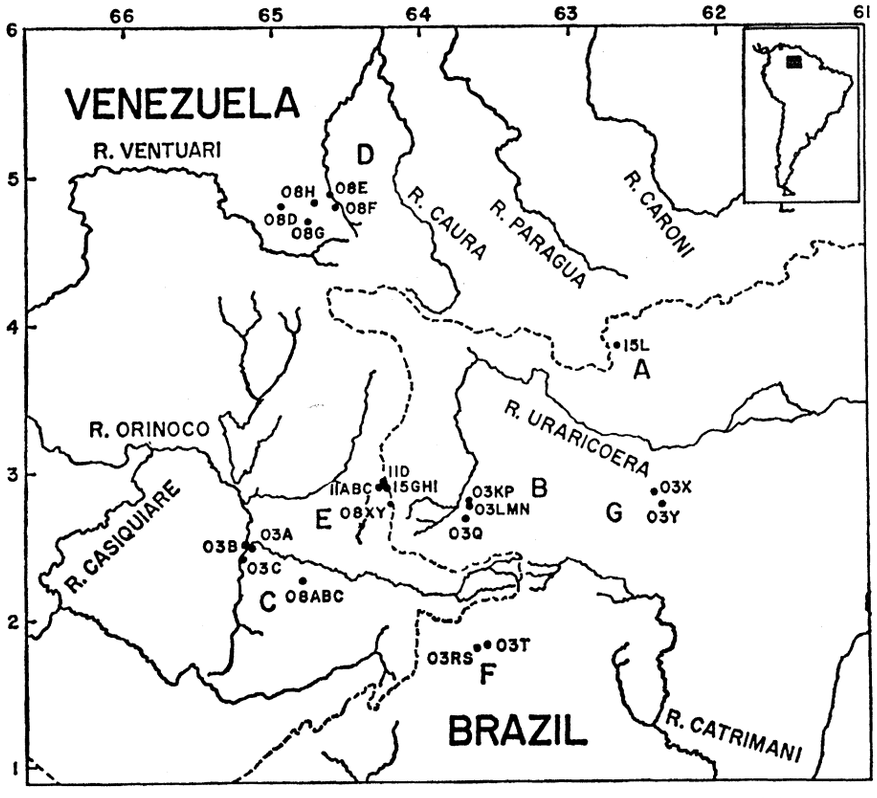

Fig. 1. Locations of the Yanomama Indian villages comprising the seven dialect areas (A–G) studied. The dialects group as follows into the four related languages studied by Migliazza (*21*): A and G, B and F, C and E, and D.

linguistics is the persistence of the variability and the differential success of the variants (selection).

For the present, our task is less ambitious: to compare observed biological and linguistic divergence in contemporary groups. Howells (*18*), Friedlaender *et al.* (*19*), and Spuhler (*20*) have carried out studies with a similar thrust. Howells' study did not include differentiation based on allele frequencies, and thus is not comparable to ours in this respect. Friedlaender *et al.* used allele frequencies for groups on Bougainville Island, as did Spuhler for North American and Mexican Indian tribes, but the scale of language difference in those groups seems vastly greater than that we encountered in the Yanomama. Correlation coefficients for genetic and linguistic differentiation were calculated in both these studies. Friedlaender *et al.* recognize some thorny problems in the statistical evaluation of their correlations and do not assign significance levels. Spuhler finds moderate negative correlations, which, regardless of significance level, cannot provide evidence for positive association between genetic and linguistic divergence. We have developed a different approach to the problem involved and will indicate the statistical significance of the correspondence we obtain between genetic and linguistic data.

Data on Linguistic and Biological Differentiation

Migliazza (*21*) has identified several varieties within the Yanomama language family and has grouped them into four languages on the basis of lexical, phonological, and syntactic differences. We have added data for dialects within three of these four languages, for a total of seven dialect areas. Data were gathered during the years indicated in the following villages (code designations in parentheses identify the villages in Fig. 1; see legend for language groupings of dialects):

A) *Yanam*—Aiwata-teri (15L), on the upper Uraricaá and Paragua rivers, from 1958 to 1960 (in extended contact with the Awake).

B) *Yanomam*—Aikama-teri (03KP), on the upper Parima River near the Surucucu Mountains, in 1965 and 1969.

C) *Yanomami*—Bisaasi-teri (03A) and Patanowa-teri (08ABC), on the upper Orinoco River, in 1968 and 1971.

D) *Sanima*—Hiiwiti-teri (08D) and Wasau-teri (08F), on the upper Erevato River, in 1968 (these villages are known to have had contact with at least one neighboring tribe).

E) *Yanoam*—Niayopa-teri (11ABC) and Mayopa-teri (08XY), at the headwaters of the Butaú and Maheko rivers, in the Parima mountain range, in 1971.

F) *Yanomai*—Hutua-teri (03S) and Horepa-teri (03R), on the Toototobi River, in 1965.

G) *Ninam*—Porai-teri (03X) and Apinas-teri (03Y), on the Mucujai River, in 1965.

The data on each dialect consisted of at least 1000 sentences and short texts that included the 750 words of the comparative vocabulary and different types of syntactic constructions. Almost all the data collected were recorded on magnetic tape and transcribed and translated in the field.

Inferences about relationships among the seven dialects are based on shared lexical items (percentage of presumed cognates) and grammatical rules. The percentage of cognates in a basic vocabulary that is presumed to change very slowly is often used as a measure of relationship (lexical divergence) among languages and families. In order to account for those morphological differences that are manifested syntactically by the addition or loss of low-level rules (*22*), we have compared not only Swadesh's basic vocabulary of 100 or 200 words (*23*), but an extended lexicon of 750 entries. The vocabulary items include 589 nouns and verbs, plus 161 minor morphemes, such as kinship terms, various classes of pronouns, and other formatives of noun and verb phrases.

Cognate percentages alone are insufficient as a measure of relationship among dialects, since dialects differ mainly in their grammar. The grammatical rules for this preliminary comparison are those accounting for syntactic and phonological differences. The major syntactic differences observed among the seven dialects included the processes of relativization, conjunction (coordinate and subordinate), noun phrases, possessive pronouns for kinship terms, the third person marker in verbs, the use of an auxiliary, and neutralization of the following features: plural and dual, inclusive and exclusive, animate and inanimate, and witnessed and unwitnessed.

Phonological differences are found in the underlying segments (systematic phonemes) and in the major processes (rules) that relate underlying phonological representations to surface phonetic ones. The underlying segments for each of the seven dialects, their sequential constraints, and the canonical patterns of their syllables were established and compared. The major phonological processes compared include: word stress assignment, vowel elision, spirantization, palatalization, /h/ deletion, vowel agreement, glide vocalization, nasalization, voicing of stops, and lateralization. Each dialect has been classified with respect to 15 syntactic and 23 phonological processes, so that every pair of dialects can be found to differ in from 0 to 38 grammatical rules.

Genetic typings for 25 systems have been carried out on blood samples from the inhabitants of 50 Yanomama villages (*24*). The eleven loci used here, each of which is polymorphic in at least one Yanomama village, are MNSs, P, Rh (four alleles only), Fy, Jk, Lewis, Hp, Gc, PGM, acid phosphatase, and Diego. For the calculation of allele frequencies, we have pooled Yanomama villages that speak indistinguishable forms of the same dialect. The resulting linguistic areas are represented in the genetic data by samples ranging in size from about 80 to about 300 individuals. The villages pooled for this purpose are identified on the map (Fig. 1) by codes used elsewhere (*25*). Allele frequencies for these seven groups were estimated using the maximum likelihood computer program, MAXLIK (*26*).

For the analysis that follows, one must be able to represent each dialect by a point in a multidimensional space; its coordinates are the variables that differ among the dialect areas. To achieve this goal, we have adapted a very popular biometrical concept, that of generalized distance [first employed by Mahalanobis *et al.* (*27*), but originally devised by Heincke (*28*) in a different form]. Consider first an example from genetics. Suppose allele frequencies at each of three loci are determined in two populations, A and B. In A, the frequencies are x_1, y_1, and z_1; in B, the corresponding frequencies are x_2, y_2, and z_2. We might obtain a single summary statistic incorporating all three differences, $x_1 - x_2$, $y_1 - y_2$, and $z_1 - z_2$, as shown in Fig. 2, by simply measuring the straight line connecting the points representing the populations. By an extension of the Pythagorean theorem, the square of the distance between the points is the sum of the squares of the differences. An analogous procedure has been developed for distance based on lexical differences, but it is necessary to imagine 750 mutually perpen-

dicular axes for the 750 potential cognates. For each word in the list, the judgment of cognate versus noncognate in the two languages is represented by the value 0 or 1. This procedure gives a list of 750 differences analogous to $x_1 - x_2$ above, from which the length of the straight line connecting each pair of points in lexical space may be calculated. By this definition, the total number of noncognates becomes the square of the distance between the two groups compared. The 21 lexical distances (squared) are given below the diagonal in Table 1. If the number of differences in the catalogue of rules (syntactic and phonological) is used, the same principle yields a distance reflecting differences in grammatical rules for each pair of villages, shown above the diagonal in Table 1.

The genetic distances, given in Table 2, were calculated by the method of Cavalli-Sforza and Edwards (*29*). For statistical reasons (*30*), the points are not represented directly by the frequencies, as in Fig. 2. The frequencies are adjusted by the "angular" transformation, and, from these transformed values, distances are calculated exactly as in Fig. 2.

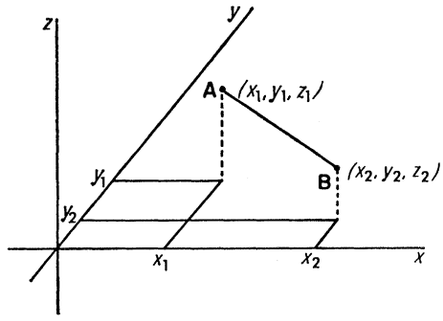

Fig. 2. A generalized distance in three-dimensional space. In this illustration, three variables (x, y, and z) have been measured in each population (A and B). Once the positions of the points representing the populations have been specified, the distance, which is the length of the line between the points, may be calculated from the coordinates by elementary geometry.

Comparisons of Two Sets of Data

Correspondence between two sets of data may be construed in various ways. By one interpretation, correspondence implies approximation to geometric congruence; that is, after a suitable transformation involving change of origin and scale and, if necessary, reflection and rotation, one set of data may be superimposed on the other. Congruence thus requires exact metric correspondence after (linear) transformation. A weaker, nonmetric form of correspondence may be defined as similarity between the cluster structure or hierarchic relations implicit in the data. Clearly, points (villages) might cluster similarly in two sets of data without being superimposable. Both interpretations of correspondence are presented below.

Comparing Cluster Structures by Using Dendrograms

It is convenient to summarize the cluster structure implicit in a set of distances with a dendrogram (tree structure, or network). For a large number of populations ($N \geqq 6$) there are many ways of connecting the points and nodes to make such a dendrogram. The technique we use, developed by Cavalli-Sforza and Edwards (*29*), is based on the principle that the most appropriate, or best, networks require the least total net length, or path length [see (*31*) for original motivation]. For seven populations there are 945 different networks of the kind used here. We may therefore identify for each set of data the best representations by examining the total length for all 945 possible nets.

In Fig. 3, a, b, and c, the best network thus obtained is shown superimposed on two-dimensional plots that represent the distances among the groups. In both the lexical and grammatical networks, two dialects of the same language (A and G, B and F, C and E) always cluster together. As a result, the two networks are structurally very similar, differing only in the way dialect D is related to the others. In spite of some differences, there are also prominent topological similarities between the linguistic and genetic networks. In both networks, areas A and G group together, as do B and F. Areas C and E, however, cluster closely in the linguistic networks, but not in the genetic. This situation illustrates the potential difficulties in reaching a conclusion from comparisons of different kinds of data by using a single diagram for each.

Recognizing that the possibly ambiguous relationships implicit in a set of distances may not be adequately represented by a single network, we have extended our comparison to a small number of the best networks for two sets of data. The procedure is sketched below; the details are described elsewhere (*1*). If two sets of data correspond in basic cluster structure, it follows that networks which represent one set well should also represent the other well. Basically, the test of association employed here determines whether networks that are good representations for one set of data are also good representations for the other. The 945 networks for each set of data are listed in order of increasing net length. Even in the absence of significant correspondence, a few networks will be, by chance, good representations for both sets. The comparison is therefore quantified by asking: Among some small fraction (say the best 50 nets) are more nets found in common than would be expected by chance alone? Unfortunately, no theoretical distribution is known for the number to be expected by chance alone.

Accordingly, the distribution of this test criterion, given only random correspondence, has been determined in a Monte Carlo computer simulation (*1*). Two sets of seven points, each set representing seven populations, are given randomly chosen coordinate positions between 0 and 1 on six axes, and the number of nets in common among the best 50 nets for each is determined. This procedure is repeated with new random coordinates for a total of 100 pairs, yielding a sample of the distribution of the "best-50" statistic (or of similarity of cluster structure) by chance alone. With the simulation results it is possible to specify the significance level at which the null hypothesis of no association may be rejected by a given comparison of real data.

For the comparison of lexical and genetic data, we in fact find 31 nets in common among the best 50. The largest number of nets in common by chance alone in the 100 simulated comparisons was 21, which occurred only once and determined the most significant probability (.01) that was directly observable; to indicate probabilities less than .01 directly requires more than 100 comparisons. However, a rough extrapolation from the shape of the observed part of the distribution indicates that 31 or more nets in common could be expected by chance alone with probability no greater than .005. Similarly, when all possible networks for the grammatical (rules) data are compared with those for the genetic data, we find 26 nets in common among the best 50. The two kinds of linguistic data yield 38 nets in common. All of these findings indicate correspondences significant at well beyond the .01 level.

Congruence of the Linguistic and Genetic Data

Having shown significant similarity in cluster structure between linguistic and genetic data, we proceed to test for a second kind of correspondence, that of congruence. The congruence of two sets of data is tested in analogy with the definition of congruent triangles, but in a space of more than two dimensions. For each set of data, we may use the pair-wise distances to find new coordinates for the N populations in as few dimensions as possible (maximum $N-1$). Using the method and program of Schönemann and Carroll (*32*) for comparing two particular sets of data, we transformed one matrix of coordinates by a combination of rotation, reflection, and uniform change of scale to yield a matrix fitted to the second, or target, matrix. The transformation sought is that which minimizes the sum of the squared deviations between elements of the fitted matrix and the target matrix. To quantify the fit of two matrices, Lingoes ond Schönemann (*33*) have developed a scaled, symmetric measure that is a function of the squared deviations, or residuals. The measure of fit, called S, has a range of 0 (perfect fit) to 1 (worst possible fit). Except for these two extremes, however, a particular value of S for two sets of data can be judged to be large or small only by comparison with some standard or with the distribution of S values obtained by chance alone, for which there is no known analytic expression. Such a distribution of S under the null hypothesis of random fit has been generated from the same kind of random sets of data described above and used for testing similarity of cluster structure. The value of S is now calculated for each pair of seven points with randomly generated coordinates. The process is repeated 500 times; the 500 values for S constitute a sample of the required distribution of the degree of congruence found by chance alone. On exactly the same principle described for testing similarity of cluster structure, a given value of S for real data may be compared with this distribution, and a value that falls in the extreme lower tail is said to be statistically significant.

When this procedure was carried out for the genetic and linguistic data, all three comparisons were found to be highly significant statistically. The respective probabilities of obtaining values as small as or smaller than the observed sample values by chance alone (as defined by the simulation) are as follows: lexical data versus genetic data, $P<.005$; grammatical versus genetic, $P<.01$; lexical versus grammatical, $P<.005$. Thus the test of geometric congruence corroborates the test for similarity of cluster structure by indicating correspondence at a statistically highly significant level.

Table 1. Linguistic differences (generalized distance, or D^2) among seven Yanomama dialects. Lexical D^2 (percent noncognate) below diagonal; rule (grammatical) D^2 above.

Language area	A	B	C	D	E	F	G
A		14	18	27	18	13	6
B	24.4		10	23	8	3	14
C	28.1	17.7		21	2	9	18
D	36.9	28.9	28.0		21	22	29
E	31.1	14.5	7.9	28.1		10	18
F	27.9	7.6	18.7	30.7	20.8		13
G	8.4	26.5	29.6	39.3	31.3	29.7	

Spuhler (*20*) was unable to demonstrate significant correspondence between genetic distances and estimated time since language divergence (glottochronological age) in 21 North and Central American Indian tribes representing ten linguistic stocks. These tribal languages are much older and more diverse than the languages within the Yanomama tribe. One may easily imagine that evidence of correspondence with genetic distances might have been obscured by undetected borrowing or imposition of language after conquest, which was not the case within the Yanomama. Nevertheless, Spuhler did show that the mean genetic distance between groups within the same language stock was significantly smaller than that between groups belonging to different stocks. This finding of smaller genetic differences between the linguistically more similiar groups may be interpreted as implying some correspondence of the kind demonstrated here for Yanomama language areas.

The very high levels of correspondence we find between genetic and a form of cultural microdifferentiation may result in part from the great degree of differentiation of Yanomama subgroups (*34*) and may not be easily demonstrated in situations where such divergence has been much less pronounced. For the biological data, the apparent rapidity of the divergence has been attributed to the population structure of the tribe, especially the frequent breaking up of villages, and to the tribe's isolation from outside influence.

The composition of our seven-region sample of the Yanomama may also increase the apparent level of correspondence. Statistically, the optimal situation for constructing a sample would approximate the following: (i) we

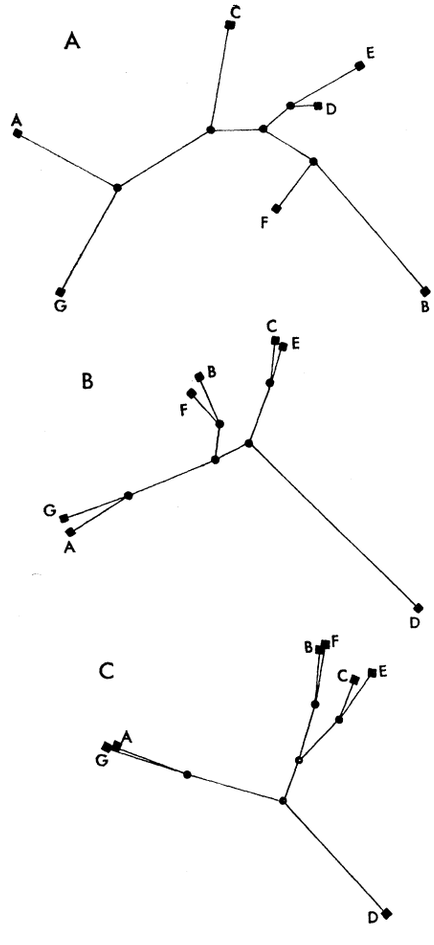

Fig. 3. Cluster structure implicit in the data on Yanomama Indian language areas. [A shows genetic data, B grammatical (rules) data, and C lexical (word list) data.] For each kind of data, the branching structure drawn is the one (out of 945 possible with seven points) that is the best representation of the groupings, according to the criterion described in the text. For this two-dimensional plot of the actual positions in six dimensions, the axes (not shown) are principal components, chosen in such a way that the projection minimizes the reduction of distances between groups. Solid circles indicate branch points; squares indicate population positions.

Table 2. Genetic distances and cognate percentages for Yanomama language areas. Genetic distances (D^2) below diagonal; percent cognates above diagonal (750-word list).

Language area	A	B	C	D	E	F	G
A		75.6	71.9	63.1	68.9	72.1	91.6
B	.391		82.3	71.1	85.5	92.4	73.5
C	.258	.264		72.0	92.1	81.3	70.4
D	.276	.181	.150		71.9	69.3	60.7
E	.258	.142	.150	.095		79.2	68.7
F	.184	.091	.157	.097	.110		70.3
G	.178	.285	.223	.229	.325	.168	

should have both genetic and language data on a very large number of villages or areas and (ii) we should be able to choose repeatedly a sample of the desired size by some randomizing scheme such that each unit is selected independently—that is, no unit by its presence in the sample alters the probability of another unit's being included. Unfortunately, the data do not exist in a form that permits this methodological purity. There simply are not a large number of Yanomama language groupings, and the logistics of work in the jungle makes random, independent sampling of the Yanomama territory unfeasible. Similar constraints are likely to hamper others who wish to test hypotheses about isolated, primitive groups.

We have pointed out how our sampling departs from the stated ideal: Each of the three most variable language areas identified in Migliazza's (*21*) treatment is represented by a pair of very similar dialects. In our set of seven, the inclusion of both members of these three pairs renders our sampling not strictly independent; thus, in effect, we are comparing somewhat fewer than seven separate groups (or 945 nets). The extent to which this effect operates in our data is very hard to gauge, but it is likely that our stated significance levels for observed degree of correspondence are overoptimistic. Such considerations do not apply, however, to the glottochronology.

Time Depth and the Process of Microdifferentiation

If the implications of the accumulated genetic data on the Yanomama are to be properly understood, we must develop some estimate of the period during which the Yanomama have been in relative isolation. Despite its limitations, the linguistic technique called glottochronology offers at present the sole means of obtaining such time depth for the Yanomama (*35*). We see two related time depths that bear directly on, and might ideally be expected to bracket, the duration of Yanomama language isolation. One is the time of separation of the most different Yanomama languages (dialects). The other is the time of separation of Yanomama from similar language families.

Although a considerably more sophisticated estimation of time depth may soon be possible (*36*), we shall use conventional glottochronology (*23, 37, 38*). The technique is based on a mathematical model in which lexical divergence of two related languages (or dialects) is considered to be exactly analogous to radioactive decay: The differential equations for the two processes are taken to be identical. Radiocarbon dating proceeds from a quantitative determination of ^{14}C and a knowledge of the general form of the process of decay. The corresponding quantity to be determined in glottochronology is the fraction of a lexical corpus still intact—the cognates remaining in a specially chosen core vocabulary of two languages. The rate of decay or retention—for example, the percent of cognates retained per millennium—must be assumed constant through time and must be known from previous work; neither of these requirements can be met perfectly. The linguistic process has been calibrated, so to speak, chiefly for written languages with well-documented histories, and there is considerable bias toward Indo-European languages. Even among this group, which might be expected to be artificially homogeneous, the estimated fraction retained per millennium varies at least from 0.65 to 0.97 (*37, 39*). Since the South American Indian languages used here have no written history, and information about retention rates in such circumstances is meager, we shall necessarily be limited to a range of time depths corresponding to the uncertainty about retention rate.

Table 3 indicates the basis for constructing such a range. The entries are time depths in years, calculated for various combinations of observed cognate percentages and assumed retention rates. Table 3 thus indicates the sensitivity of the conclusions to change (or errors) in the data or retention rate, as well as giving the range of values desired. From Table 2, we see that the smallest percentage of cognates retained (about 60 percent) is for languages D and G. If we assume the standard retention rate within each language—80.5 percent per millennium, the mean for 13 languages studied by Lees (*37*)—the corresponding estimate of time depth is a little less than 1200 years. It seems likely that the retention

Table 3. Estimated time depths for representative combinations of cognate percentages observed and retention rate assumed, calculated with the glottochronological model. [Entries are time depths (t_{ij}) in years, calculated from $t_{ij} = (\ln C_j)/(2 \ln r_i)$; the expression is derived by Lees (*37*).]

Cognate (C_j) (%)	Retention rate (r_i)					
	.60	.65	.70	.75	.805	.85
20	1,575	1,868	2,256	2,797	3,710	4,952
25	1,357	1,609	1,943	2,409	3,196	4,265
30	1,178	1,397	1,688	2,093	2,775	3,704
35	1,028	1,219	1,472	1,825	2,420	3,230
40	897	1,064	1,284	1,593	2,112	2,819
45	782	927	1,119	1,388	1,841	2,457
50	678	805	972	1,205	1,598	2,133
55	585	694	838	1,039	1,378	1,839
60	500	593	716	888	1,177	1,572
65	422	500	604	749	993	1,325
70	349	414	500	620	822	1,097
75	282	334	403	500	663	885
80	218	259	313	388	514	687
85	159	189	228	282	375	500
90	103	122	148	183	243	324
95	50	60	72	89	118	158

rate in illiterate groups like the Yanomama, which lack the conservative influence of written language, is lower than in the civilized groups that provided the standard rate. On the other hand, the Yanomama dialects have not developed in complete isolation from each other, a factor that must operate to increase the apparent retention rate, as might the existence of the lingua franca mentioned above. Guessing that the former effect is more important than the latter, we have decided to present time depths as the range corresponding to retention rates between 0.65 and 0.805. Consequently, the maximum duration of separation between Yanomama dialects is estimated at 600 to 1200 years. The minimum lexical divergence between Yanomama dialects, represented by cognate counts of about 92 percent, corresponds by the same convention to a time separation of 75 to 200 years.

For the time of isolation of Yanomama from other South American language groups we have only a first estimate. Ideally, we should have cognate counts for comparing the Yanomama language with a large number of representative languages from various parts of South and Central America. Since such language studies have been very limited, we now have data for only a modest sample of major groupings. The native languages for which we have word lists are all located in Panama or northern South America east of the Cordillera, a relatively small part of the continent. Other languages studied, and the percent of cognates with Yanomama in Swadesh's 100-word list, are Shipibo (Peru), 27 percent; Warao (Orinoco delta, Venezuela), 27 percent; Guaymí (Panama), 25 percent; Macushi (Brazil), 25 percent; Makiritare or Yekuana (Venezuela), 23 percent; Trio (Brazil), 21 percent. Given the imprecision of these figures as measures of language similarity, no one language group studied so far stands out on the basis of cognates as strikingly similar to Yanomama; the restricted geographical spread perhaps accounts for some of the similarity among these figures.

The Shipibo-Yanomama and Warao-Yanomama comparisons indicate a time depth of 1500 to 3000 years (using retention rates of 0.65 and 0.805, respectively). These times, taken at face value, imply that the internal linguistic differentiation of Yanomama began about 1000 to 2000 years after the separation of Yanomama as a whole from the most similar of the other South American languages so far tested. This conclusion regarding relative ages remains valid, even if the absolute time depths are inaccurate. For any given retention rates, the glottochronological time depths are necessarily underestimates (*23, 38*), since effects of contact between languages are inevitably ignored. The estimates obtained, however, are consistent with the considerable genetic divergence from other South American Indian tribes, in conjunction with the substantial genetic microdifferentiation within the Yanomama tribe.

The time depths, although approximate, also provide a chronological yardstick for measuring the rate at which allele frequencies change—that is, the rate of evolution, both within and between tribes. For example, the Yanomama have a variant form of serum albumin not seen in any neighboring tribes; in a few villages, the allele for this protein has a frequency greater than 0.35. From its apparent absence in nearby tribes, we infer that all contemporary instances of the allele are descendants of a single mutant that arose after the ancestors of the Yanomama and their neighbors were separated (*40*). The time depths supplied by linguistic data should permit us to calculate the rate of increase necessary to reach the present frequency from a single, original allele. Given the necessary statistical theory, it will then be possible to ask how plausible this rate of increase would be in the absence of natural selection. This hypothesis, that the allele is "neutral" with respect to selection, may also be compared with the possibility of various kinds and intensities of selection. Inferences from linguistic data should thus make possible conclusions about genetic processes.

Summary

The extent of linguistic differentiation and its correspondence to genetic differentiation have been examined in seven language areas of the Yanomama Indians, a relatively isolated South American tribe. Using an objective method for quantifying the degree of similarity in the cluster relationships implicit in different kinds of data, we find that the linguistic divergence of the seven areas corresponds significantly with the pattern of genetic microdifferentiation. This conclusion is corroborated by a test that measures geometric congruence of the patterns of village positions specified by two different kinds of data. The analogy between linguistic and genetic divergence, first recognized well before 1900, is redeveloped in the light of more recent population genetics.

Approximate time depths, necessary for an estimate of the *rate* of genetic divergence, have been calculated from cognate counts among the dialects. The time depths suggest that the Yanomama dialects have been diverging for about 1000 years. Cognate percentages with non-Yanomama languages lead to the tentative conclusion that Yanomama languages as a whole diverged at least 1000 years earlier still. These linguistic findings are consistent with the well-documented genetic finding that the Yanomama show a high degree of microdifferentiation in allele frequencies within the tribe, with a marked genetic distinctiveness when compared to other South American tribes. The use of time depths and linguistic distances in genetic studies is illustrated with an example from South American Indians.

References and Notes

1. R. S. Spielman, *Am. J. Phys. Anthropol.* **38**, 461 (1973).
2. J. V. Neel, F. Rothhammer, J. C. Lingoes, *Am. J. Hum. Genet.*, in press.
3. N. A. Chagnon, *Yąnomamö: The Fierce People* (Holt, Rinehart & Winston, New York, 1968), pp. xiv, 142; J. V. Neel, *Science* **170**, 815 (1970).
4. E. C. Migliazza, *Bol. Mus. Paraense Emilio Goeldi Nova Ser. Antropol.* No. 22, 1 (30 June 1964); *ibid.*, No. 25, 1 (8 March 1965).
5. N. A. Chagnon, J. V. Neel, L. R. Weitkamp, H. Gershowitz, M. Ayres, *Am. J. Phys. Anthropol.* **32**, 339 (1970).
6. A. Schleicher, quoted by J. H. Greenberg, in *Language and Evolutionary Theory*, J. H. Greenberg, Ed. (Univ. of Chicago Press, Chicago, 1957), p. 56.
7. R. W. Gerard, C. Kluckhohn, A. Rapoport, *Behav. Sci.* **1**, 6 (1956).
8. L. L. Cavalli-Sforza, in *Mathematics in the Archaeological and Historical Sciences*, F. R. Hodson, D. G. Kendall, P. Tăutu, Eds. (Edinburgh Univ. Press, Edinburgh, 1971), p. 535.
9. ——— and M. W. Feldman, *Theor. Popul. Biol.* **4**, 42 (1973).
10. N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, Mass., 1965), pp. 58–59; *Language and Mind* (Harcourt, Brace & World, New York, 1968), p. 69.
11. E. H. Lenneberg, *Biological Foundations of Language* (Wiley, New York, 1967).
12. E. Mayr, *Animal Species and Evolution* (Harvard Univ. Press, Cambridge, Mass., 1963).
13. P. B. Medawar, *The Future of Man* (Basic Books, New York, 1960), p. 98.
14. W. Labov, *Word* **19**, 273 (1963); *The Social Stratification of English in New York City* (Center for Applied Linguistics, Washington, D.C., 1966).
15. P. B. Medawar, *Nature* **207**, 1327 (1965).
16. R. D. King, *Historical Linguistics and Generative Grammar* (Prentice-Hall, Englewood Cliffs, N.J., 1969), p. 189.
17. W. Labov, *Stud. Gen.* **23**, 30 (1970). See also, U. Weinreich, W. Labov, M. I. Herzog, in *Directions for Historical Linguistics*, W. P. Lehmann and Y. Malkiel, Eds. (Univ. of Texas Press, Austin, 1968), p. 97.
18. W. W. Howells, *Curr. Anthropol.* **7**, 531 (1966).

19. J. S. Friedlaender, L. Sgaramella-Zonta, K. K. Kidd, L. Y. C. Lai, P. Clark, R. J. Walsh, *Am. J. Hum. Genet.* **23**, 253 (1971).
20. J. N. Spuhler, in *The Assessment of Population Affinities in Man*, J. S. Weiner and J. Huizinga, Eds. (Clarendon, Oxford, 1972), p. 72.
21. E. C. Migliazza, *Yanomama Grammar and Intelligibility*, thesis, Indiana University, Bloomington (1972).
22. Low-level rules are rules that are "low" in the derivation—that is, near the syntactic surface structure and accordingly language specific.
23. M. Swadesh, *Proc. Am. Philos. Soc.* **96**, 452 (1952).
24. H. Gershowitz, M. Layrisse, Z. Layrisse, J. V. Neel, N. Chagnon, M. Ayres, *Ann. Hum. Genet.* **35**, 261 (1972); L. R. Weitkamp, T. Arends, M. L. Gallango, J. V. Neel, J. Schultz, D. C. Shreffler, *ibid.*, p. 271; L. R. Weitkamp and J. V. Neel, *ibid.*, p. 433; R. Tanis, J. V. Neel, H. Dovey, M. Morrow, *Am. J. Hum. Genet.* **25**, 655 (1973).
25. R. H. Ward, *Ann. Hum. Genet.* **36**, 21 (1972).
26. T. E. Reed and W. J. Schull, *Am. J. Hum. Genet.* **20**, 579 (1968).
27. P. C. Mahalanobis, D. N. Majumdar, C. R. Rao, *Sankhyā* **9**, 89 (1949).
28. F. Heincke, *Abh. Deutsch. Seefischerei-Vereins* **2**, 1 (1898).
29. L. L. Cavalli-Sforza and A. W. F. Edwards, *Am. J. Hum. Genet.* **19**, 233 (1967).
30. A. W. F. Edwards, *Biometrics* **27**, 873 (1971); the variance of a distribution of frequencies or proportions is not independent of the mean. This undesirable property is eliminated or reduced to negligible importance by the so-called angular transformation of the frequencies (*41*) which we have used throughout.
31. A. W. F. Edwards, in *Mathematics in the Archaeological and Historical Sciences*, F. R. Hodson, D. G. Kendall, P. Tăutu, Eds. (Edinburgh Univ. Press, Edinburgh, 1971), p. 347.
32. P. H. Schönemann and R. M. Carroll, *Psychometrika* **35**, 245 (1970).
33. J. C. Lingoes and P. H. Schönemann, *ibid.*, in press.
34. T. Arends *et al.*, *Proc. Natl. Acad Sci. U.S.A.* **57**, 1252 (19
J. V. Neel and R. H. Ward, *ibid.* **65**, 323 (1970).
35. Conventional historical linguists have been extremely critical of both the theory (*42*) and results (*39*) of glottochronology. A group of statisticians has recently shown (*43*) that the criticisms of the principles arise from misunderstandings of the statistical basis of glottochronology, thereby invalidating the most serious objections. As a way of acknowledging the imprecision in practice of the results of any glottochronological study, we use a wide range of values for each estimate of time depth.
36. D. Sankoff, in *Mathematics in the Archaeological and Historical Sciences*, F. R. Hodson, D. G. Kendall, P. Tăutu, Eds. (Edinburgh Univ. Press, Edinburgh, 1971), p. 381; J. B. Kruskal, I. Dyen, P. Black, in *ibid.*, p. 361.
37. R. B. Lees, *Language* **29**, 113 (1953).
38. D. Hymes, *Curr. Anthropol.* **1**, 3 (1960).
39. K. Bergsland and H. Vogt, *Curr. Anthropol.* **3**, 115 (1961).
40. R. J. Tanis, R. E. Ferrell, J. V. Neel, M. Morrow, *Ann. Hum. Genet.*, in press.
41. C. Eisenhart, in *Techniques of Statistical Analysis*, C. Eisenhart, M. W. Hastay, W. A. Wallis, Eds. (McGraw-Hill, London, 1947), p. 397.
42. C. D. Chrétien, *Language* **38**, 11 (1962).
43. A. J. Dobson, J. B. Kruskal, D. Sankoff, L. J. Savage, *Anthropol. Linguist.* **14**, 205 (1972).
44. We thank A. L. Becker, B. Dyke, J. MacCluer, W. Morrill, P. Smouse, A. Templeton, and R. H. Ward for critical comments. This research was supported by the Atomic Energy Commission and the National Science Foundation.

NEWS AND COMMENT

The Sloan-Kettering Affair: A Story without a Hero

There is no sin in science more grievous than falsifying data. There is no accusation that can be made against a man more serious than that he is guilty of such a sin. The very thought of fakery threatens the powerful mystique of the purity of science. It stirs deep and contradictory feelings of incredulity, outrage, and remorse among the entire scientific community—feelings it is experiencing now in the very complex and unresolved matter of William T. Summerlin.

The case has received widespread attention in the press, and by now it is well known that Summerlin, a young investigator at the Sloan-Kettering Institute for Cancer Research in New York, is alleged to have falsified the results of an experiment intended to prove that skin, when grown in tissue culture, loses its ability to provoke an immune response. Late in March, he was "temporarily relieved of his responsibilities" by institute president Robert A. Good, who has promoted Summerlin for the past couple of years. Good appointed a committee of Sloan-Kettering scientists to investigate the situation.

It is understood that the committee is investigating not only the mouse painting incident, which is alleged to have happened during the last few weeks, but the whole of Summerlin's work which has been cast into doubt.

A spokesman for the institute says there will be a "full disclosure" of the review committee's findings when they are complete, but, as yet, the full facts of the case are unknown. The institute and the committee are unwilling to discuss the matter, even declining comment on precisely what it is they are investigating. According to Summerlin, even he has not been informed about what is going on. "I know nothing about the review committee, and I don't know what they are reviewing," he told *Science*. He does, however, expect to appear before the committee before it concludes its work. "There are two sides to this story, and I want to tell mine [to the public] after the committee is done and I've had a chance to talk to them," Summerlin declared.

The Summerlin case raises issues that go far beyond the question of whether one man has or has not literally falsified data. That is important, but the stakes are even higher.

First, there is Summerlin's own reputation, already seriously damaged by an accusation, reportedly made by a Sloan-Kettering laboratory attendant, that he painted black patches on the skin of white mice to make it appear that he had successfully transplanted skin between genetically incompatible animals. When asked by *Science* whether he had painted mice, Summerlin said, "I have never willfully misrepresented my data. I look forward to continuing in science."

Beyond the alleged mouse painting incident is the question, now on many persons' minds, of the validity of the whole of Summerlin's work during the last 4 years, work which potentially has enormous implications for research in immunology and cancer.

Second, there is the reputation of Good, Summerlin's boss and mentor, a tremendously powerful and persuasive man who has lent the prestige of his own stature to the work Summerlin has been doing.

Third, there is the reputation and internal stability of Sloan-Kettering itself, a troubled institution that has been struggling for the last couple of years to find its identity and its future.

The present crisis could turn out to be ruinous for all three.

The simplest of the questions to be resolved in the Summerlin case is whether or not he really painted mice for the purpose of deceiving his colleagues. Possibly the answer to that is already known to the review committee.

The other questions are not even potentially easy to answer. They have to do with the environment in which research is conducted, with pressures to succeed in a spectacular way, and with who properly gets credit for what. They also have to do with how research should be presented to the scientific community and to the public and what one does when, seemingly all of a sudden, one can no longer repeat one's own experiments and it is possible