AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 61:411–422 (1983)

# Intertribal Gene Flow Between the Ye'cuana and Yanomama: Genetic Analysis of an Admixed Village

JEFFREY C. LONG AND PETER E. SMOUSE
*Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109*

*KEY WORDS* Admixture, Multiallelic estimation, model evaluation

*ABSTRACT* Genetic exchange with a neighboring village of Ye'cuana Indians had introduced two alleles, $Di^a$ and $ACP^a$, into the Yanomama Indian village of Borabuk. After several generations, these alleles had reached frequencies of 0.08 and 0.10, respectively. These frequencies are puzzling because they are higher in Borabuk than in the Ye'cuana village from which they were derived. Single allele estimates of ancestral proportions obtained from either of these traits are biologically unrealistic and suggest that admixture is not a good explanation for genetic variation in Borabuk. Nevertheless, multiallelic admixture models are seen to produce credible estimates of ancestral proportions and to explain a large amount of allele frequency variation in Borabuk. When these results are compared with expectations derived from a formal pedigree analysis, good agreement is seen. Comparison of single allele estimates of ancestral proportions obtained from alleles at 11 loci, with multiallelic estimates obtained from the same 11 loci and with the pedigree-derived estimates, demonstrates the superiority of the multiallelic approach.

The Yanomama Indians of southern Venezuela and northwestern Brazil are one of the most remote aboriginal Amerindian populations still existing. They are also one of the few anthropological populations for which there exist reasonably good collections of both ethnological and biological data. These features have made the Yanomama the focus of a variety of studies of processes affecting biological and cultural variation in human populations at the tribal level of social and demographic organization (cf. Chagnon, 1968; Neel, 1978). Recent analyses have revealed the Yanomama to be quite distinct from surrounding tribes, both linguistically and culturally (Chagnon et al., 1970; Spielman et al., 1974). The genetic differences between the Yanomama and their neighbors have been demonstrated to parallel in magnitude the linguistic and cultural differences, suggesting that they have been isolated from other South American Indian tribes for a considerable period of time (Ward et al., 1975).

During the past century, there has been increased contact between the Yanomama and Ye'cuana (Makiritare), a neighboring Amerindian tribe for which good biological and cultural data have also been acquired (Ward and Neel, 1970; Arvello-Jimenez, 1971; Ward, 1972). Genetic exchange accompanying this increased contact has introduced into the Yanomama village of Borabuk two alleles that are typically absent from Yanomama villages but are present in polymorphic frequencies among the Ye'cuana. These alleles, the $Di^a$ variant of the Diego blood group system and the $ACP^a$ allele of the red cell acid phosphatase locus, had achieved frequencies in Borabuk of 0.08 and 0.10, respectively, by approximately 1970. These frequencies are puzzling because they are higher in Borabuk than in the Ye'cuana village whence they were derived. Chagnon et al. (1970) have explained these unusual results in terms of how social organization in the two tribes affected the survival probabilities of new alleles introduced into Borabuk by Ye'cuana immigrants.

Received January 19, 1982; accepted March 1, 1983.

Single allele estimates of admixture proportions obtained from either of these systems suggest that Borabuk has approximately 200% Ye'cuana ancestry, which is clearly impossible. Taken at face value, these results suggest that an admixture model explaining the genetic structure of Borabuk is inappropriate for the data. Nevertheless, it is of interest to determine whether a broader treatment of allele frequencies would yield a more credible reflection of the role of admixture in the evolutionary history of this population.

For this purpose, we have adopted a multiallelic approach to analysis of hybrid population allele frequencies and propose several alternative admixture models to account for the distribution of genetic traits observed in Borabuk. Our objective will be to find a model (or models) that is (are) consistent with the genetic marker data. This strategy differs from that of most admixture studies because it emphasizes the fit of the genetic marker data to the model, rather than focusing on estimates of ancestral proportions. The results from this procedure will then be compared with expectations derived from a formal pedigree analysis. Borabuk is unusual among admixed populations, because pedigree information exists. This information spans six generations of Borabuk residents and includes all known immigrants from the Ye'cuana population. To facilitate the comparison of the genetic marker analysis and the pedigree structure, an admixture measurement derived from kinship considerations will be developed. This measure can be viewed as the expected average result from a large number of genetic loci.

## ANALYTICAL FORMULATION

When a population arises as the consequence of admixture between two parental groups, the allele frequencies of the hybrid population will be a linear combination of the parental frequencies, providing that the frequencies are not subsequently altered by other evolutionary processes. The frequency of the $i^{th}$ allele in the hybrid population, $y_i$, will then be

$$y_i = \sum_{j=1}^{2} x_{ij}u_j \qquad (1)$$

where $x_{ij}$ denotes the frequency of the $i^{th}$ allele ($i = 1, \ldots, p$) in the $j^{th}$ parental group ($j = 1,2$) and the value of $u_j$ is equivalent to the proportionate contribution of the $j^{th}$ parental group to the hybrid population; by construction $u_1 + u_2 = 1.0$. The process of interpopulation admixture has the same expectation for all loci, and the same set of $u_j$ values are applicable to alleles at all loci. Consequently, equation (1) may be generalized to more than one allele:

$$\mathbf{y} = \mathbf{Xu} \qquad (2)$$

where $\mathbf{X} = \{x_{ij}\}$ is a $p \times 2$ matrix of parental population allele frequencies, $\mathbf{u}$ is a $2 \times 1$ vector composed of the proportionate contributions of the two parental populations, and $\mathbf{y}$ is a $p \times 1$ vector of the frequencies of the p alleles in the hybrid population.

### *Estimation procedure*

When admixture is the only process affecting the evolution of the parental and hybrid populations and allele frequencies for these populations are known without error, $\mathbf{u}$ may be determined exactly, using the frequency of a single allele drawn from any locus and the standard Bernstein formula (c.f. Cavalli–Sforza and Bodmer, 1971):

$$u_1 = \frac{y_1 - x_{12}}{x_{i1} - x_{i2}}; \quad u_2 = 1 - u_1. \qquad (3)$$

However, the allele frequencies in neither the hybrid population nor the parental populations will be known without error because of sampling processes (including genetic drift), uncertainty in ascertainment of true population histories, and perhaps selection. Consequently, $\mathbf{u}$ must be estimated. A complete solution of the problem of estimating ancestral proportions requires the simultaneous estimation of allele frequencies in the parental populations, allele frequencies in the mixed population, and $\mathbf{u}$. The efficiency of the estimate will improve when information from a large number of loci, including multiallelic loci, is used. Most current techniques available for the estimation of ancestral proportions do not fulfill these requirements in one way or another. The maximum likelihood (Krieger et al., 1965; Elston, 1971) and least squares methods (Roberts and Hiorns, 1962; 1965; Elston, 1971) consider parental population allele frequencies to be known without error. The weighted average method suggested by Cavalli–Sforza and Bodmer (1971) allows for error in ancestral frequencies but with a single exception (Szath-

mary and Reed, 1972) has not been extended to multiallelic loci.

We have investigated the complete solution to the problem (including multiallelic loci) using the method of maximum likelihood (cf. Li, 1976). The likelihood function for one locus, $l = 1, \ldots, L$, allowing for error in parental population frequencies, is proportional to

$$\prod_{j=1}^{2} \prod_{k=1}^{K} (P_{kj})^{n_{kj}} \prod_{k=1}^{K} (P_{kh})^{n_{kh}}. \quad (4)$$

where the locus contains $k = 1, \ldots, K$ phenotypic classes, $P_{kj}$ is the probability of the $k^{th}$ phenotype in the $j^{th}$ parent population, and $n_{kj}$ is the observed number of individuals in the $k^{th}$ phenotypic class in the $j^{th}$ parental group. $P_{kh}$ is the *expected* probability of the $k^{th}$ phenotype in the hybrid population and $n_{kh}$ is the observed number of individuals in that class. The probability of each phenotypic class is taken as the sum of probabilities of each genotype composing that class, assuming Hardy–Weinberg proportions. The *expected* probabilities of phenotypic classes in the hybrid population are generated from model expected allele frequencies, obtained according to $\mathbf{X}_l\mathbf{m}_l = \hat{\mathbf{y}}_l$, where $\mathbf{m}_l$ is the maximum likelihood estimate of $\mathbf{u}$ obtained from the $l^{th}$ locus.

Collectively, the parental population allele frequencies and $\mathbf{m}_l$ are the parameters, $\theta_r$, to be estimated from the likelihood function. It will be necessary to estimate $r = 1, \ldots, 2p + 1$ parameters. The maximum likelihood estimates of these parameters are obtained by setting the equations $\delta L/\delta\theta_r = 0$ and solving simultaneously, where L is the logarithm of the likelihood function. The maximum likelihood equations are given by the general formulae

$$\frac{\delta L}{\delta\theta_r} = \sum_{j=1}^{2} \sum_{k=1}^{K} \frac{n_{kj}}{P_{kj}} \frac{\delta P_{kj}}{\delta\theta_r} + \sum_{k=1}^{K} \frac{n_{kh}}{P_{kh}} \frac{\delta P_{kh}}{\delta\theta_r} = 0 \quad (5a)$$

with $r = 1, \ldots, 2p + 1$. The solutions are provided by standard iteration techniques (see Li, 1976), using the second partial derivatives

$$-\frac{\delta^2 L}{\delta\theta_r\delta\theta_s} = \sum_{j=1}^{2} \sum_{k=1}^{K} \left[\frac{1}{P_{kj}^2} \frac{\delta P_{kj}}{\delta\theta_r} \frac{\delta P_{kj}}{\delta\theta_s} - \frac{1}{P_{kj}} \frac{\delta^2 P_{kj}}{\delta\theta_r\delta\theta_s}\right] + \sum_{k=1}^{K} \left[\frac{1}{P_{kh}^2} \frac{\delta P_{kh}}{\delta\theta_r} \frac{\delta P_{kh}}{\delta\theta_s} - \frac{1}{P_{kh}} \frac{\delta^2 P_{kh}}{\delta\theta_r\delta\theta_s}\right] = I_{\theta_r\theta_s}. \quad (5b)$$

The $I_{\theta_r\theta_s}$ values are arranged in matrix form; the elements of the inverse of this matrix are the estimated variances and covariances of the estimated parameters. This general approach permits one to deal with any number of alleles per locus and any array of dominance relationships among them.

Investigation of this model reveals that estimates of parental population allele frequencies, their variances, and covariances are unaffected by their simultaneous estimation with admixture proportions. In the special case where $\mathbf{u}$ is estimated from a two allele locus, the solution of equation (5a) is identical to equation (3) where $\mathbf{y}_i$ and $\mathbf{X}_{ij}$ are taken as the maximum likelihood estimates of the respective allele frequencies; the variance of the estimate (m) is identical to that given by Cavalli–Sforza and Bodmer (1971:492).

This procedure can be extended to multilocus problems by taking the product of expression (4) over all loci and altering equations (5a) and (5b) accordingly. This is a very tedious and cumbersome operation. Instead, we recommend following the strategy of Cavalli–Sforza and Bodmer (1971) and Szathmary and Reed (1972); the essence of that strategy is to take a weighted average of $\mathbf{m}_l$ values, estimated locus by locus

$$\mathbf{m} = \frac{\sum_{l=1}^{L} w_l\mathbf{m}_l}{\sum_{l=1}^{L} w_l}, \quad (6)$$

where each weight, $w_l$, is the inverse of the variance of the admixture proportion, $\mathbf{m}_l$, estimated for the $l^{th}$ locus. Equation (6) can thus be rewritten in matrix form

$$\mathbf{m} = (\mathbf{X}^{*\prime}\,\mathbf{V}^{*-1}\,\mathbf{X}^{*})^{-1}\,\mathbf{X}^{*\prime}\mathbf{V}^{*-1}\,\mathbf{y}^{*}, \quad (7)$$

where the vectors $\mathbf{X}^*$ and $\mathbf{y}^*$ are simple alterations of the vector $\mathbf{y}$ and matrix $\mathbf{X}$ from equation (2). Letting $\mathbf{X}$ be represented by its column vectors $\mathbf{x}_1$, $\mathbf{x}_2$, the vector $\mathbf{X}^*$ is found by subtracting $\mathbf{x}_2$ from $\mathbf{x}_1$ (eg, $\mathbf{X}^* = \mathbf{x}_1 - \mathbf{x}_2$). Similarly, $\mathbf{y}^*$ is found by subtracting $\mathbf{x}_2$ from $\mathbf{y}$. The entries of $\mathbf{X}^*$ and $\mathbf{y}^*$ correspond to maximum likelihood estimates of $p - L$ alleles belonging to L allelic systems, where one arbitrarily chosen allele from each system has been discarded

from analysis. The choice of which allele to discard or which population to subtract will not affect the outcome. The matrix $\mathbf{V}^*$ is a block diagonal matrix,

$$\mathbf{V}^* = \begin{bmatrix} \mathbf{V}_1^* & 0 & \cdots & 0 \\ 0 & \mathbf{V}_2^* & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & \mathbf{V}_L^* \end{bmatrix} \tag{8}$$

where there is a block, $\mathbf{V}_l^*$, corresponding to each locus. These blocks will be scalars for diallelic loci and small matrices for multiallelic loci. Each block is of the form

$$\mathbf{V}_l^* = \mathbf{V}_{lh} + m_f^2\mathbf{V}_{l1} + (1 - m_l)^2\mathbf{V}_{l2} \tag{9}$$

where $\mathbf{V}_{lh}$ is the *expected* variance–covariance matrix of hybrid population allele frequencies (using $\hat{\mathbf{y}}_l = \mathbf{X}_l\mathbf{m}_l$), and $\mathbf{V}_{l1}$ and $\mathbf{V}_{l2}$ are the variance–covariance matrices at the $l^{th}$ locus in the parental populations.

If two allele loci are used and parental populations are considered to be known without error, equation (7) is algebraically identical with Elston's (1971) weighted least-squares method. If multiallelic loci are used and parental populations are known with error, equation (7) differs from Elston's weighted least-squares method only in that the expected variance–covariance matrix of hybrid population allele frequencies is used rather than the observed variance–covariance matrix. If multi- or diallelic loci are considered and parental population allele frequencies are assumed known without error, it can be shown that equation (7) is equivalent to a weighted average of maximum likelihood estimates taken locus by locus according to Elston's (1971) method. If parental population allele frequencies are estimated with more confidence (ie, larger sample sizes) than hybrid population allele frequencies, error in their estimation will have little effect on the estimation of $\mathbf{u}$ or its standard error.

The use of $\mathbf{V}^*$ in equation (7) is equivalent to a transformation of the original variables to a new set of variables with homogeneous variance. In particular, equation (2) can be redefined as

$$\mathbf{z} = \mathbf{W}\mathbf{y}^* = \mathbf{W}\mathbf{X}^*\mathbf{u} \tag{10}$$

where the transformation matrix is $\mathbf{W} = \mathbf{V}^{*-1/2}$. Following standard procedures (Graybill, 1961; Neter and Wasserman, 1974), the variance of $\mathbf{m}$ can be estimated from normal distribution theory, provided that the data fulfill certain requirements: (a) the elements of the vector $\mathbf{z}$ must be independent, (b) they must have homogeneous variance over their range, and (c) they must maintain a uniform normal distribution over their range. Because allele frequency data are multinomially distributed, these requirements cannot be realized exactly. Nevertheless, the treatment suggested above will result in these conditions being approximately met in most cases. As the number of alleles becomes large, the elements of $\mathbf{z}$ will be expected to approximate normal form, as a consequence of the Central Limit Theorem. Given that the data adequately meet these conditions, the variance of $\mathbf{m}$ is given by $V_m$ where

$$V_m = \text{MSE}\,(\mathbf{X}^{*\prime}\mathbf{V}^{*-1}\mathbf{X}^*)^{-1} \tag{11a}$$

and

$$\text{MSE} = \frac{(\mathbf{z} - \mathbf{W}\mathbf{X}^*\mathbf{m})'\,(\mathbf{z} - \mathbf{W}\mathbf{X}^*\mathbf{m})}{(p - L - 1)} = \frac{(\mathbf{y}^* - \mathbf{X}^*\mathbf{m})'\mathbf{V}^{*-1}\,(\mathbf{y}^* - \mathbf{X}^*\mathbf{m})}{(p - L - 1)} \tag{11b}$$

A confidence interval for $\mathbf{m}$ can be obtained using standard procedures (Neter and Wasserman, 1974:159).

### *Model evaluation*

The reliability of admixture analysis depends heavily on the degree to which the requirements of the model have been met; i.e., the degree to which good estimates of parental and hybrid population allele frequencies are available and drift and selection have been absent (Reed, 1969). Establishment of these conditions has traditionally amounted to a reconstruction of population histories and the assertion that effective population sizes have remained large enough that random processes have not resulted in significant departures from the original allele frequencies (eg, Workman et al., 1963; Thompson, 1973). It is difficult in practice to assess the truth of such statements, and the results of such analyses have consequently been controversial, especially where estimates of admixture proportions have been used to measure the action of natural selection (cf. Adams and Ward, 1973; Mandarino and

Cadien, 1974). The most common method for formally assessing the adequacy of an admixture model uses a chi-square test for heterogeneity among the $m_I$ values (Krieger et al., 1965; Cavalli–Sforza and Bodmer, 1971). While this approach is useful for detecting alleles that deviate from an expected mean value, it does not deal with our situation. The primary objective of this paper is to determine whether the role of admixture in the evolution of Borabuk is accurately reflected in the allele frequency data, taken as a whole.

For this purpose we suggest using the standard $R^2$ criterion from regression analysis. $R^2$ will be a measure of the efficacy of parental population allele frequencies in predicting allele frequencies in the hybrid population. Transformed allele frequencies in the hybrid population will be predicted by equation (10). Provided that the model requirements are met, there will be no variation about these predictions. If even minor violations of the model requirements occur, variability in predicted transformed allele frequencies will not be exactly zero. We suggest that the model should be evaluated in terms of the proportion of allele frequency variability in the mixed group that is accounted for by using the best estimates of hybrid population allele frequencies derived from the corresponding allele frequencies in the parental populations. This proportion is given by the $R^2$ criterion (Neter and Wasserman, 1974)

$$R^2 = \frac{[(\mathbf{y}^* - \bar{\mathbf{y}})' \mathbf{V}^{*-1} (\hat{\mathbf{y}} - \bar{\hat{\mathbf{y}}})]^2}{(\mathbf{y}^* - \bar{\mathbf{y}})' \mathbf{V}^{*-1} (\mathbf{y}^* - \bar{\mathbf{y}}) (\hat{\mathbf{y}} - \bar{\hat{\mathbf{y}}})' \mathbf{V}^{*-1} (\hat{\mathbf{y}} - \bar{\hat{\mathbf{y}}})} \quad (12)$$

$$= \frac{[(\mathbf{z} - \bar{\mathbf{z}})' (\hat{\mathbf{z}} - \bar{\hat{\mathbf{z}}})]^2}{(\mathbf{z} - \bar{\mathbf{z}})' (\mathbf{z} - \bar{\mathbf{z}}) (\hat{\mathbf{z}} - \bar{\hat{\mathbf{z}}})' (\hat{\mathbf{z}} - \bar{\hat{\mathbf{z}}})}$$

where $\mathbf{y} = \mathbf{X}^*\mathbf{m}$, $\bar{\mathbf{y}}$ is a vector containing of the average of the elements of $\mathbf{y}^*$ for all entries, $\bar{\hat{\mathbf{y}}}$ is a vector consisting of the average value of the elements of $\hat{\mathbf{y}}$ for all entries, and similar definitions apply to the various $\mathbf{z}$ measures. The $R^2$ criterion given in equation (12) is the square of the Pearson product moment correlation coefficient between the elements of $\mathbf{z}$ and $\hat{\mathbf{z}}$. It may be interpreted simultaneously as a measure of correspondence between the elements of $\mathbf{z}$ and $\hat{\mathbf{z}}$ and as the proportion of variability in hybrid population allele frequencies that is accounted for by the admixture model under investigation.

## GENE FLOW BETWEEN THE YE'CUANA AND YANOMAMA

The Ye'cuana occupy a contiguous territory to the north of the Yanomama, but they speak a language belonging to the Carib family that is not closely related to Yanomama dialects. The Ye'cuana have an extensive riverfaring tradition, a fact that resulted in their establishing an early trade network with European settlers in the ninteenth century, long before permanent European contact with the Yanomama was established in the mid-twentieth century. Through trade contact with the Ye'cuana, the Yanomama acquired steel and other valuable European goods even though they had not yet met Europeans. To facilitate these trade arrangements, the Yanomama village of Borabuk and the Ye'cuana village of Huduaduña settled at contiguous locations on the upper reaches of the Auris river about 1875. Ethnographic accounts reveal that this situation was stressful; because the Ye'cuana trade contact afforded them an advantage in social relations with the Yanomama, the Huduaduña men are purported to have been able to maintain persistent affairs with Yanomama women. Just prior to 1900 the situation became so stressful that a severe intervillage fight erupted, and the Yanomama village moved a considerable distance downstream (Chagnon et al., 1970). The Yanomama headman at the time of fission had been sired by a Ye'cuana father and raised by his Yanomama mother amongst her own people.

Further social disharmony between the people of Borabuk and Huduaduña resulted in the abduction of three Huduaduña women and their incorporation into Borabuk as wives. A unique series of social events, coupled with demographic and sociocultural processes affecting marriage and reproduction in the two tribes resulted in the captive Ye'cuana women, the half-Ye'cuana headman, and his half- and full-Ye'cuana brothers contributing an extraordinary number of descendants to the Yanomama village. A detailed account of this sequence has been given by Chagnon et al. (1970).

In examining Borabuk for admixture, the identification of the correct parental populations and estimation of their allele frequencies is highly problematic. The Yanomama parental population of Borabuk was the antecedent of the modern village and consequently no longer exists as such. Two possibilities exist for the estimation of the parental Yanomama frequencies. The first involves determination

of admixed and unadmixed individuals in the village from the pedigree data available. The unadmixed portion of the population represents a genetic sample of the original parental group and would consequently provide an estimate of that parental group, unbiased by the introduction of Ye'cuana alleles. We do not favor this approach, because only a small fraction of the present village is not admixed, and any estimates derived from this portion are not likely to be reliable. The second approach is developed from the subtribal genetic structure of the Yanomama; Yanomama villages can be grouped into subtribal clusters on the basis of historical and linguistic relationships (cf. Migliazza, 1972; Ward, 1972; Smouse, 1982). These clusters represent groups of villages descended essentially from a single ancestral population and are the result of numerical and territorial expansion of the Yanomama over at least the past 100 years. Villages belonging to a single cluster show a high degree of cultural and genetic similarity (cf. Smouse, 1982). Therefore, we have chosen to utilize cluster averages as the best estimates of the ancestral parental population frequencies. Since it is not known with any confidence to which cluster of villages Borabuk belonged prior to its difficulties with Huduaduña, we have used the frequencies from several different Yanomama village clusters as different estimates of the ancestral Borabuk frequencies.

Ascertainment of the correct Ye'cuana parental population poses a different set of problems. While all the Ye'cuana individuals making genetic contributions to Borabuk were from a single village, Huduaduña, many factors suggest that the present day frequencies are not accurate estimates of those at the time of genetic exchange with Borabuk. This village is uncharacteristic of more stable villages, because it has a surprisingly high rate of exogamy and has absorbed many groups of Ye'cuana who have been displaced by European contact (Ward and Neel, 1970). In addition to these factors, the total number of individuals composing the group has remained low enough that genetic drift may have played an important role in determining the allele frequencies observed in this group. Unfortunately, Huduaduña does not show close affiliation with any of the established Ye'cuana village clusters (Smouse and Ward, 1978), so cluster averages cannot credibly be used as estimates. Lacking a reasonable alternative, we will assume that the frequencies observed in modern Huduaduña are representative of those at the time of admixture. Any variation this introduces into the model should be evident in the evaluation criterion, equation (12), resulting in lowered $R^2$.

## GENETIC MARKER ANALYSIS

Allele frequency data for eleven loci found to be polymorphic in the Yanomama and Ye'cuana are presented in Table 1. These data were gathered from several sources (Gershowitz et al., 1967, 1970, 1972; Shreffler and Steinberg, 1967; Tashian et al., 1967; Arends et al., 1970; Weitkamp and Neel, 1970, 1972; Tanis et al., 1973; Ward et al., 1975), and have been partially summarized by Smouse and Neel (1977) and Smouse and Ward (1978). They form the basis for five different admixture models. Model I uses the average frequencies observed in six villages belonging to the Yanam village cluster to represent the ancestral Borabuk allele frequencies. Models II–IV are developed from the averages of a single village belonging to the Ninam cluster, five villages belonging to the Sanema cluster, and four villages belonging to the Shamatari cluster. (Borabuk is currently classified as a member of the Ninam cluster on the basis of geographic contiguity (cf. Smouse and Neel, 1977) but this assignment is not very firm.) Model V uses allele frequencies observed in the Saõ Marcos Xavante (Gershowitz et al., 1967; Shreffler and Steinberg, 1967; Tashian et al., 1967) to represent the ancestral Borabuk allele frequencies. The Xavante are a Gê-speaking Amerindian tribe from the Mato Grosso of Brazil. They are not geographically close to the Yanomama or Ye'cuana and show little cultural or genetic similarity with either tribe. They have been included in this analysis to illustrate the results expected when an inappropriate ancestral group is used. For each model, the frequencies observed in modern Huduaduña are taken as representative of the ancestral Ye'cuana frequencies. The appropriate data for Huduaduña and Borabuk are also provided in Table 1, along with sample sizes for each group.

Weighted maximum likelihood estimates of the proportion of Yanomama ancestry in Borabuk and model evaluation statistics ($R^2$) are provided in Table 2. All estimates of $\mathbf{u}$ were obtained according to equation (7). Inspection of the $R^2$ values in Table 2 reveals that Model IV, with the Shamatari cluster as the Yanomama parental population, accounts for the greatest portion of Borabuk allele frequency variation. Because this model accounts for 80% of the variance of present day Borabuk allele frequencies, we may conclude that an evolutionary model based on admixture provides a

*TABLE 1. Allele and haplotype frequencies for the seven populations used in models I through V*[1]

| Genetic locus | Allele/ haplotype | Group designations and sample (N) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Hu (70) | Bk (74) | Yn (351) | Nm (24) | Sn (196) | Sh (302) | Xv (113) |
| P | :$P^1$ | 0.51 | 0.47 | 0.76 | 0.46 | 0.53 | 0.47 | 0.63 |
| | :$P^2$ | 0.49 | 0.53 | 0.24 | 0.54 | 0.47 | 0.53 | 0.37 |
| Duffy | :$Fy^a$ | 0.64 | 0.40 | 0.55 | 0.80 | 0.50 | 0.68 | 0.53 |
| | :$Fy^b$ | 0.36 | 0.60 | 0.45 | 0.20 | 0.50 | 0.32 | 0.47 |
| Kidd | :$Jk^a$ | 0.36 | 0.52 | 0.53 | 0.54 | 0.76 | 0.49 | 0.40 |
| | :$Jk^b$ | 0.64 | 0.48 | 0.47 | 0.46 | 0.24 | 0.51 | 0.60 |
| Lewis | :$Le^+$ | 0.77 | 0.43 | 0.50 | 0.46 | 0.49 | 0.40 | 0.67 |
| | :$Le^-$ | 0.23 | 0.57 | 0.50 | 0.54 | 0.51 | 0.60 | 0.33 |
| Diego | :$Di^a$ | 0.04 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 |
| | :$Di^b$ | 0.96 | 0.92 | 1.00 | 1.00 | 1.00 | 1.00 | 0.85 |
| MNS | :MS | 0.14 | 0.09 | 0.31 | 0.35 | 0.10 | 0.04 | 0.37 |
| | :Ms | 0.53 | 0.69 | 0.37 | 0.65 | 0.43 | 0.76 | 0.36 |
| | :NS | 0.27 | 0.00 | 0.00 | 0.00 | 0.03 | 0.10 | 0.15 |
| | :Ns | 0.06 | 0.22 | 0.32 | 0.00 | 0.43 | 0.10 | 0.12 |
| Rhesus | :$R^z$ | 0.00 | 0.10 | 0.10 | 0.13 | 0.01 | 0.20 | 0.03 |
| | :$R^1$ | 0.42 | 0.58 | 0.76 | 0.56 | 0.95 | 0.76 | 0.57 |
| | :$R^2$ | 0.53 | 0.20 | 0.20 | 0.31 | 0.04 | 0.04 | 0.36 |
| | :$R^0$ | 0.05 | 0.12 | 0.02 | 0.00 | 0.00 | 0.00 | 0.04 |
| Haptoglobin | :$Hp^1$ | 0.27 | 0.79 | 0.80 | 0.74 | 0.89 | 0.78 | 0.48 |
| | :$Hp^2$ | 0.73 | 0.21 | 0.20 | 0.26 | 0.11 | 0.22 | 0.52 |
| Group specific component | :$Gc^1$ | 0.81 | 0.81 | 0.82 | 0.44 | 0.79 | 0.89 | 0.30 |
| | :$Gc^2$ | 0.19 | 0.19 | 0.18 | 0.56 | 0.21 | 0.11 | 0.70 |
| Phospho-glucomutase 1 | :$PGM^1$ | 0.96 | 0.94 | 0.95 | 0.96 | 0.92 | 0.92 | – |
| | :$PGM^2$ | 0.04 | 0.06 | 0.05 | 0.04 | 0.08 | 0.08 | – |
| Red cell acid phosphate | :$ACP^a$ | 0.05 | 0.10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 |
| | :$ACP^b$ | 0.95 | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 | 0.80 |

[1]Hu, Huduaduña; Bk, Borabuk; Ym, Yanam; Nm, Ninam; Sn, Sanema; Sh, Shamatari; and Xv, Saõ Marcos Xavante.

*TABLE 2. Estimated proportion of Yanomama ancestry, standard error of estimate, model evaluation criterion, and approximate significance level for each of the admixture models*

| Model | Population sampled | Estimated proportion of Yanomama ancestry | Standard error | $R^2$ | Approximate significance level |
|---|---|---|---|---|---|
| I | Yanam | 0.6089 | 0.1190 | 0.6397 | 0.0020 |
| II | Ninam | 0.4408 | 0.2179 | 0.2080 | 0.0874 |
| III | Sanema | 0.7026 | 0.1751 | 0.5534 | 0.0013 |
| IV | Shamatari | 0.8157 | 0.1249 | 0.7989 | 0.0000 |
| V | Xavante | 0.2688 | 0.2227 | 0.0747 | 0.2488 |

reasonable explanation for the genetic structure of Borabuk.

Features of the normal error model relevant to the estimation of the standard error of **m** are evaluated graphically in Figure 1. The individual error terms $e_i = (z_i - \hat{z}_i)$, are plotted against the predicted values, $\hat{z}_i$, for the Model IV (Shamatari) in Figure 1a, and show no consistent trend, as required by the estimation procedure. Figure 1b, which is a cumulative frequency plot of the standardized error terms for Model IV on normal probability paper, shows that the error terms closely approximate a normal distribution, also required by the procedure. Kolmogorov–Smirnov tests for goodness of fit to the normal distribution (Sokal and Rohlf, 1969) failed to establish significant departures for any of the five models. Given that clear violations of the requirements of the normal error model are not observed in these data and that the normal error model is known to be robust to even moderate departures from its assumptions (Neter and Wasserman, 1974), we suggest that estimation of standard error terms by this procedure is justified. These estimates are also provided in Table 2, along with the admixture estimates.

## PEDIGREE ANALYSIS OF ADMIXTURE PROPORTIONS

Detailed ethnographic accounts of Borabuk history collected by Dr. Napoleon Chagnon have

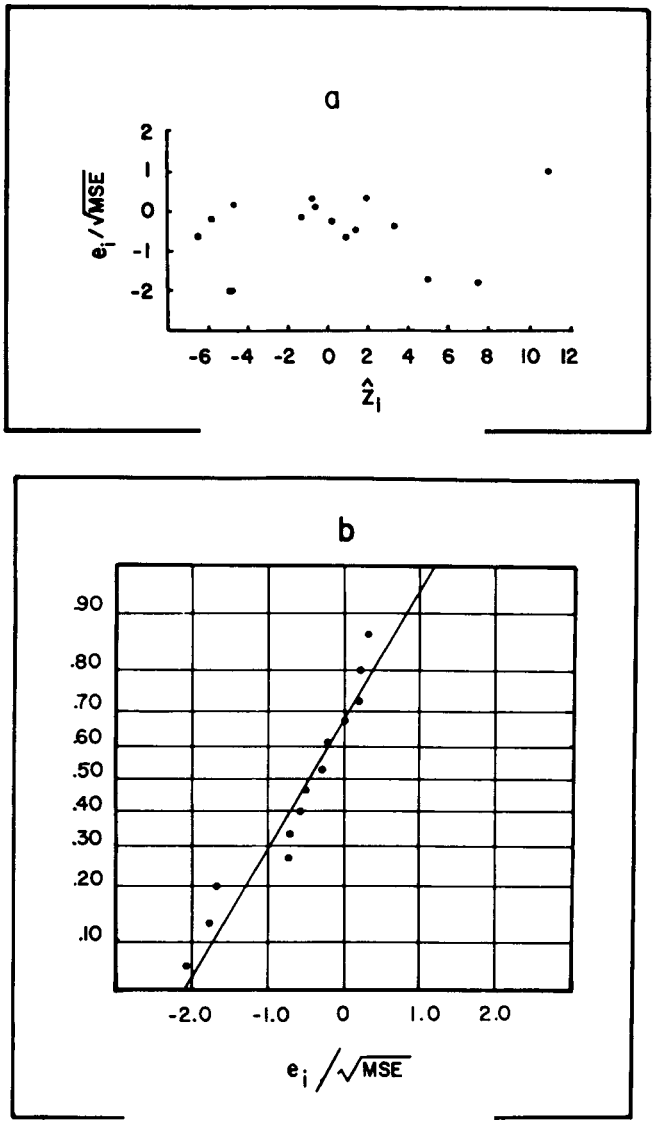


Fig. 1. (a) A plot of standardized allele frequency error terms $e_i/\sqrt{MSE}$ against their predicted values, $\hat{z}_i$, where $\mathbf{e}_i = \mathbf{z}_i - \hat{\mathbf{z}}_i = \mathbf{W}\mathbf{y}_i^* - \mathbf{W}\mathbf{X}_i^*\mathbf{m}$. (b) Cumulative frequency plot of standardized error terms on normal probability coordinates. Neither graph provides any evidence for violation of normal error model assumptions.

allowed the construction of a partial village pedigree spanning six generations (Chagnon et al., 1970). This cultural information is complete enough to allow estimation of ancestral proportions in Borabuk. These estimates are independent of those derived from the genetic marker data and provide a convenient reference frame against which to evaluate the genetic analysis.

The Borabuk pedigree is presented in Figure 2; it includes the known immigrants into Borabuk and the present village members who are their descendants. (We have slightly modified the original pedigree in Chagnon et al. (1970) to improve visualization. Occasional individuals who were identical with respect to their Ye'cuana ancestry and whose descendants were also identical have been coalesced. All pedigree

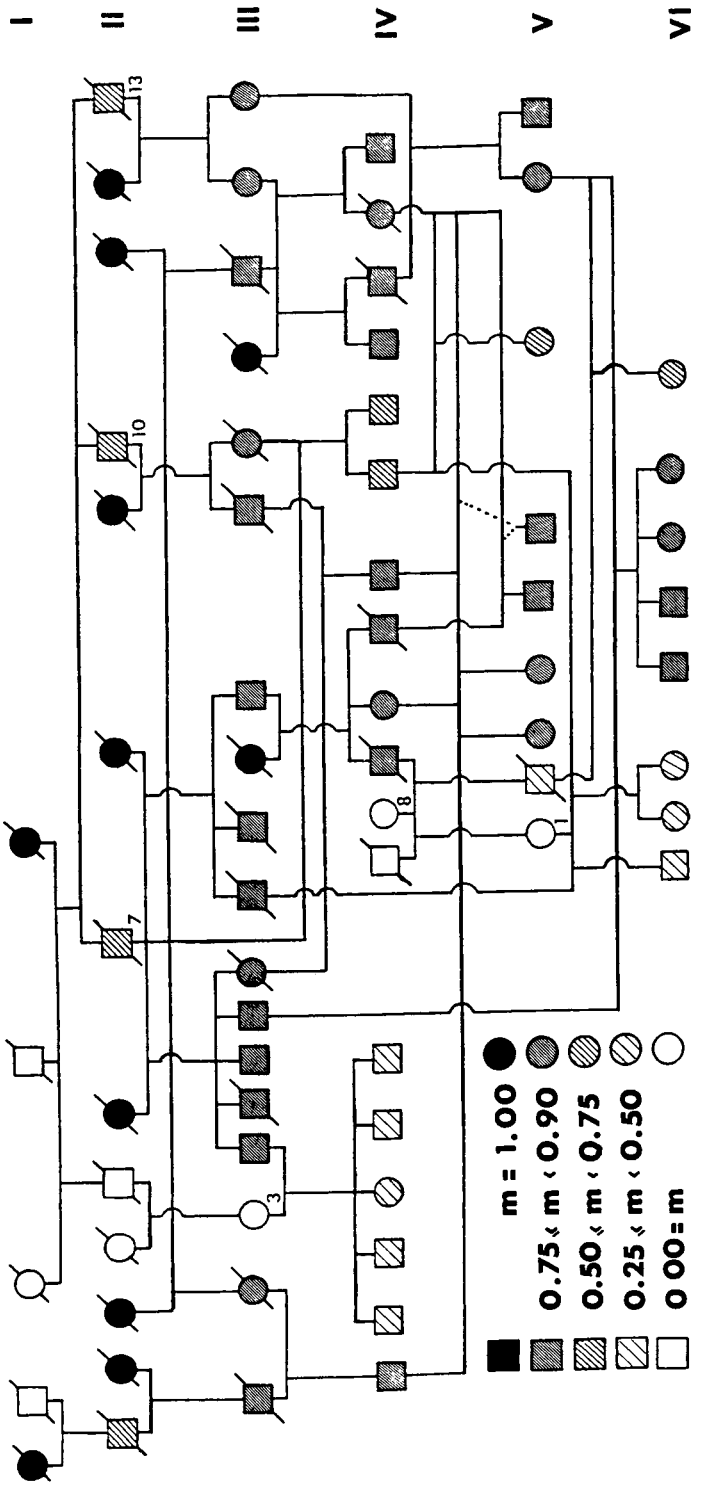


Fig. 2. Borabuk partial pedigree spanning six generations. The percentage of Yanomama ancestry is provided for each individual. A slashed line indicates the individual is either deceased or unsampled for some other reason.

ambiguities except one have been visually "resolved" in such a way that the net Ye'cuana ancestry of the village was not altered. The remaining ambiguity cannot be resolved, and its coefficient is simply the average of that with each of the possible fathers.) This visual simplification has not been used for the formal analysis described below, for which each individual has been separately treated. Examination of Figure 2 attests to the considerable reproductive success of the half-Ye'cuana headman (II7), his brothers (II10, II13), and two of the three abducted Ye'cuana women (III3, V1). Estimates of admixture proportions can be derived as the expected proportion of alleles in the population that descended from full-blooded Yanomama and Ye'cuana ancestors.

For this purpose, the proportion of Yanomama ancestry possessed by each individual, $m_i$, is estimated as the average of the proportions of Yanomama ancestry possessed by that individual's parents

$$m_i = 1/2(m_{P1} + m_{P2}) \qquad (13)$$

where $m_{P1}$ is the proportion of Yanomama ancestry possessed by individual i's first parent and $m_{P2}$ corresponds to the second parent. To obtain an estimate of the Yanomama contribution to the Borabuk gene pool, the $m_i$ values for each living individual who was included in the genetic marker sample is summed, and this total is then divided by the numbers of individuals in the genetic sample.

Following the above procedure, the *sampled* residents of Borabuk are estimated to have 20% Ye'cuana and 80% Yanomama ancestry. This method assumes that members of the Borabuk village not shown in the pedigree are not admixed, and that all Ye'cuana alleles possessed by individuals shown in the pedigree have descended from known immigrants. These assumptions are supported by the fact that all persons in Borabuk carrying the Ye'cuana marker traits, $Di^a$ and $ACP^a$, are accounted for by the pedigree, but ignores the possibility of small amounts of undetected admixture elsewhere in the village. The pedigree-derived estimates of ancestral proportions are very close to those estimated from the genetic marker data in Model IV, (0.82 ± 0.12). It is of interest that Model IV also has the highest $R^2$ value (0.80).

## DISCUSSION

A comparison of single allele and multiallelic approaches to genetic admixture analysis is presented in Figure 3, and is very instructive. Single allele estimates of Yanomama ancestry were obtained from equation (3) using the frequencies of 26 alleles belonging to the 11 loci employed for the weighted maximum likelihood (WML) analysis given in equation (7). For Figure 3, the ancestral Yanomama allele frequencies were estimated as those observed in the Shamatari village cluster (Model IV), and the ancestral Ye'cuana allele frequencies were estimated as those observed in

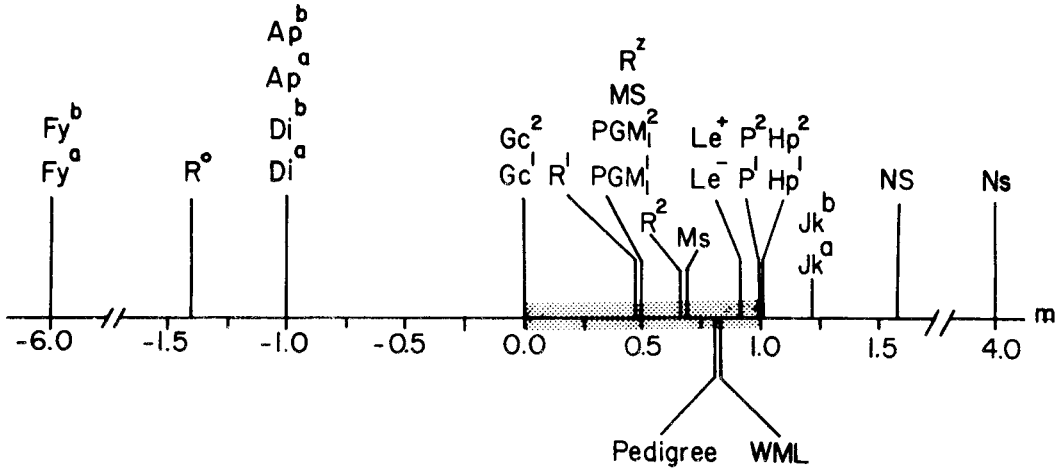


Fig. 3. Graph of estimated proportion (m) of Yanomama ancestry by single allele method (above the line), weighted maximum likelihood method (WML) and pedigree method (below the line). Only the sampled individuals in Figure 2 were used for these analyses.

modern Huduaduña. Also shown on Figure 3 are the multilocus estimates of Yanomama ancestry derived from Model IV and from the pedigree analysis. The shaded region between 0.0 and 1.0 represents the range of admixture proportions that are biologically possible. The values obtained from single allele analyses range from negative 600% Yanomama ancestry to positive 400% Yanomama ancestry. The $Di^a$ and $ACP^a$ frequencies discussed at the outset are not even particularly extreme. The single allele estimates are highly heterogeneous, and many are observed to lie outside the biologically expected range. This is because various genetic marker loci differ greatly in their utility in estimating admixture proportions. The model specified in equation (6) gives the loci with the most information (ie, smallest variances) the most importance for the combined estimate. The information available for obtaining an admixture estimate at a locus depends on several factors: (a) the allele frequency variances in the parental and hybrid populations, (b) the genetic divergence between parental population allele frequencies, (c) and $m_l$ itself. The alleles yielding the most unrealistic estimates of $m_l$ also tend to have the largest variances. Of the factors contributing to the information at a locus, separation of parental population allele frequencies seems to be the most important. Parental population allele frequencies are separated by only a few percent for most of the alleles, yielding $m_l$ values outside the biologically permissible range; these $m_l$ estimates also have extraordinarily high variances and contribute little to the estimation **m**. Consequently, a suite of characters analyzed individually can yield widely varying results, while the same suite of characters can yield entirely reasonable results when analysed simultaneously and allowing for differences in information.

The estimates of **u** resulting from all five multiallelic models (Table 2) are in accordance with the requirements of the mathematical formulation of the genetic admixture process ($\Sigma m_j = 1.0$) stipulated in equation (2). This condition is ensured by the estimation procedure equation (7). This characteristic of the estimation procedure, however, does not ensure that the model itself is appropriate. Model V, which erroneously used allele frequencies observed in the Saõ Maros Xavante as estimates of the allele frequencies in the Yanomama parental population, gives a value of 0.22 for the proportion of Yanomama ancestry. This is not a very plausible value, given what we know, but probably could not have been excluded on the basis of ethnohistorical consideration alone, without the aid of the pedigree. The possession of cultural information complete enough to allow construction of a village pedigree containing all known immigrants is a luxury of this particular situation that cannot be generally expected. However, the low $R^2$ value, 0.07, associated with this model is sufficient to reject Model V as an adequate explanation for allele frequency variation in the Borabuk population. The $R^2$ evaluation procedure will be available with any genetic marker data set and should prove especially useful in the absence of detailed historical information.

The $R^2$ values for the four models using Yanomama allele frequencies all greatly exceed the value of $R^2$ for Model V. The Yanomama models all explain a considerably larger portion of allele frequency variation in Borabuk than does Model V. This can be attributed to the fact that allele frequencies in the different subtribal village clusters are highly correlated as a result of having originated from a single ancestral population. While any one of these cluster averages may or may not reflect the actual ancestral Yanomama population from which Borabuk was drawn, it is likely that the allele frequencies in the true ancestral population would have been similar. The relatively poor fit of Model II can probably be assigned to the fact that the Ninam allele frequencies were estimated from a small sample ($N = 24$) in a single village. Because of this, the Ninam are inadequate representations of either the modern population or the ancestral group.

At the outset of this investigation, our objective was to determine whether a multiallelic approach to the analysis of Borabuk allele frequencies would serve to balance the seemingly bizarre results obtained from infiltration of two tribally restricted alleles, $Di^a$ and $ACP^a$. Application of this approach demonstrated that a model explaining a large (80%) portion of allele frequency variation in Borabuk could be developed, despite the fact that several of the individual marker alleles included in the analysis would yield widely aberrant results if considered in isolation. The values of the $u_i$ parameters estimated in the multiallelic analyses were then seen to be in close agreement with those suggested by a partial village pedigree.

The superiority of a multiallelic approach to the analysis of hybrid population allele frequencies is clearly shown by these findings.

ACKNOWLEDGMENTS

We wish to thank Dr. Harvey Mohrenweiser, Dr. James V. Neel, and Dr. James W. Wood, Department of Human Genetics, University of Michigan, for their comments on earlier versions of this paper. A pair of anonymous reviewers also contributed useful critique.

This study was supported by grants from the National Institutes of Health, 2-T32-GM-07123 (to J.C.L.), and the National Science Foundation, NSF-DEB-7823293 (to P.E.S.).

LITERATURE CITED

Adams, J, and Ward, RH (1973) Admixture studies and the detection of selection. Science *180*:1137–1143.

Arends, T, Weitkamp, LR, Gallango, ML, Neel, JV, and Schultz, J (1970) Gene frequencies and microdifferentiation among the Makiritare Indians. II. Seven serum protein polymorphisms. Am. J. Hum. Genet. *22*:526–532.

Arvello-Jimenez, N (1971) Political Relations in a Tribal Society: a study of the Ye'cuana Indians of Venezuela. Unpubl. Ph.D. dissertation, Dept. of Anthropology, Cornell University, University of Microfilms 71-18:901.

Cavalli–Sforza, LL, and Bodmer, W (1971) The Genetics of Human Populations. Freeman: San Francisco.

Chagnon, NA (1968) Yąnomamö: The Fierce People. New York: Holt, Rinehart and Winston.

Chagnon, NA, Neel, JV, Weitkamp, LR, Gershowitz, H, and Ayres, M (1970) The influence of cultural factors on the demography and pattern of gene flow from the Makiritare to the Yanomama Indians. Am. J. Phys. Anthropol. *32*:339–349.

Elston, RC (1971) The estimation of admixture in racial hybrids. Ann. Hum. Genet. *35*:9–17.

Gershowitz, H, Junqueira, PC, Salzano, FM, and Neel, JV (1967) Further studies on the Xavante Indians. III. Blood groups and ABH–Le$^a$ secretor types in the Simões Lopes and Saõ Marcos Xavantes. Am. J. Hum. Genet. *19*:502–513.

Gershowitz, H, Layrisse, M, Neel, JV, Brewer, C, Chagnon, N, and Ayres, M (1970) Gene frequencies and microdifferentiation among the Makiritare Indians. I. Eleven blood group systems and the ABH–Le$^a$ secretor traits: A note on Rh gene frequency determinations. Am. J. Hum. Genet. *22*:515–525.

Gershowitz, H, Layrisse, M, Layrisse, Z, Neel, JV, Chagnon, N, and Ayres, M (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.

Graybill, FA (1961) An Introduction to Linear Statistical Models. Vol. I. New York: McGraw-Hill.

Krieger, H, Morton, NE, Mi, MP, Azevedo, E, Freire-Maia, A, and Yasuda, N (1965) Racial Admixture in northeastern Brazil. Ann. Hum. Genet. *29*:113–125.

Li, CC (1976) First Course in Population Genetics. Pacific Grove: Boxwood.

Mandarino, L, and Cadien, JD (1974) Use of ranked migration estimates for detecting natural selection. Am. J. Hum. Genet. *26*:108–112.

Migliazza, EC (1972) Yanomama Grammar and Intelligibility. Unpubl. Ph.D. Dissertation, Department of Linguistics, Indiana University. University Microfilms 72-30:432.

Neel, JV (1978) The population structure of an Amerindian tribe, the Yanomama. Ann. Rev. Genet. *12*:365–413.

Neter, J, and Wasserman, W (1974) Applied Linear Statistical Models. Homewood, IL: Richard D. Irwin, Inc.

Reed, TE (1969) Caucasian genes in American Negroes. Science *165*:762–768.

Roberts, DF, and Hiorns, RW (1962) The dynamics of racial admixture. Am. J. Hum. Genet. *14*:261–277.

Roberts, DF, and Hiorns, RW (1965) Methods of analysis of the genetic constitution of a hybrid population. Hum. Biol. *37*:38–43.

Shreffler, DC, and Steinberg, AG (1967) Further studies on the Xavante Indians. IV. Serum protein groups and the $SC_1$ trait of saliva in the Simões Lopes and Saõ Marcos Xavantes. Am. J. Hum. Genet. *19*:514–523.

Smouse, PE (1982) Genetic architecture of swidden agricultural tribes from the lowland rainforests of South America. In M Crawford and J Mielke (eds): Current Developments in Anthropological Genetics: Ecology and Population Structure, vol. II. New York: Plenum Press. pp. 139–178.

Smouse, PE, and Neel, JV (1977) Multivariate analysis of gametic disequilibrium in the Yanomama. Genetics *85*:733–752.

Smouse, PE, and Ward, RH (1978) A comparison of the genetic infrastructure of the Ye'cuana and Yanomama: A likelihood analysis of genotypic variation among populations. Genetics *88*:611–631.

Sokal, R, and Rohlf, FJ (1969) Biometry. San Francisco: Freeman and Co.

Spielman, RS, Migliazza, EC, and Neel, JV (1974) Regional linguistic and genetic differences among Yanomama Indians. Science *184*:637–644.

Szathmary, EJE, and Reed, TE (1972) Caucasian admixture in two Ojibwa Indian communities in Ontario. Hum. Biol. *44*:655–671.

Tanis, RJ, Neel, JV, Dovey, H, and Morrow, M (1973) The genetic structure of a tribal population, the Yanomama Indians. IX. Gene frequencies for 17 serum protein and erythrocyte enzyme systems in the Yanomama and five neighboring tribes: nine new variants. Am. J. Hum. Genet. *25*:655–676.

Tashian, RE, Brewer, GJ, Lehmann, H, Davies, DA, and Rucknagel, DL (1967) Further studies on the Xavante Indians. V. Genetic variability in some serum and erythrocyte enzymes, hemoglobin, and the urinary excretion of β-aminoisobutyric acid. Am. J. Hum. Genet. *19*:524–531.

Thompson, EA (1973) The icelandic admixture problem. Ann. Hum. Genet. *37*:69–80.

Ward, RH (1972) The genetic structure of a tribal population the Yanomama Indians. V. Comparisons of a series of genetic networks. Ann. Hum. Genet. *36*:21–42.

Ward, RH, and Neel, JV (1970) Gene frequencies and microdifferentiation among the Makiritare Indians. IV. A comparison of a genetic network with ethnohistory and migration matricies; a new index of genetic isolation. Am. J. Hum. Genet. *22*:538–561.

Ward, RH, Gershowitz, H, Layrisse, M, and Neel, JV (1975) The genetic structure of a tribal population, the Yanomama Indians. XI. Gene frequencies for 10 blood groups and the ABH–Le secretor traits in the Yanomama and their neighbors: The uniqueness of the tribe. Am. J. Hum. Genet. *27*:1–30.

Weitkamp, LR, and Neel, JV (1970) Gene frequencies and microdifferentiation among the Makiritare Indians. III. Nine erythrocyte enzyme systems. Am. J. Hum. Genet. *22*:533–537.

Weitkamp, LR, and Neel, JV (1972) The genetic structure of a tribal population, the Yanomama Indians. IV. Eleven erythrocyte enzymes and summary of protein variants. Ann. Hum. Genet. *35*:433–443.

Workman, PL, Blumberg, BS, and Cooper, AJ (1963) Selection, gene migration and polymorphic stability in a US White and Negro population. Am. J. Hum. Genet. *15*:429–437.