Differences among Yanomama Indian Villages: Do the Patterns of Allele Frequencies, Anthropometrics and Map Locations Correspond?

RICHARD S. SPIELMAN [1]
Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48104

KEY WORDS Microdifferentiation · Genetic and anthropometric distance · South American Indians.

ABSTRACT In order to determine the degree of correspondence between sets of multivariate observations based on different kinds of traits, two new methods, derived from fundamentally different notions of "correspondence," are adopted here and compared. Using networks or trees to represent contemporary relationships, the first method tests the similarity of the cluster or hierarchic structures implicit in two sets of data. The second approach tests the departure from perfect geometric congruence or superimposability. Computer simulation was used to generate the distributions needed for significance tests under the null hypothesis.

By the first technique, we find significant correspondence among the cluster structures for geographic, allele frequency, and anthropometric data on 19 Yanomama Indian villages. The results are similar and more precise for a subset consisting of seven villages. Some of these results differ from the conclusions which would be reached with the conventional correlations based upon entries in distance tables.

The direct test of congruence, used only for the data on the subset of seven villages, gives results which differ substantially from those based on cluster-structure. There are, however, similarities between the measure of congruence and the simple correlations based on entries in the distance tables.

The significant correspondences observed call for some explanation. Cultural and demographic features determine the particular non-random allocation of individuals to village fragments when a village splits. These social phenomena are invoked in tentative explanation of the agreement among historical, biological, and geographic relationships of villages.

The classical study of animal evolution and speciation examines the results of major genetic changes which require thousands of generations. Over short periods of five to ten generations, the process of evolution reveals itself, if at all, as subdivision and differentiation within a species. In the present paper and a companion piece (Spielman, '73) I have tried to elucidate this differentiation in a tribe of "our contemporary ancestors," by bringing together the materials of physical anthropology with those of population genetics.

Small human groups with a common origin but only restricted mutual contact or exchange are expected to become increasingly dissimilar over time. The present paper examines the resulting differentiation in biological traits among villages of the Yanomama Indians, and quantifies by two techniques the correspondence of the pattern of differentiation in various traits. The descriptive study of this dispersion process has long been a concern of physical anthropologists. The classic attempt to determine whether observed differences in physique reflect his-

[1] Supported in part by U.S.P.H.S. Training grant 5-T01-GM00071-10, the U.S. Atomic Energy Commission, and the National Science Foundation. The Computing Center of the University of Michigan provided computer time. Part of a doctoral dissertation submitted by R. S. Spielman to the Graduate School of the University of Michigan.

torical and contemporary ethnological relations was that of Mahalanobis, Majumdar and Rao ('49); the basic technique was to calculate generalized distance, a composite measure of morphological difference between groups. More recently, as it has become possible to define differences between populations in strictly genetic, quantitative terms, the same sort of problem has been approached from the viewpoint of population genetics, with the additional goal of identifying the forces of evolution contributing to differentiation. Following Sanghvi ('53), the usual approach has been to use generalized distances to answer the question: "Do different systems of variables (genetic, morphological) reflect the relationships between groups in the same way?"

Implicit in this question is a notion of correspondence between sets of data. Although earlier studies (Sanghvi, '53; Howells, '66; Friedlaender, '69) have never defined this concept in a rigorous way, at least two reasonable interpretations may be provided. Ignoring the absolute magnitude of distances between groups, we may cluster them on the basis of relative distance; then "correspondence" may be construed as similarity of cluster or hierarchic structures so derived for different sets of data. Cluster similarity in this sense is a kind of non-metric correspondence. On the other hand, "correspondence" may be viewed differently and taken as exact geometric congruence. First the positions of the populations in multidimensional space are specified by the coordinates for each set of data. Then two sets should be understood to correspond if the points are congruent, or can be made congruent by a linear transformation.

In what follows, various sets of data are tested for correspondence using both the definitions given above. In both cases, significance tests are constructed empirically, by simulating with a computer the two types of comparisons under the respective null hypotheses of no correspondence. It should be apparent *a priori* that sets of data which are found to correspond under one definition will not *necessarily* do so under the other. Since cluster structure ignores important metric differences, examples may easily be imagined where no linear transformation to achieve congruence is possible, but where cluster similarity is nevertheless substantial.

Methodological issues

Howells ('66) has tried to cast the comparisons of different kinds of variation, including geographic, anthropometric, and genetic variation, in a way which might directly yield biologically meaningful results. In a lucid review of the difficulties with this kind of inference. Friedlaender ('69) has stressed that it is not apparent what kind of correspondence one should expect when comparing marker gene (i.e., blood group, serum protein, and erythrocyte enzymes) and morphological differentiation with each other and with geographic separation. First, unlike marker gene traits, morphological features are highly susceptible to environmental influence during development. As a result, even when two groups are genetically indistinguishable for both marker and morphological traits, environmental (developmental) effects on the latter may result in prominent morphological differences, with consequent discrepancies between marker and morphological patterns of differentiation (Hiernaux, '56).

The marker gene phenotypes also differ from the morphological features in aspects other than susceptibility to environmental modification. The former traits are determined by alleles at a single locus or a few closely linked loci, while the determination of metric traits is polygenic. For this reason, it is usually presumed that marker gene traits and traits determined genetically by many loci might be influenced by selection or genetic drift in different ways. One might thus doubt that anthropometric and marker gene frequency differences will correspond significantly, or that they will reflect geographic relationships.

Apart from the methodological problems in extracting biological meaning from the correspondence of different sets of variables, all previous studies have suffered from the lack of an appropriate objective technique to specify the degree of correspondence. For two univariate sets of observations, there exist numerous appropriate measures of correlation with analytically derived distributions. There

is no analogous statistical procedure for making comparisons between sets of multivariate observations. In the absence of an appropriate technique, previous studies have resorted to two approaches. The simpler, essentially intuitive solution, especially applicable with no more than five to ten groups, is to present two-dimensional plots as approximate representations of distance relationships derived from each kind of variable, and to encourage the reader to reach the author's conclusion concerning the similarity of two such structures to each other or to the geographic distribution (Pollitzer, '58; Majumdar and Rao, '60; Chai, '67).

The second approach is an elaboration of the methods of Cavalli-Sforza and Edwards ('64, '67), who introduced a technique of "phylogenetic analysis" which although essentially heuristic has simplified and improved the representation of group relationships. Evolutionary relationships are inferred from generalized distances and represented by a network or tree-diagram. In this way a set of populations representing every inhabited continent has been analyzed (Cavalli-Sforza and Edwards, '64). On a much smaller geographic scale, the technique has been applied to tribal populations in detailed comparisons of gene frequency differences with known historical relationships and other kinds of data on differentiation (Ward and Neel, '70; Sinnett, Blake, Kirk, Lai and Walsh, '70; Friedlaender, Sgaramella-Zonta, Kidd, Lai, Clark and Walsh, '71; Ward, '72).

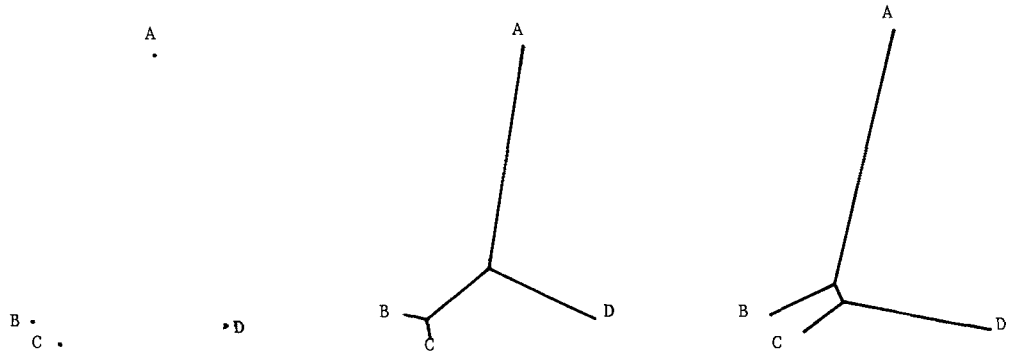
With a large number of populations, a comparison of two phylogenetic tree-diagrams by inspection may be very difficult. Even with small numbers of groups, the overall correspondence between different sets of data may not be apparent. The diagrams have usually been supplemented or replaced therefore with a measure of correlation applied directly to two tables of distances, for example the correlation of morphological distance with marker gene distance over all pair-wise distances (Howells, '66; Workman and Niswander, '70; Friedlaender et al., '71). The statistical shortcomings of this approach, in which the degrees of freedom for a significance test of the comparison are likely to be exaggerated, have been emphasized repeatedly (Ward and Neel, '70:541; Friedlaender et al., '71: 267–268), but no better alternative was available. In addition, the correspondence of the pair-wise distances is not equivalent to, and may not always reflect, correspondence of the points; some practical difficulties which result are illustrated in a later section.

Cluster correspondence

The approach developed here takes from Cavalli-Sforza and Edwards the idea of representing group relationships by trees or networks, but uses a network only as a scheme of contemporary relationships. No attempt is made to develop evolutionary inferences; in this context it is unimportant whether the actual evolutionary process meets the assumptions of the model of Cavalli-Sforza and Edwards ('67). Although in the nomenclature of graph theory, the representations used here are called "trees" and are identical to the trees used by Kidd and Sgaramella-Zonta ('71), I use "nets" or "networks" to avoid the phylogenetic connotations of the tree terminology, and for consistency with Prim's ('57) original usage.

We begin with the observation that different topologies or branching structures applied to the same set of data require different total path lengths; the amount of "string" necessary to connect a set of points depends on the way in which the points are connected. In principle, some unusual sets of data or points might be connected by different nets with the same path length — the four vertices of a regular tetrahedron provide an example. In practice, such cases are rare. We follow Edwards and Cavalli-Sforza ('63) and Cavalli-Sforza and Edwards ('67) on partly pragmatic grounds and define as *better* representations of the true relationships those topologies having small total lengths. (See fig. 1 for an intuitive justification.) The same points may be connected using less string if points which "belong" together are grouped together. It takes more string, i.e., total net length, when the groupings are discordant with the distances. Edwards ('71) gives the theoretical motivation for this argument.

For N populations ($N \geq 3$), the number of different topologies is given by $\prod_{i=3}^{N} (2i-5)$.



Artificial example: positions of four populations (Visualize in three dimensions)

Shorter total path length: better representation groups B with C before adding D

Longer total path length: poorer representation groups D with C before adding B

Fig. 1 Rationale for path length criterion: example with four populations. Figure exaggerates difference in total length.

The task of finding for each set of data the best representation of the relationships of the points, reduces to finding that network with minimum path length. Unfortunately, there is no algorithm or constructive technique to produce the desired net, and for eight or more populations, the number of possible topologies is greater than 10,000. Therefore only in those cases where it is possible to enumerate and evaluate all possible nets, that is, when there are seven or fewer populations, will we be certain of finding the single best one. For 19 populations, the largest set treated here, total enumeration is not feasible; for this set it will be necessary to work with the best net we can find, knowing that there are probably still better ones not identified.

We now re-phrase the goal of comparing relationships inferred from different sets of data, e.g., anthropometric, marker gene and geographic distances (or the corresponding coordinates). If two sets of data have similar cluster-structures, it follows that nets which are good representations for one set should also be good representations for the second. Accordingly, among the large number of possible topologies for the first set, we choose the best we can find (see below) and evaluate how well it represents the data of the second; i.e., we ask, "does it yield a relatively small total path length on the second set too?" Strictly speaking, the technique proposed here for evaluating the correspondence of entire sets of genetic and anthropometric data compares the best topology or net implied by one set with the distribution of possible nets for the other or "reference" set.

A possible misunderstanding of this method for comparing different kinds of data must be anticipated here. The similarity of two networks for the same or different sets of data cannot be established from the similarity of their path lengths. Indeed, *the degree of similarity of two networks* which are not identical is not defined by the techniques used here. It does follow however from the fact that a particular net for one set of data has a total path length which is more than three or four standard deviations lower than the mean path length of all nets for those data, that the net is a "good" representation. Thus, two such nets would both be good, but no assertation is made about their similarity. Throughout the present paper we infer correspondence between cluster structures of two sets of data by evaluating representation or fit; no attempt is made to establish the similarity of two nets.

Correspondence as congruence

For the interpretation of "correspondence" as geometrical congruence, I have taken over directly the least squares method of Schönemann and Carroll ('70) which fits one matrix to another by linear transformation, and is apparently equivalent to the technique sketched by Gower ('71). As described by the former authors,

one set of data A is fitted to another set B so that

$$B = c\,A\,T + Jx' + E,$$

where A and B are pxq matrices (of coordinates), and where T is the qxq transformation matrix defining a rigid rotation, x is the vector defining the translation of the origin, $J' = (1, 1, \ldots 1)$, and E is the residual matrix, i.e., the matrix of differences between the elements of B and the matrix $c\ A\ T + Jx'$. The least squares solution sought is that choice of T, J, c, and x which minimizes the sum of squared elements of E, given by the trace of E'E.

Schönemann and Carroll ('70) point out that this sum of squares may also serve as a measure of fit. In general, however, fitting A to B will not give the same residual matrix E as fitting B to A, so that the measure of fit is not symmetric: the fit of A to B is not the same as the fit of B to A. In addition, the trace of E'E depends of the magnitude of the elements in A and B as well as on their fit to each other, so that values for the least squares measure based on matrices with very different norms are not directly comparable. Schönemann and Carroll ('70) defined a "normalized symmetric error," still based on the matrix E, which is symmetric but does not solve the problem of non-comparable norms. For this reason, Lingoes and Schönemann (in press) advocate norming each matrix by adjusting the terms to have unit variance before fitting. The normalized symmetric error calculated for matrices which have first been normed is called S by Lingoes and Schönemann. S must lie in the interval 0 to 1 and, given matrices of the same dimensions and rank, S should be more nearly comparable for all matrix fits, regardless of differences in the norms of the original matrices. The criterion S, which has a value of zero only when A and B can be made to superimpose exactly, is the measure of congruence used in the present study.

The distances

The Yanomama Indians of northern Brazil and southern Venezuela live in approximately 100–150 villages ranging in size from 40 to 250 (Chagnon, '70). The 19 villages on which the present study is based occupy an area about 150 miles (east-west) by 200 miles (north-south); genetic distance data for them have appeared in Ward ('72) and anthropometric data and distances are taken from Spielman, da Rocha, Weitkamp, Ward, Neel, and Chagnon ('72). Most of the data on genetic variation (the "marker gene" data: allele frequencies for blood groups, erythrocyte enzymes, and serum protein types) may be found in Gershowitz, Layrisse, Layrisse, Neel, Chagnon and Ayres ('72), Weitkamp, Arends, Gallango, Neel, Schultz, and Shreffler ('72), and Weitkamp and Neel ('72); additional allele frequencies are available (Gershowitz et al., unpublished; Tanis et al., unpublished).

Genetic distances, the kind called "G distances" by Kidd and Sgaramella-Zonta ('71), were derived from the allele frequencies by the method of Cavalli-Sforza and Edwards ('67). This technique uses the angular transformation (with an approximation of chord to arc length) to stabilize multinomial variances. For a substantial fraction of the loci, allele frequencies in at least one of the 19 villages fall outside the range (0.05 to 0.95) where the transformation is most effective. The exclusion of these loci would have meant an enormous loss of data, so even allele frequencies outside this preferred range were retained. The course adopted follows the established precedent set by the inventors of the method, Cavalli-Sforza and Edwards ('67), and by Ward and Neel ('70), and Friedlaender et al. ('71).

The geographic distances (in arbitrary units equal to approximately 100 km) are taken along straight lines connecting the villages on the map in figure 1 of Spielman et al. ('72). Large regions shown on the map have never been surveyed. Consequently geographic distances are not very precise, although the relative magnitudes are probably reliable. Because the degree of contact desired by villages partly determines their proximity, the geographic distance may also be taken as a rough inverse indication of inter-village exchange of goods and members. The pair-wise genetic and geographic distances are given in table 1.

In addition to the basic distances listed, a derivative set was obtained from the marker gene data. Since blood samples were taken from children who were not measured, and because occasional other

TABLE 1

Distances [1] between Yanomama villages: genetic distances below diagonal, geographic distances above. Underlining indicates distances between villages used in 7-population analysis

Villages	03AB	03C	03D	03KP	03LMN	03RS	03T	08E	08F	08K	08XY	11ABC	11D	11HI	11LQ	11S	11T	11U	11X
03AB		0.09	0.28	1.52	1.49	1.69	1.72	2.53	2.44	0.82	0.99	1.00	0.97	0.52	0.35	0.98	0.82	0.86	0.88
03C	0.33		0.20	1.61	1.58	1.75	1.78	2.59	2.50	0.90	1.08	1.08	1.06	0.47	0.39	1.07	0.89	0.94	0.97
03D	0.46	0.35		1.78	1.75	1.84	1.87	2.76	2.67	1.10	1.27	1.26	1.24	0.34	0.56	1.27	1.02	1.07	1.17
03KP	0.47	0.58	0.61		0.07	1.02	0.98	2.32	2.22	0.87	0.57	0.53	0.55	1.81	1.53	0.77	0.86	0.80	0.91
03LMN	0.54	0.69	0.72	0.24		0.95	0.91	2.38	2.28	0.88	0.57	0.52	0.54	1.77	1.52	0.79	0.81	0.75	0.92
03RS	0.41	0.57	0.61	0.26	0.29		0.07	3.30	3.20	1.51	1.26	1.15	1.14	1.68	1.90	1.52	0.90	0.88	1.61
03T	0.43	0.55	0.53	0.40	0.41	0.30		3.27	3.17	1.50	1.24	1.13	1.13	1.72	1.92	1.51	0.92	0.89	1.60
08E	0.45	0.55	0.43	0.44	0.53	0.45	0.43		0.10	1.95	2.10	2.20	2.22	3.06	2.21	1.86	2.66	2.64	1.84
08F	0.42	0.50	0.45	0.43	0.49	0.37	0.41	0.29		1.85	2.00	2.10	2.12	2.96	2.12	1.76	2.56	2.53	1.74
08K	0.40	0.46	0.47	0.32	0.41	0.38	0.47	0.42	0.33		0.30	0.39	0.38	1.27	0.70	0.16	0.72	0.71	0.11
08XY	0.44	0.60	0.59	0.28	0.28	0.32	0.40	0.35	0.36	0.31		0.11	0.12	1.37	0.96	0.27	0.59	0.55	0.37
11ABC	0.44	0.57	0.55	0.35	0.41	0.36	0.44	0.39	0.38	0.26	0.24		0.03	1.34	1.00	0.37	0.50	0.46	0.47
11D	0.51	0.59	0.56	0.38	0.45	0.46	0.52	0.39	0.41	0.30	0.27	0.20		1.31	0.98	0.38	0.48	0.44	0.47
11HI	0.41	0.24	0.39	0.61	0.73	0.58	0.53	0.59	0.55	0.52	0.63	0.62	0.67		0.86	1.43	0.98	1.03	1.36
11LQ	0.45	0.56	0.46	0.39	0.45	0.34	0.34	0.43	0.41	0.45	0.46	0.50	0.56	0.56		0.85	1.00	1.03	0.73
11S	0.48	0.57	0.51	0.46	0.47	0.45	0.49	0.42	0.37	0.28	0.32	0.31	0.33	0.60	0.53		0.80	0.78	0.14
11T	0.37	0.43	0.47	0.39	0.45	0.41	0.49	0.47	0.38	0.24	0.40	0.33	0.39	0.53	0.48	0.39		0.06	0.83
11U	0.33	0.46	0.48	0.44	0.48	0.43	0.48	0.46	0.34	0.28	0.41	0.40	0.47	0.52	0.45	0.36	0.22		0.81
11X	0.54	0.67	0.65	0.46	0.48	0.52	0.63	0.51	0.45	0.36	0.42	0.46	0.43	0.75	0.55	0.33	0.47	0.40	

[1] Genetic (or marker gene) distances are generalized distances representing composite difference in allele frequencies for 11 loci. The computation is described by Cavalli-Sforza and Edwards ('67). The geographic distances are from the map of Spielman et al. ('72).

TABLE 2

Distances between Yanomama villages: SFA [1] distances below diagonal, anthropometric distances above (see text). Underlining indicates distances between villages used in 7-population analysis

Villages	03AB	03C	03D	03KP	03LMN	03RS	03T	08E	08F	08K	08XY	11ABC	11D	11HI	11LQ	11S	11T	11U	11X
03AB		0.96	1.47	2.95	2.84	2.53	2.89	5.07	4.88	2.03	2.40	3.09	2.93	1.96	1.48	1.75	2.81	2.97	2.18
03C	0.35		0.88	3.18	3.39	2.58	3.01	5.71	5.60	2.25	2.77	3.15	3.24	1.92	1.65	1.81	3.00	3.18	2.57
03D	0.44	0.38		3.09	3.48	2.60	3.21	5.83	5.70	2.32	2.78	2.86	3.13	2.18	1.79	1.80	3.02	2.93	2.63
03KP	0.47	0.56	0.59		1.58	2.35	2.77	5.19	5.51	2.47	2.69	2.95	3.33	2.94	2.76	2.37	2.60	2.33	1.94
03LMN	0.52	0.65	0.67	0.30		2.53	2.70	4.55	4.75	2.42	2.40	3.19	3.11	3.05	2.87	2.53	2.49	2.35	2.09
03RS	0.57	0.76	0.80	0.50	0.43		1.20	4.72	5.21	2.58	3.05	3.64	3.69	2.89	2.78	2.36	2.78	2.82	2.64
03T	0.65	0.69	0.69	0.52	0.60	0.68		4.87	5.38	2.94	3.44	4.18	4.10	3.17	3.09	2.88	3.24	3.32	2.95
08E	0.69	0.74	0.63	0.61	0.65	0.88	0.79		2.17	5.43	4.79	5.73	5.19	5.98	5.76	5.43	5.77	5.71	5.32
08F	0.34	0.46	0.46	0.43	0.41	0.57	0.66	0.56		5.40	4.82	5.90	5.19	5.76	5.52	5.47	5.87	5.56	5.44
08K	0.47	0.53	0.43	0.38	0.43	0.63	0.72	0.53	0.34		1.76	2.26	1.99	1.21	1.22	1.21	1.47	1.51	1.33
08XY	0.47	0.57	0.53	0.28	0.33	0.61	0.53	0.51	0.35	0.38		1.59	1.11	2.46	2.14	1.47	1.82	2.35	1.92
11ABC	0.50	0.57	0.55	0.31	0.39	0.58	0.61	0.68	0.39	0.38	0.32		1.27	2.99	2.59	2.02	2.35	2.58	2.28
11D	0.53	0.58	0.54	0.32	0.41	0.66	0.63	0.53	0.37	0.34	0.25	0.21		2.83	2.42	2.06	2.18	2.52	2.23
11HI	0.34	0.26	0.43	0.55	0.65	0.75	0.65	0.73	0.49	0.56	0.53	0.62	0.62		1.26	1.63	2.20	2.16	2.13
11LQ	0.40	0.49	0.44	0.40	0.48	0.55	0.60	0.62	0.41	0.34	0.44	0.53	0.54	0.44		1.29	2.03	2.09	1.45
11S	0.40	0.49	0.42	0.50	0.48	0.67	0.72	0.61	0.32	0.38	0.42	0.38	0.36	0.52	0.53		1.52	2.00	1.51
11T	0.42	0.39	0.47	0.36	0.48	0.67	0.71	0.70	0.41	0.34	0.45	0.39	0.43	0.45	0.44	0.44		1.69	1.75
11U	0.31	0.46	0.48	0.42	0.47	0.61	0.75	0.71	0.31	0.32	0.45	0.41	0.44	0.48	0.43	0.36	0.26		1.91
11X	0.52	0.68	0.62	0.56	0.55	0.71	0.86	0.61	0.50	0.42	0.52	0.53	0.45	0.69	0.62	0.37	0.56	0.43	

[1] SFA distances are based on marker genes in measured subjects only. See also footnote to table 1.

individuals were bled or measured but not both, a discrepancy between the genetic and anthropometric results could be due in part to non-identity of the samples. The allele frequencies were therefore estimated again after the total marker gene or "serological" sample was reduced to a subset consisting of only those individuals who were also measured. The distances based on these frequencies are called "SFA": "Serological For Anthropometrics." They are given in table 2, which also contains the anthropometric (Mahalanobis) distances. There remain 49 individuals, distributed through 13 villages, who were measured but not bled. The effect of ignoring these few is presumed to be small.

In principle it is possible to compare the historical dispersion process with the divergence observed in biological variables; but in the case of the Yanomama, the historical relationships are not known over the large geographical areas covered by the 19 villages represented. Although within restricted geographical regions the recent history of some villages is known (Chagnon, '66), it was decided that there was no way to choose among the various possible evolutionary relationships of the major geographically defined groups (Chagnon, '66, '70).

With the possible exception of a documented genetic contribution by non-Yanomama neighbors to a village not included in the present study (Chagnon et al., '70), there is no indication that the Yanomama have a heterogeneous origin, at least in the last six or seven generations. We presume therefore that we are dealing with a process of dispersion from a relatively homogeneous origin, like the situation described in the introduction.

RESULTS

The methods for comparing anthropometric and marker gene differentiation differ sufficiently in the 7-population and 19-population cases to require separate presentation of the results.

Comparisons using 19 populations

For contrast with the new technique presented below, we first give a comparison of distance tables by one of the customary methods. The correlation coefficient for the $19 \times 18/2 = 171$ entries in each triangular distance table was calculated for each pair of tables; we use the Spearman rank correlation since our interest is restricted to association of rank, not necessarily linear association. The correlation found in this way for anthropometric distance and marker gene distance is small: $r = 0.19$. (As indicated above, there must be substantially fewer than $N - 2 = 169$ degrees of freedom for this test, so its significance is doubtful. A correlation of 0.19 requires about 105 degrees of freedom for significance at the 0.05 level.)

For the kind of cluster comparison described above, we must obtain the probability density of path lengths for a particular reference set. Genetic and marker gene distances, viewed as independent or causal variables, are the reference cases. The problem then is to estimate distributions, each of which consists, for the case of 19 populations, of more than 6.3×10^{18} values (the number of different nets connecting 19 points).

Random networks

Since it is impossible to examine any appreciable fraction of such a large number of nets, it was necessary to represent the total with a sample of 1,000, drawn so that every one of the possible networks has equal probability of inclusion at each draw. We have followed a suggestion attributed to Cavalli-Sforza by Kidd and Sgaramella-Zonta ('71), and constructed a net by adding branches sequentially, repeating the process with new random numbers 1,000 times. Figure 2 illustrates the procedure. At each step indicated in the figure by an arrow, the branch to which the next population will be added is determined ("equiprobably") by drawing a random number from a uniform distribution; i.e., the next population may be added to any one of the pre-existing branches with equal probability.

Following the method outlined earlier we now evaluate the fit of each of the 1,000 nets to a given set of data, the reference set, using an algorithm due to Edwards (unpublished) and described briefly by Kidd and Sgaramella-Zonta ('71). By this technique one sample distribution of path lengths was obtained for the

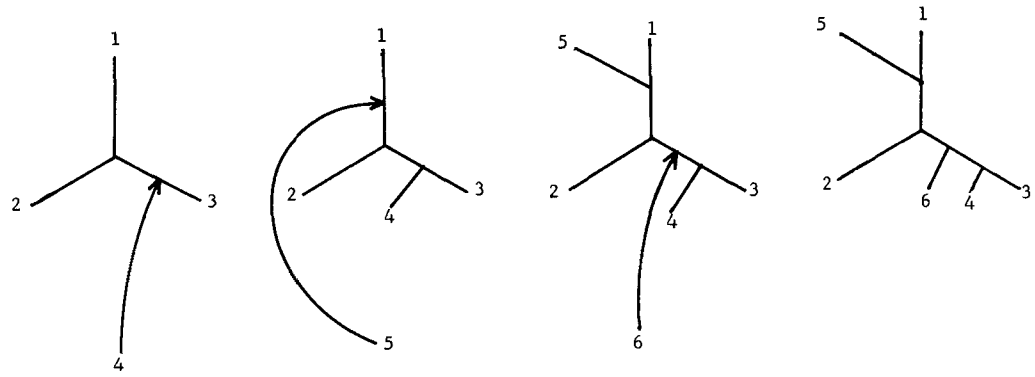


Fig. 2 Construction of random network: illustration for six populations. Successive populations added to branch chosen at random until network contains desired number of populations.

marker gene distances as a reference set, and another distribution for the geographic distances.

Preliminary experience with the much smaller distributions of path lengths for all possible nets of seven populations (Kidd and Sgaramella-Zonta, '71: 249) suggested that the distributions might be approximately Normal. When tested, the distributions of 1,000 randomly generated nets for 19 Yanomama villages showed only slight departures from Normality (Spielman, '71). We therefore infer that to a good approximation in these data, the properties of the Normal distribution will apply to the path lengths of randomly generated sets.

The best (i.e., shortest) nets for each set of biological data were sought by a combination of algorithmic and intuitive techniques. For the former, we follow Edwards and Cavalli-Sforza ('65) in generating a plausible initial candidate by cluster analysis and using Edwards' program to rearrange the relationships around that segment until a new network is found which has no zero-length internal segments. This is the "best" network the algorithm can generate, given the input candidate.

In addition, the best networks identified in the exhaustive 7-population treatment (below) have proved a source of excellent suggestions for candidates, when expanded to 19 populations. In general the strategy followed has been to infer the structure of the topology relating the major branches from the exhaustive treatment. These often differ from the major splits derived by the clustering technique, presumably because the latter's sequential splitting permits optimal clustering only for each split considered separately, and not for the cluster structure considered as a whole (Edwards and Cavalli-Sforza, '65). After the basic relationships are defined, the best branching structure for relatively closely related groups is found by trying several suggested by the cluster analysis.

By a combination of such techniques we identify for each set of biological data a "best net found." When compared to the distributions of randomly generated nets for the same kind of data, these best nets found have path lengths ranging from 5.7 to 7.5 standard deviations below the mean path length for the 1,000 random nets. To the extent that the distributions are Normal, we are therefore justified in arguing from the properties of the Normal distribution indicated in table 3, that these "best networks found" are among the best 10^{-8} to 10^{-12} of all possible networks. For the 19-population treatment, we indicate in this way the degree to which the best net found is a good representation of the data. While the best 10^{-12} part of a distribution is very small by conventional standards, in a distribution of 10^{18} nets it is composed of 10^{6} (a million).

The comparison of these networks with the distribution of path lengths for (1) the geographic distances, and (2) the marker gene distances, is given in table 3. By

TABLE 3

19-population comparisons: cluster-structure correspondence between geographic and various biological data expressed as difference (in standard deviation units) between path length for best net of one kind and mean path length of 1,000 random nets for the other (reference set).[1] *"SFA" distances are based on marker genes in measured subjects only. All entries are standard deviations below the mean*

Best net for	Reference set		
	Geographic distance	Marker gene distance	Anthropometric distance
Marker gene	4.64	—	4.83
SFA	7.45	5.47	5.54
Anthropometric	5.58	4.30	—

[1] The probability of obtaining by chance alone a network whose path length is x standard deviations below the mean for the reference set is approximated by the fraction of the Normal distribution x or more standard deviations less than the mean:

x	Cumulative normal distribution
4.0	3.2×10^{-5}
4.5	3.4×10^{-6}
5.0	2.9×10^{-7}
5.5	1.9×10^{-8}
6.0	1.0×10^{-9}
7.0	7.8×10^{-11}

this technique, the biological distances with the highest correspondence to geographic distribution are SFA distances; the poorest correspondence with the map is shown by marker gene distances. The correspondence of anthropometric and SFA data with the marker gene distances is also substantial. Only one of these comparisons between different kinds of data indicates a fit as good as that of the best nets found for a given set of data (compare values in table 3 with those given in the preceding paragraph). The implication is nevertheless that the best anthropometric net is a good representation of the marker gene distance relationships, and both are excellent representations of the geographic relationships. Given the prior doubts described earlier, and the known contribution of measurement error (Spielman et al., '72), the finding that the best anthropometric net found is among the best 10^{-5} of the possible marker gene nets indicates significant correspondence between these two kinds of biological divergence. In addition, as anticipated, the best SFA net found yields a better fit (it is among the best 2×10^{-8} part of the distribution of anthropometric nets) than the best marker gene net found (best 7×10^{-7} part); the appropriate results are in the last column of table 3.

The strength of the conclusions obviously depends on our confidence that one of a very few best networks has been identified for the biological variables, but the size of the distribution for 19 populations implies that a relatively large number of nets have small path lengths more than four or six or even eight standard deviations below the mean net length, if the distributions are truly Normal. For this reason, the corresponding analysis with only seven populations has also been carried out.

Seven-population analysis

The seven villages were selected to represent all the major geographic regions, covering the entire distribution of the Yanomama villages so far sampled. Within each such grouping the village with the largest anthropometric sample was chosen. The possibility of pooling villages to represent a region was rejected because the results would be less comparable with those of the 19-population analysis, and to preserve the culturally defined population unit. The distances and villages constituting this sample are indicated in tables 1 and 2.

As in the case of 19-population comparisons, the rank correlation coefficient for $7 \times 6/2 = 21$ entries in the tables of anthropometric and marker gene dis-

tances — the appropriate 21 entries from the distance tables above — was calculated. The value of r_s is -0.246, which of course does not indicate positive association, and would not be significantly different from zero, even on 19 degrees of freedom (which is surely too many). Now however, since it is possible to evaluate the net length for all the 945 possible networks connecting seven populations, we can select the single net which is truly best — the minimum length network. We might proceed in analogy with the 19-population case; select the best net for one set of variables, and evaluate the representation, or net length, of that net when used for the other distances. In some cases, the result of this kind of comparison is easily interpreted: for example, the best (shortest) net for the marker gene data is also the best net for the geographic distances.

It is possible, however, to make better use of the exhaustively evaluated sets of networks. Our interest is not really confined to the single best net for each set of variables. There are compelling reasons for abandoning the notion that the relationships embodied in a set of distances may be compared using a *single* best network. Suppose, for example, that the first and second best networks for the anthropometric data are not among the best ten as representatives of the marker data, but that the third best from the anthropometric data is second best for the marker data. Clearly this situation indicates some correspondence in the best 0.5% of the possible networks, which would be ignored when only the best network is considered. This example suggests that we must examine the distribution of networks in common among some best fraction of the entire list.

Comparisons in the totally enumerated case

The following procedure has been developed to compare relationships among the same seven villages based on different sets of variables. First the path lengths for the 945 nets evaluated on each set of data are listed in order of increasing magnitude. The comparison of some fixed fraction of the lists, say the best 50 or about 5%, may be put rigorously as the question: of the nets constituting the first 5% in one list, are more found in the first 5% of the second list than would be expected by chance alone, i.e., if the second list were randomly (in the sense defined by the distribution given below) ordered with respect to the first? Thus the step corresponding to the evaluation of the distribution of net lengths for 19 populations becomes for seven populations the calculation of the probability (under the null hypothesis of no association) of x or more nets in common among the first 50 in two such lists of $N = 945$.

After some initial misplaced optimism about an analytic solution for the probability density of this "best 50" statistic, it has become clear that a complex form of correlation exists within each of the two lists, making an analytic solution unlikely. If a particular net appears in common along the best 50, other nets which are (intuitively speaking) similar, are more likely also to appear than by chance alone, even in the absence of underlying cluster similarity. Although an analytic solution would of course be preferable, I have abandoned it in favor of computer simulation of the distribution of the best-50 statistic under the null hypothesis of no cluster correspondence.

In the simulation, seven populations were given coordinates on each of six axes, using the uniform random number generator FRAND to assign locations in the unit hypercube. Two hundred such sets were constructed. For each, the 945 possible networks were evaluated and ranked. From the 200 sets of data on seven "villages," 100 random pairs were formed and tested to give a sample of the distribution for the best-50 criterion. Table 4 shows the results. Among the best 50 nets in two such lists, 16 or more in common are encountered twice in this set of 100 trials, 21 or more once in 100 trials. With this distribution, which is at best a small sample, the 5% level of significance is put at approximately 14 or more in common.

These significance levels may be compared with the results from the actual data in table 5. All the comparisons except that between anthropometric and geographic data appear significant — i.e., most would be expected to occur by chance

TABLE 4

Simulation results for "best 50" statistic. Cluster-structure similarity between 100 simulated pairs of sets of data, expressed as number of nets in common among best 50 (out of 945 possible). Each set of data consists of seven populations given six coordinate values chosen randomly (i.e., from the uniform distribution) in the range 0 to 1.0

No. of nets in common among best 50	Frequency
0	0.56
1	0.06
2	0.06
3	0.04
4	0.09
5	0.04
6	0.03
7	0
8	0.01
9	0.01
10	0.01
11	0.02
12	0.02
13	0
14	0.03
15	0
16	0.01
21	0.01
	1.00

TABLE 5

7-population comparisons: cluster-structure similarity between geographic and various biological data expressed as number of nets in common among the best 50 out of 945 (the total possible). "SFA" designates marker gene data for measured subjects only; statistical significance may be inferred by comparison with table 4

	Geographic	Marker gene	SFA
Marker gene	30		
SFA	27	31	
Anthropometric	7	16	17

alone with frequency less than 0.01. In particular, the cluster-structure correspondence of genetic data with map distances is significant well beyond the 0.01 level. Of the significant associations, that of marker gene and anthropometric data appears the weakest, but even it would be conventionally labeled statistically significant ($P \leq 0.02$). Some reduction in significance would of course be needed to form a multiple comparisons type of overall significance level. It is also possible that other definitions of the null hypothesis represented by table 4, e.g., the scale-matched random sets generated below for tests of congruence, would yield slightly different probabilities. For lack of computer time, this possibility has not yet been explored.

The 7-population treatment does not reproduce perfectly the results for 19 populations; in particular, although geographic relationship is more nearly approximated by anthropometric than by marker gene distance in the former (table 3), the opposite is true for the latter (table 5). It seemed possible from the outset that a small sub-sample (7 villages) might fail through sampling errors alone to reproduce the properties of the larger group. Among the 19 populations are two related villages, 08E and 08F, both located at the northern extreme of the Yanomama territory, and found to be the two most divergent from the overall anthropometric means by Spielman et al. ('72). In the 7-population treatment, however, only 08F appears, representing the area where both are located. When the composition of the 7-population sample was altered to include both 08E and 08F (11X was removed), the relationships between different sets of variables reproduced the results for 19 populations (Spielman, '71). These results confirm that the discrepancy between the 7- and 19-population analyses is due in part to the inevitable failure of the smaller sample to represent the larger perfectly.

Another discrepancy in the 7-population treatment requires comment. The 50 best nets for SFA include only 31 from the best 50 for complete marker gene data (table 5). As described in more detail in Spielman ('71), the two villages (11T, 11X) whose allele frequencies are changed the most by restricting the marker gene sample to measured individuals, are also among the three which have the smallest samples in SFA. As a result, the standard errors of the allele frequency estimates for some loci are very high (0.08 to 0.11 for the Lewis and Kidd systems, for example). The imprecision of such estimates, inevitable when the population unit is the natural village, accounts for the discrepancy between marker gene and SFA networks.

In their instructive review, Kidd and

Sgaramella-Zonta ('71) distinguished between "additive" and "spatial" models of evolutionary divergence, corresponding to two different methods for estimating trees (networks). Their least-squares method presupposes the additive model, which assumes that the amount of evolution separating two groups equals the sum of the evolution from each to their common ancestor. Estimation by the criterion of minimum path length corresponds to their spatial model. Kidd and Sgaramella-Zonta prefer least-squares estimation on the pragmatic grounds that it requires less computer time; they apparently feel that conceptually neither approach is clearly preferable. Elsewhere, Kidd, Astolfi, and Cavalli-Sforza (in press) conclude from simulation that least-squares estimation is marginally better at identifying the "true" phylogeny.

In Kidd and Sgaramella-Zonta's data, the least-squares solutions for *most* of the 945 nets possible with seven populations contain at least one negative segment. Consequently, the authors ('71; 240) reject those nets as inadequate representations of the data. If least-squares estimation had been applied to the Yanomama data, presumably many of the best 50 nets for each data set would have contained negative segments. Rejecting these nets would have complicated enormously the exhaustive comparison based on the first 50 in two ordered lists. For this reason, networks were evaluated by the minimum path method instead of least-squares.

Tests of congruence: problems of scale and dimensionality

Just as a value for the best-50 statistic for cluster-similarity must be evaluated against a distribution, the value of the normalized symmetric error (S), describing the fit of two normed coordinate matrices, is only large or small in the context of a distribution corresponding to some null hypothesis of no congruence. In the application of this statistic, two difficulties arose. I shall call them the problems of (1) scale and (2) dimensionality.

The first attempt to generate a null distribution for S used the data sets simulated earlier in a unit hypercube of 6 dimensions. Recall that along each axis the distribution chosen was uniform, and that only the range of 0 to +1.0 was allowed. It quickly became clear that the resulting distribution of points did not simulate the real data well: in the anthropometric data, for example, some dimensions span an order of magnitude more than others, while in the hypercube, all dimensions tended to be homogeneous. It therefore does not seem likely that all comparisons can be referred to a single null distribution. To test this assertion, separate null distributions, randomly generated as described below, were tested for homogeneity of the mean value of S, and shown to be heterogeneous and different in mean from the null distribution generated in a hypercube. It follows that the distribution of S for data sets randomly generated in the hypercube (or any such single reference distribution) is inappropriate for metric comparisons with the real ones, since the real data incorporated gross discrepancies in scale.

For each kind of data listed in table 5, 120 new random sets were generated. These were constructed so that the average of the 120 sets matched the real set in mean and variance (±5%) along each dimension, with the distribution in each dimension approximately normal. Now the S obtained by fitting, say, the real anthropometric and real marker gene data, could be compared appropriately with the distribution of S values from 120 such fits with data sets matched in scale to the real ones.

The comparisons of biological data with geographic relationships brought out the problem of dimensionality. Although geographic distances lie only in a plane, the marker gene, SFA and anthropometric data include substantial variation in dimensions beyond the first two. It is clearly futile to seek a good fit (congruence) to points in a plane starting with points which vary substantially in more than two dimensions. For the test of congruence with geographic data, I reluctantly decided to discard the variation in all but the two most variable dimensions. The projections of village positions on the first two principal components of the *between*-villages covariance matrix were used for the comparisons with map relationships. Appropriate two-dimensional random sets

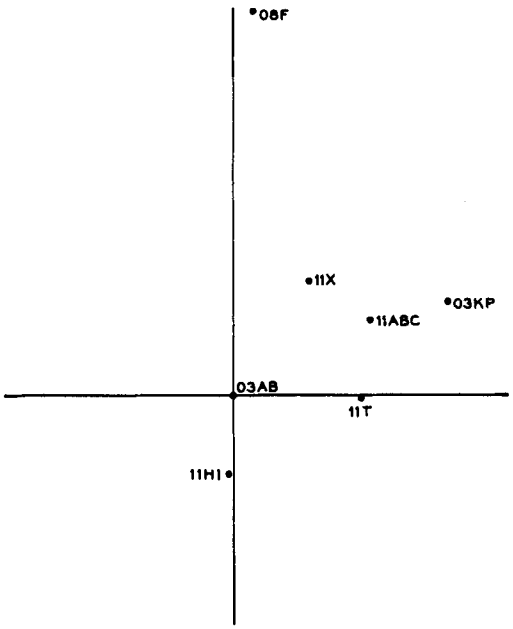


Figure 3A

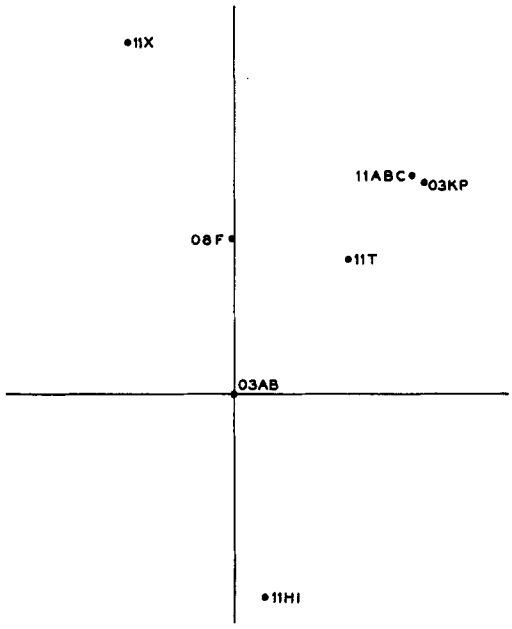


Figure 3B

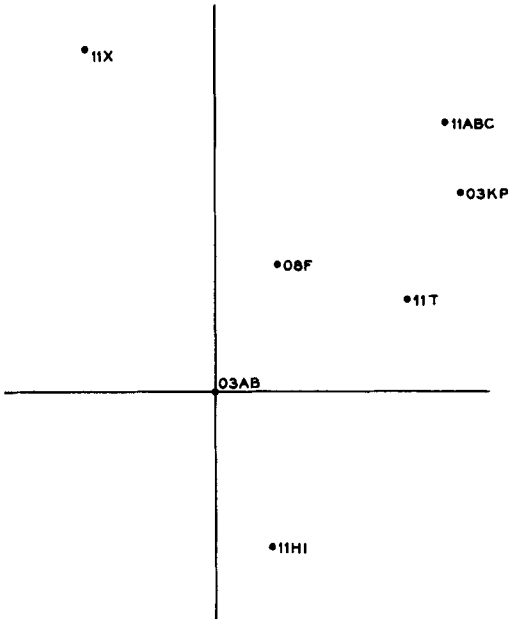


Figure 3C

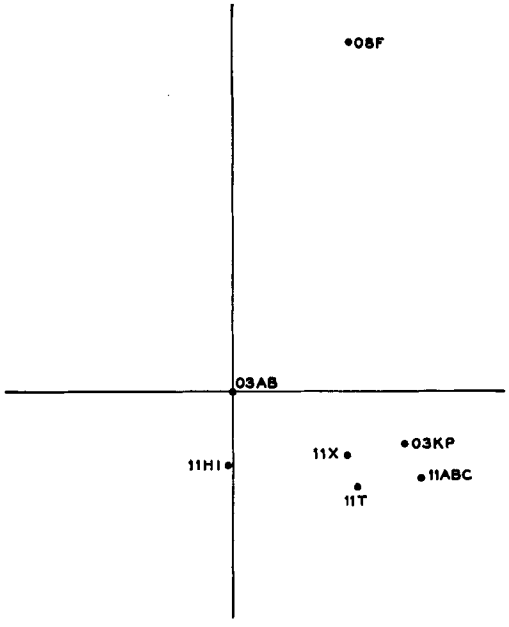


Figure 3D

Fig. 3 The relative positions of seven Yanomama villages plotted on the first two principal components of the between-groups covariance matrix for each of the four sets of data. (A) Geographic (identical to map relationships); (B) Marker gene; (C) SFA (marker gene data for measured subjects only); (D) Anthropometric.

of data were matched to the real ones as above. (The first two principal components account for 64, 67, and 78% of the between-groups variance in the marker gene, SFA, and anthropometric data respectively.) Fig. 3 (A, B, C, and D) shows the relative positions of the seven groups on the first two principal components of each set of data.

The results of applying Schönemann and Carroll's method with the S criterion are given in table 6. As in table 5 for cluster-structures, the similarity of marker gene and its derivative SFA data dominates, with a probability by chance alone less than 0.01. But in the test of congruence, marker gene and SFA seem to fit the geographic data rather poorly ($P > 0.10$), reversing the situation seen with networks (table 5). More startling still is the degree of congruence demonstrated in table 6 for anthropometric and geographic data ($P < 0.01$), the weakest of all correspondences in the comparison of cluster-structures. Whatever the ultimate interpretation of the differences from cluster-similarity, the values for S in table 6 show clearly that matched random sets provide reasonable null distributions against which to test observed values of S. However, the pattern of significant values does not necessarily parallel that obtained by the network-comparison method.

DISCUSSION

The present study addresses the question: Is the biological dispersion process, represented by the marker gene and anthropometric data, similar in these two kinds of traits, and do the resulting patterns of divergence of villages reflect geographic and historical separation? A related goal is to formulate appropriate objective measures of association between different sets of measurements on a single set of villages. Such a measure has been devised for similarity of cluster structure and used to demonstrate correspondence between various kinds of dispersion. The method of Schönemann and Carroll for fitting two matrices to each other has been elaborated into a test for significant degree of congruence in two sets of data. These two criteria for correspondence are not equivalent; in particular, similarity of cluster-structure is possible without significant congruence. In addition, as shown for cluster-similarity in the data presented here, and inferred for congruence, the answers to these questions depend in part on the choice of villages for the comparisons.

It is important to note, in addition, that the techniques described may give different results from the more conventional use of correlation coefficients, as indicated earlier for marker gene and anthropometric networks. To permit a compari-

TABLE 6

7-population comparisons. Multi-dimensional (and two-dimensional) congruence between geographic and various biological data, expressed as normalized symmetric error (S; Lingoes and Schönemann, in press) obtained by fitting normed data matrices. First entry is value of S; entry in parentheses is fraction of randomly generated S that small or smaller.

	Geographic	Marker gene	SFA
Marker gene	0.5768 (0.13)		
SFA	0.6618 (0.21)	0.0445 (<0.01)	
Anthropometric	0.1435 (<0.01)	0.3795 (0.13)	0.4021 (0.08)

TABLE 7

Spearman rank correlations of distance table entries based on four kinds of variables. Below diagonal: 7-population treatment. Above diagonal: 19-population treatment. ("SFA" designates marker gene data for measured subjects only)

	Geographic	Marker gene	SFA	Anthropometric
Geographic		0.39	0.54	0.80
Marker gene	0.27		0.61	0.19
SFA	0.06	0.82		0.39
Anthropometric	0.73	−0.25	−0.32	

son of the two approaches for all the data used in the present study, table 7 gives the Spearman rank correlation coefficients for various pairs of variables, calculated from the entries in the triangular distance matrices. Since the entries in such matrices are not independent, the correct degrees of freedom for the comparisons are not known. Rather than comparing significance levels, which would therefore be of questionable accuracy for the correlation case, we simply indicate some outstanding discrepancies between the implications of tables 5, 6, and 7 (lower triangular matrix) for the 7-population analysis.

(1) The salient feature of the lower triangular matrix in table 7 is that the two most prominent positive associations parallel those in table 6. Both the correlation and Schönemann and Carroll's matrix fitting seem to detect the same correspondence of anthropometric with geographic data and the marker gene with SFA data, as against all other comparisons. (This is roughly true for the 19-populations correlations also, shown in the upper triangular matrix of table 7.)

(2) On the other hand, the anthropometric data show significant similarity in cluster structure with marker gene and SFA (table 5) and must be construed as showing a positive (though not significant) tendency to congruence (table 6). In the comparison using the correlation coefficient however (table 7), these data show a weak inverse relationship — if any. It thus appears that the Schönemann and Carroll matrix fitting method gives results not unequivocally like either the cluster-structure or the correlation results.

The correlation approach can only measure the correspondence between pair-wise village differences rather than the correspondence of the village measurements themselves, and the statistical problem of non-independence of pair-wise comparisons introduces additional difficulties in interpretation. Although the new measures of association avoid the problem of unspecifiable degrees of freedom encountered with the correlation coefficient, they are not without drawbacks. In particular, for the 19-population treatment, it is not obvious how to compare results for different numbers of populations. The best few nets in a Normal distribution of 1,000 cannot be as many standard deviations below the mean in net length as those in a distribution of 10^{18} with the same variance. On the other hand, it is not clear that the correlation technique copes any more successfully with the enormous number of possible relationships, some of which appear to be ignored in the transformation of multidimensional data to pair-wise distances.

Since the network technique and the matrix-fitting method test different kinds of correspondence, they may be expected to give different answers when applied to the same data. This kind of divergence has been seen in the comparison of tables 5 and 6. Although appreciable cluster similarity in the absence of congruence is not surprising, the apparent congruence of anthropometric and geographic data without cluster similarity is unexpected. It must be recalled, however, that the anthropometric data for the two tests are not identical. To preserve equality of dimensionality in the test of congruence, only the two axes of largest between-group variability were used. It is possible that the 20% of the variance thereby excluded, but present for network comparisons, greatly obscures in the latter case a basic similarity in cluster structure.

Should we expect comparable correspondence if the net technique is applied to data from other populations? The anthropometric data are subject to considerable measurement error (Spielman et al., '72), as are the geographic distances. In view of such imprecision and the sensitivity of the comparisons to various sampling errors as illustrated earlier, the highly significant associations demonstrated here might not have been expected. The Yanomama, however, may be a particularly favorable case for detecting correspondences. Compared to similar subdivided populations, the Yanomama villages show unusually large heterogeneity in gene frequencies, as measured by values of F_{ST} (Neel and Ward, '72). The analysis of the anthropometric data in Spielman ('73) indicates a corroborating homogeneity within villages. It is possible that the high degree of differentiation

among villages facilitates, or is even a prerequisite for, the demonstration of correspondence between anthropometric and marker gene data.

On the other hand most previous studies have also concluded that different variables or networks agree, to some subjectively acceptable degree. (For a review, see Friedlaender, '69). The contribution of the present analysis is an objective measure of correspondence. Schull ('72) has criticized the lack of quantification or precision in previous attempts to evaluate the correspondence of distance based on different variables. Noting that most previous studies have found different networks "to agree at least generally," he asks, "By what criterion could one reach a different conclusion. . . ?" The treatment developed here for the Yanomama data provides and answer.

CONCLUSIONS

At the outset, we gave grounds for suspecting that patterns of differentiation in anthropometrics and marker genes might not correspond well. This pessimism was not relieved by the known measurement error associated with the anthropometric data (Spielman et al., '72). The mobility of Indian villages (Chagnon et al., '70) led to similar doubts for correspondences with geographic relationships (Ward and Neel, '70). What then are the causes for the observed correspondence of anthropometric and marker gene data on villages, and for their agreement with the map?

In the first place, the distorting effects of village movements are presumably much less important for very distant villages than for villages in close proximity. Villages which are separated by only a few days' walk may change their distances and relative positions easily, obscuring the relationship of geographic distance to biological differentiation (Ward and Neel, '70). When villages are separated by hundreds of kilometers of jungle, as are the major village clusters used here, a few kilometers displacement does not alter relative distances appreciably. Movement on the scale reported by Chagnon et al. ('70) is quite unusual. Thus for the present data village movements are not expected to influence greatly any potential correspondence with map positions.

Secondly, aspects of tribal demography provide a plausible explanation for the biological and geographic correspondences observed. In an expanding population like the Yanomama, new villages arise by the splitting or fissioning of an older one. The members of the old village, however, are not randomly distributed between the fragments produced. The tendency for village splits to occur in a manner which preserves lineage integrity has been described by Neel ('67), who called the phenomenon "lineal effect." To the extent that members of one lineage are more similar to each other than to those of other lineages, it is likely therefore that each of the daughter-villages or immediate products of a split is more homogeneous than the parent group (Spielman, '73) in various measurable traits. It follows that descendants of one daughter-village are more similar to each other than to descendants of other villages produced by the split. The extent of differences between villages thus reflects the historical development and may be expected to do so in all variables (e.g., dermatoglyphic and linguistic) for which lineages might be relatively homogeneous, irrespective of any genetic determination of these traits. In this view, the correspondence of different systems of variables is seen as the consequence of their common dependence on the historical process; any features distributed non-uniformly by lineages may thereby become associated with village differences. To the extent that closely related Yanomama villages remain in geographic proximity, the same process would of course account for correspondence with the map. One of the goals of future work in this area is to specifiy in detail how cultural and demographic features determine the village relationships whose correspondences have been demonstrated here.

ACKNOWLEDGMENTS

I am grateful for critical advice from G. F. Estabrook, K. K. Kidd, J. W. MacCluer, J. V. Neel, W. J. Schull and C. F. Sing. R. M. Carroll kindly supplied the program used to test for congruence. E. A. Thompson provided the impetus for the simulations. R. H. Ward emphasized for me that networks and matrix-fitting

test for very different kinds of correspondence, and by his scrupulous scrutiny, identified errors of omission and commission. Those remaining are my responsibility alone.

LITERATURE CITED

Cavalli-Sforza, L. L., and A. W. F. Edwards 1964 Analysis of human evolution. Proc. XI Int. Cong. Genet., 3: 923–933.

——— 1967 Phylogenetic analysis: models and estimation procedures. Amer. J. Hum. Genet., *19*: 233–257.

Chagnon, N. A. 1966 Yanomamö Warfare, Social Organization and Marriage Alliances, (Ph.D. Thesis) University of Michigan, Ann Arbor.

——— 1970 The culture-ecology of shifting (pioneering) cultivation among the Yanomamö Indians. Proc. VIII Int. Cong. Anthropol. and Ethnol. Sciences, 3: 249–255.

Chagnon, N. A., J. V. Neel, L. R. Weitkamp, H. Gershowitz and M. Ayres 1970 The influence of cultural factors on the demography and pattern of gene flow from the Makiritare to the Yanomama Indians. Am. J. Phys. Anthrop., 32: 339–349.

Chai, C. K. 1967 Taiwan Aborigines. Harvard University Press, Cambridge.

Edwards, A. W. F., and L. L. Cavalli-Sforza 1965 A method for cluster analysis. Biometrics, *21*: 362–375.

——— 1963 The reconstruction of evolution. Ann. Hum. Genet., *27*: 104–105 (Abstract).

Edwards, A. W. F. 1971 Mathematical approaches to the study of human evolution. In: Mathematics in the Archaeological and Historical Sciences. F. R. Hodson, D. G. Kendall and P. Tăutu, eds. Edinburgh University Press, Edinburgh, pp. 347–355.

Friedlaender, J. S., L. Sgaramella-Zonta, K. K. Kidd, L. Y. C. Lai, P. Clark and R. J. Walsh 1971 Biological divergences in South-Central Bougainville: An analysis of blood polymorphism gene frequencies and anthropometric measurements utilizing tree models, and a comparison of these variables with linguistic, geographic, and migrational "distances." Amer. J. Hum. Genet., 23: 253–270.

Friedlaender, J. S. 1969 Biological Divergences over Population Boundaries in South-Central Bougainville. (Ph.D. Thesis) Harvard University.

Gershowitz, H., M. Layrisse, Z. Layrisse, J. V. Neel, N. Chagnon and M. Ayres 1972 The genetic structure of a tribal population, the Yanomama Indians. II. Eleven blood-group systems and the ABH-Le secretor traits. Ann. Hum. Genet., 35: 261–269.

Gower, J. C. 1971 An illustration of a new technique for comparing different distance analyses. Am. J. Phys. Anthrop., 35: 280–281.

Hiernaux, J. 1956 Analyse de la variation des charactères physiques humains en une région de l'Afrique centrale: Ruanda — Urundi et Kivu. Annales du Musée Royal du Congo Belge. Série en 8e. Sciences de l'homme, Vol. 3, Belgium, Tervuren.

Howells, W. W. 1966 Population distances: Biological, linguistic, geographical, and environmental. Current Anthropology, 7: 531–540.

Kidd, K. K., and L. A. Sgaramella-Zonta 1971 Phylogenetic analysis: Concepts and methods. Amer. J. Hum. Genet., 23: 235–252.

Kidd, K. K., P. Astolfi and L. L. Cavalli-Sforza Error in the reconstruction of evolutionary trees. In: Genetic Distance, (Ed, Crow, J. F.), in press.

Lingoes, J. C., and P. H. Schönemann Alternative measures of fit for the Schönemann-Carroll matrix fitting algorithm. Psychometrika, in press.

Mahalanobis, P. C., D. N. Majumdar and C. R. Rao 1949 Anthropometric survey of the United Provinces, 1941: A statistical study. Sankhya, 9: 89–324.

Majumdar, D. N., and C. R. Rao 1960 Race elements in Bengal: A quantitative study. Indian Statistical Institute, Asia Pub. House, London.

Neel, J. V. 1967 The genetic structure of primitive human populations. Jap. J. Hum. Genet., *12*: 1–16.

Neel, J. V., and R. H. Ward 1972 The genetic structure of a tribal population, the Yanomama Indians. VI. Analysis by F-statistics (including a comparison with the Makiritare and Xavante). Genetics, 72: 639–666.

Pollitzer, W. S. 1958 The Negroes of Charleston (S. C.): A study of hemoglobin types, serology, and morphology. Am. J. Phys. Anthrop., *16*: 241–263.

Prim, R. C. 1957 Shortest connection networks and some generalizations. Bell Syst. Techn. J., *36*: 1389–1401.

Sanghvi, L. D. 1953 Comparison of genetical and morphological methods for a study of biological differences. Am. J. Phys. Anthrop., *11*: 385–404.

Schönemann, P. H., and R. M. Carroll 1970 Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, 35: 245–255.

Schull, W. J. 1972 Primitive populations: Some contributions to the understanding of human population genetics. In: Proc. IV Int. Cong. Hum. Genet., Excerpta Medica, Amsterdam, pp. 112–123.

Sinnett, P., N. M. Blake, R. L. Kirk, L. Y. C. Lai and R. J. Walsh 1970 Blood, serum protein and enzyme groups among Enga-speaking people of the Western Highlands, New Guinea, with an estimate of genetic distance between clans. Archaeol. Phys. Anthropol. Oceania, 5: 236–252.

Spielman, R. S. 1971 Anthropometric and Genetic Differences among Yanomama Villages. (Ph.D. Thesis) University of Michigan, Ann Arbor.

——— 1973 Do the natives all look alike? Size and shape components of anthropometric differences among Yanomama Indian villages. Amer. Nat., *107*: 694–708.

Spielman, R. S., F. J. da Rocha, L. R. Weitkamp, R. H. Ward, J. V. Neel and N. A. Chagnon 1972. The genetic structure of a tribal population, the Yanomama Indians. VII. Anthropometric differences among Yanomama villages. Am. J. Phys. Anthrop., 37: 345–356.

Ward, R. H. 1972 The genetic structure of a tribal population, the Yanomama Indians. V. Comparisons of a series of genetic networks. Ann. Hum. Genet., 36: 21–43.

Ward, R. H., and J. V. Neel 1970 Gene frequencies and microdifferentiation among the Makiritare Indians. IV. A comparison of a genetic network with ethnohistory and migration matrices; a new index of genetic isolation. Amer. J. Hum. Genet., 22: 538–561.

Weitkamp, L. R., T. Arends, M. L. Gallango, J. V. Neel, J. Schultz and D. C. Shreffler 1972 The genetic structure of a tribal population, the Yanomama Indians. III. Seven serum protein systems. Ann. Hum. Genet., 35: 271–279.

Weitkamp, L. R., and J. V. Neel 1972 The genetic structure of a tribal population, the Yanomama Indians. IV. Eleven Erythrocyte enzymes and summary of protein variants. Ann. Hum. Genet., 35: 433–444.

Workman, P. L., and J. D. Niswander 1970 Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Amer. J. Hum. Genet., 22: 24–49.