 Manufactured in the United States of America. Library of Congress Catalog Card No. 72–94660. International Standard Book No. 0–8263–0274–2.
First Edition

9

Genetic Distance Among Southern African Populations

HENRY HARPENDING
Department of Anthropology
Yale University[1]

TREFOR JENKINS
South African Institute for Medical Research
Johannesburg

Analysis of variation in gene frequencies within an area may have several purposes, and much of the ongoing disagreement about methodology in human population genetics may reflect as much a lack of clear perception of the purposes of the methodology as it does real conflict over substantive theoretical issues. A method of analysis should be directed toward answering questions, and these questions should be explicit.

There are two broad and overlapping areas of inquiry for which studies of regional gene frequency variation are pertinent. One is population structure—the study of the effects of internal migration, group composition, mating practices, and other factors on the amount and pattern of genetic drift within an area. The second is really population history. Here,

the questions concern the degree of similarity among populations, where similarity may reflect either common ancestry or mate exchange. These two aspects of genetic similarity are ordinarily inseparable.

Studies of population structure typically focus on small and/or homogeneous areas. Under the assumption that the study population is near "equilibrium," that is, that a stationary distribution of gene frequencies has been reached, variation in gene frequencies or genotype frequencies is compared with predictions made by considering various demographic parameters of the population. Each marker locus yields a measure of the amount of drift. This measure is either Wahlund's coefficient of inbreeding (F_{ST} in Workman and Niswander [1970], F_W in Cavalli-Sforza [1969b], F in Harpending and Jenkins [1972b]), or the genotypic disequilibrium inbreeding coefficient of Li and Horvitz (1953) (F_{IT} in Workman and Niswander [1970], α in Yasuda [1968b] and Morton et al. [1971c]). In addition, the form of the regression or normalized genetic covariance—r in Harpending and Jenkins (1972b), $y(d)$ in Morton et al. (1971c)—is of interest, although it now seems that this regression is primarily a function of sample size, distribution, and the local inbreeding coefficient, rather than a function of the underlying population structure. This biological measure of drift is compared with predictions from any or all of the following demographic items: pedigrees, isonymy, the root mean square distance between birthplaces of parents or between parents and offspring (Malécot 1948, Azevêdo et al. 1969), a matrix of frequencies of gene exchange among areas or villages (Bodmer and Cavalli-Sforza 1968, Smith 1969, Morton 1969, Friedlaender 1971a, Harpending and Jenkins 1972b), and local effective population size.

Evaluation of the amount of concordance between predictions from demography and observed marker gene frequency variation has been equivocal, because of uncertainty about what constitutes reasonable agreement, and because of uncertainty in the interpretation of the predictive models. For example, Friedlaender (1971a) and Morton et al. (1971c) assume that inbreeding statistics given by the migration matrix model should correspond to Wahlund F statistics, while Harpending and Jenkins (1972b) modify their results to account for finite sample size. The difference between the corrected and uncorrected predictions is very large. As this approach is refined and applied to diverse groups it may lead to inferences about selection at loci which deviate significantly from predictions.

The second, or historical, approach to studying gene frequency variation has been applied to populations of all sizes from single tribes (Ward and Neel 1970a) to the whole world (Cavalli-Sforza and Edwards 1967). In these studies, group gene frequencies are converted into genetic "distances" among groups, and these distances are used to make a "genetic map" (Sanghvi, Kirk, and Balakrishnan 1971) or a cladogram, both of which provide a visible if anecdotal summary of the intergroup differences. This summary diagram is then compared intuitively with knowledge of mating patterns, ancestry, linguistic relationships, or other heuristic indicators of similarities among the groups.

There are many measures of genetic distance, all of which are reasonable. They are all related to one another, and none is likely to seriously mislead an investigator. Consideration of the notion of genetic similarity should, however, lead to some measure that has specifiable advantages. These may be that it lend itself equally well to the construction of cladograms and to the construction of genetic maps and that it be clearly related to genetic theory—that is, that it have more than anecdotal interpretation.

The genetic distance between two groups should be small if their gene frequencies are similar. There are a number of reasons why two groups should have similar gene frequencies, the more conspicuous of which are: (1) they shared a recent common ancestor; (2) they exchange genes; (3) they are large, so that little drift has occurred since their separation; and (4) their loci are or have been subject to similar selection pressure. Of these, the third reason is often overlooked. Since drift of mean gene frequencies of a subdivided group is nearly independent of mating patterns within the group (Ewens 1969), consideration of size makes reasonable the finding, for example, that genetic distances between villages within American Indian linguistic groups are as large as distances between linguistic groups.

The fourth cause of similar gene frequencies (similar selection histories of populations) is, in practice, not often cited to explain the results of studies. Many studies of genetic distance are of restricted homogeneous areas where there is no reason to suspect heterogeneity in selection pressures. Other selection environments, such as homogeneous selection over an area or random changes in the magnitude and intensity of directional selection, will not, in effect, be very different from drift. For these reasons and for others discussed in Cavalli-Sforza (1969), selection is usually not

explicitly considered in studies of genetic distance, in which genetic drift is the presumed agent responsible for observed gene frequency differentiation. Within small areas, many kinds of selection would have little effect on the differentiation caused by drift, and even regional heterogeneity would be swamped by drift and migration unless it was very strong.

Genetic drift is described by kinship statistics, and these should be the basis of measures of genetic distance. The coefficient of kinship between two groups that are labeled i and j is written f_{ij} (or φ_{ij} in Morton et al. [1971c] and Harpending and Jenkins [1972b]). This coefficient has two interpretations that are often used interchangeably: (1) the probability that a random allele at a specified locus in population i is identical by descent to a random allele from the same locus in population j is f_{ij}; (2) the populations are undergoing genetic drift. There is some gene frequency P that is either the initial gene frequency before drift or else the gene frequency toward which systematic pressure is directing the frequencies in groups i and j. Then, f_{ij} is the normalized covariance between gene frequencies of any allele described by the model, that is,

$$f_{ij} = E\left[\frac{(p_i - P)(p_j - P)}{P(1-P)}\right],$$

where p_i and p_j are the gene frequencies in groups i and j, and $E(\cdot)$ means expectation or average value of the term on which it operates.

The first definition of f_{ij} is applicable to individuals as well as to groups, and i may be the same or different from j. When i is the same as j, the second definition is a form of Wahlund's principle (Li 1955), and it also is applicable, with slight modification, to individuals as well as to groups (Harpending and Jenkins 1972b).

The second definition of the coefficient of kinship immediately suggests that a reasonable measure of genetic distance is

$$\Delta_{ij} = f_{ii} + f_{jj} - 2f_{ij}.$$

This is simply the squared Euclidean distance between populations i and j in a hyperspace whose axes are allelic frequencies scaled by dividing by the normalizing factor $\sqrt{[P(1-P)]}$ as is appropriate for genetic drift.

COMPUTATION

The observed or sample coefficient of kinship between groups i and j is given as

$$r_{ij} = \frac{(P_i - \bar{P})(P_j - \bar{P})}{\bar{P}(1-\bar{P})}$$

for any allele. (Morton et al. [1971c] use y instead of r for a similar coefficient.) In this expression, $\bar{P}$ is the weighted mean gene frequency of the allele in the study array; it provides an estimate of the "underlying" mean gene frequency P. The matrix of sample kinship coefficients is calculated for each allele, and these matrices are averaged to yield one overall matrix of sample coefficients. If all the alleles studied are subject to genetic drift with the same systematic pressure (imagined to be immigration from the outside world), then all alleles should give estimates of the same "true" r coefficients, subject only to chance deviations. Hence, Harpending and Jenkins (1972b) simply averaged the r matrices from all alleles they studied. It is probably better to give unequal weight to alleles by weighting estimates from different loci by the degrees of freedom at the locus, as when genetic distance is calculated as a chi-squared statistic. Morton et al. (1971c) propose a weighting method that depends on their own special programs but that may be a better way to reconcile the following kinds of consideration: (1) loci without dominant alleles are much more informative and should be given greater weight than loci like ABO, where much of the gene frequency variation may reflect estimation error; and (2) common alleles are more informative than rare alleles whose frequencies are much more subject to sampling error. For samples of the size common in anthropological studies, the difference between the gene frequencies of 0.3 and 0.4 is meaningful, while the difference between 0.03 and 0.04 is not. Whatever the method used to combine information from the various loci, the resulting matrix of sample relationship coefficient r_{ij} is then amenable to analysis in several ways; one may make a "tree" or one may examine its principle axes and make a "map."

TREES

Harpending and Jenkins (1972b) suggest that the expected or average value of a sample relationship coefficient is

$$E(r_{ij}) \cong \frac{f_{ij} + \bar{f} - \bar{f}_i - \bar{f}_j}{1 - \bar{f}}.$$

Here, $\bar{f}$ is random kinship within the sample. Writing w_i as the proportion of the total sample that is the ith population ($w_i = N_i / \sum_k N_k$, where N refers to census and not sample sizes), random kinship is

$$\bar{f} = \sum_{i,j} w_i w_j f_{ij}.$$

This is interpretable as the probability of identity by descent of two random alleles from the same locus or as a measure of the drift of sample mean gene frequencies $\bar{p}$ away from the prior mean P, that is,

$$\bar{f} = E\left[\frac{(\bar{p} - P)^2}{P(1 - P)}\right].$$

Similarly $\bar{f}_i$ is the random kinship of population i, that is,

$$\bar{f}_i = \sum_j w_j f_{ij}.$$

This is interpretable as the probability of identity by descent of a random allele from population i with a random allele from anywhere in the sample or, alternatively, as a measure of the similarity of the gene frequencies of population i to the sample mean gene frequencies. If no population is much more isolated than the others, it measures the relative size of population i.

The sample relationship coefficients then give a measure of distance between populations i and j as

$$d_{ij} = r_{ii} + r_{jj} - 2r_{ij}.$$

This measure has expected value

$$E(d_{ij}) \cong \frac{f_{ii} + f_{jj} - 2f_{ij}}{1 - \bar{f}} = \frac{\Delta_{ij}}{1 - \bar{f}},$$

which is the measure of distance suggested above, apart from the constant $(1 - \bar{f})$ in the denominator. Note that the circulation of this distance measure is like the calculation of a chi-squared statistic. With the summation referring to summation over all alleles,

$$d_{ij} \cong \sum\left(\frac{1}{\bar{p}(1 - \bar{p})}\right)[(p_i - \bar{p})^2 + (p_j - \bar{p})^2 - 2(p_i - \bar{p})(p_j - \bar{p})]$$

$$\approx \sum \frac{(p_i - p_j)^2}{\bar{p}},$$

since $\sum \bar{p}$ is an integer equal to the number of loci. However, it is preferable to compute first the matrix of kinship coefficients, because this matrix gives Wahlund F as the average diagonal element and because the eigenvectors of this matrix give a "genetic map" of the sample.

There are many routines that convert a table of distances into dendrograms or trees. Attempts to justify any particular routine as superior to others because of its "reconstruction" of evolution seem unsatisfactory and inapplicable to interbreeding human populations. We prefer a simple "maximum-linkage" technique, which is economical of computer time (Jenkins et al. 1971). The trees given by various techniques are usually similar in broad outline but differ in detail. It is difficult to evaluate or to say anything meaningful about differences among trees. Figure 27 shows trees produced from our material (see below) by the maximum and minimum linkage techniques (Jenkins et al. 1971).

MAPS

Much more satisfactory visual aids for the interpretation of genetic distances are provided by genetic maps (Morton et al. 1971c). A genetic map is simply the result of a principal components analysis of the kinship matrix. To do this, we transform gene frequencies into new, imaginary gene frequencies (more precisely, imaginary scaled deviations from the sample mean) that have the following properties: (1) a population's frequency of any of the imaginary genes has no relationship to its frequency on any of the others, that is, the gene frequencies are uncorrelated; and (2) the variability of the imaginary gene frequencies among the populations may be ranked in descending order, so that a plot of the populations on axes representing the two or three most variable genes gives a good picture of the biological relationships or distances among the populations.

For convenience, we work with three alleles, which may or may not be at the same locus, in three populations, but the procedure is perfectly general. We write, p, q, and r for the gene frequencies and label the populations with subscripts i, j, and k. Then, the kinship matrix is, apart from a scalar divisor (that is, three, because there are three alleles pooled):

$$R = \begin{matrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33}, \end{matrix}$$

where, for example,

$$r_{11} = \frac{(p_i - \bar{p})^2}{\bar{p}(1-\bar{p})} + \frac{(q_i - \bar{q})^2}{\bar{q}(1-\bar{q})} + \frac{(r_i - \bar{r})^2}{\bar{r}(1-\bar{r})}$$

$$r_{12} = \frac{(p_i - \bar{p})(p_j - \bar{p})}{\bar{p}(1-\bar{p})} + \frac{(q_i - \bar{q})(q_j - \bar{q})}{\bar{q}(1-\bar{q})} + \frac{(r_i - \bar{r})(r_j - \bar{r})}{\bar{r}(1-\bar{r})}$$

$$r_{22} = \frac{(p_j - \bar{p})^2}{\bar{p}(1-\bar{p})} + \frac{(q_j - \bar{q})^2}{\bar{q}(1-\bar{q})} + \frac{(r_j - \bar{r})^2}{\bar{r}(1-\bar{r})}.$$

This may be written as the product of a matrix and its transpose:

$$R = \begin{pmatrix} \frac{(p_i - \bar{p})}{\sqrt{\bar{p}(1-\bar{p})}} & \frac{(q_i - \bar{q})}{\sqrt{\bar{q}(1-\bar{q})}} & \frac{(r_i - \bar{r})}{\sqrt{\bar{r}(1-\bar{r})}} \\ \frac{(p_j - \bar{p})}{\sqrt{\bar{p}(1-\bar{p})}} & \frac{(q_j - \bar{q})}{\sqrt{\bar{q}(1-\bar{q})}} & \frac{(r_j - \bar{r})}{\sqrt{\bar{r}(1-\bar{r})}} \\ \frac{(p_k - \bar{p})}{\sqrt{\bar{p}(1-\bar{p})}} & \frac{(q_k - \bar{q})}{\sqrt{\bar{q}(1-\bar{q})}} & \frac{(r_k - \bar{r})}{\sqrt{\bar{r}(1-\bar{r})}} \end{pmatrix}$$

$$\begin{pmatrix} \frac{(p_i - \bar{p})}{\sqrt{\bar{p}(1-\bar{p})}} & \frac{(p_j - \bar{p})}{\sqrt{\bar{p}(1-\bar{p})}} & \frac{(p_k - \bar{p})}{\sqrt{\bar{p}(1-\bar{p})}} \\ \frac{(q_i - \bar{q})}{\sqrt{\bar{q}(1-\bar{q})}} & \frac{(q_j - \bar{q})}{\sqrt{\bar{q}(1-\bar{q})}} & \frac{(q_k - \bar{q})}{\sqrt{\bar{q}(1-\bar{q})}} \\ \frac{(r_i - \bar{r})}{\sqrt{\bar{r}(1-\bar{r})}} & \frac{(r_j - \bar{r})}{\sqrt{\bar{r}(1-\bar{r})}} & \frac{(r_k - \bar{r})}{\sqrt{\bar{r}(1-\bar{r})}} \end{pmatrix}$$

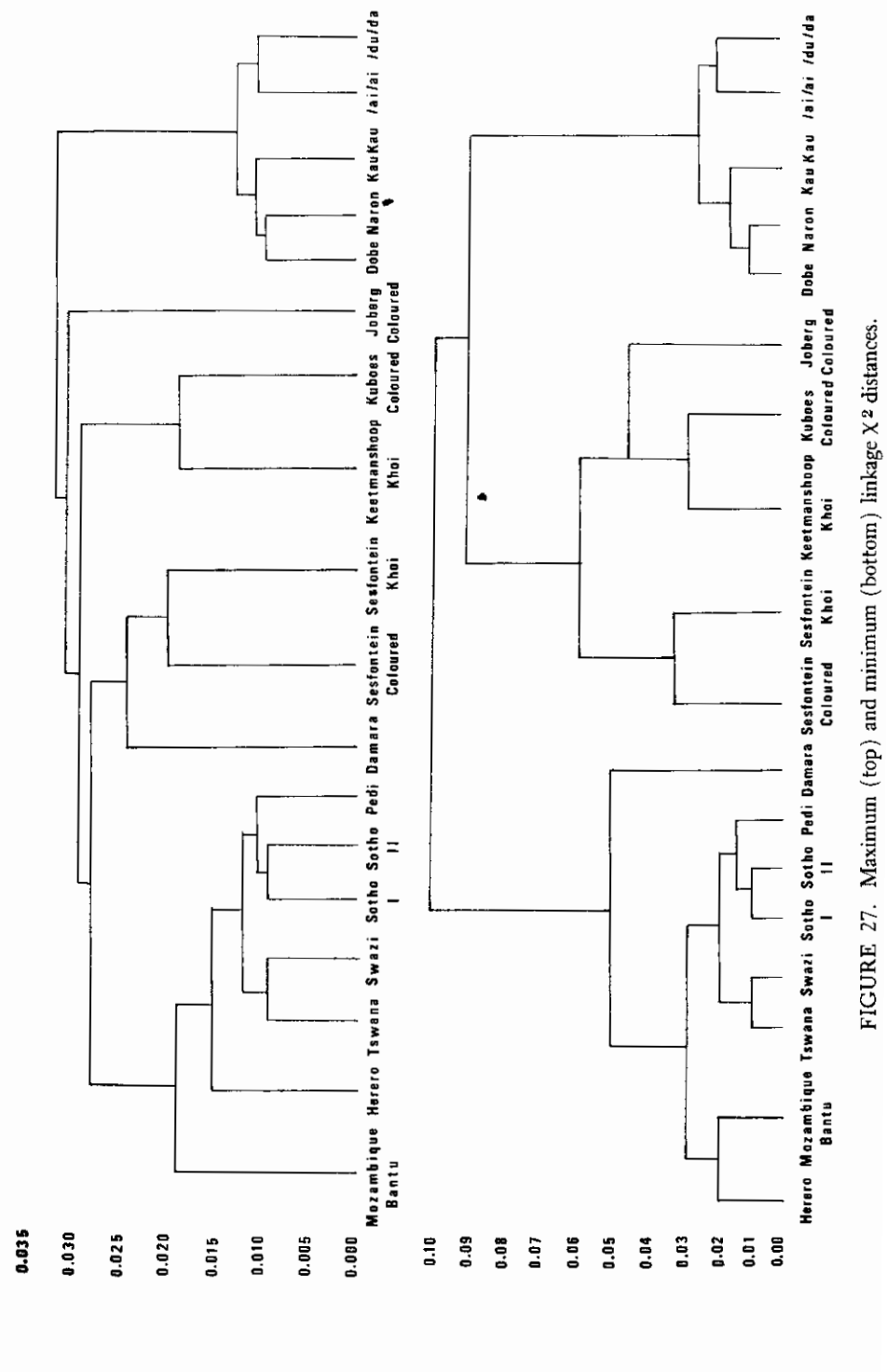


FIGURE 27. Maximum (top) and minimum (bottom) linkage X^2 distances.

For future reference, we call the matrices on the right of this equation Z and Z^T, so $R = ZZ^T$.

We write our new, imaginary gene frequencies for population i as

$$e_{1i}\sqrt{\lambda_1},\ e_{2i}\sqrt{\lambda_2},\ e_{3i}\sqrt{\lambda_3}.$$

The reasons for this notation will be apparent. We wish our new frequencies to reconstruct the original kinship matrix R, that is,

$$R = \begin{pmatrix} e_{1i}\sqrt{\lambda_1} & e_{2i}\sqrt{\lambda_2} & e_{3i}\sqrt{\lambda_3} \\ e_{1j}\sqrt{\lambda_1} & e_{2j}\sqrt{\lambda_2} & e_{3j}\sqrt{\lambda_3} \\ e_{1k}\sqrt{\lambda_1} & e_{2k}\sqrt{\lambda_2} & e_{3k}\sqrt{\lambda_3} \end{pmatrix} \begin{pmatrix} e_{1i}\sqrt{\lambda_1} & e_{1j}\sqrt{\lambda_1} & e_{1k}\sqrt{\lambda_1} \\ e_{2i}\sqrt{\lambda_2} & e_{2j}\sqrt{\lambda_2} & e_{2k}\sqrt{\lambda_2} \\ e_{3i}\sqrt{\lambda_3} & e_{3j}\sqrt{\lambda_3} & e_{3k}\sqrt{\lambda_3} \end{pmatrix}$$

Write E as a matrix consisting of the column vectors e_{1-}, e_{2-}, e_{3-}, and scale these vectors so that the sum of squares of their elements is one and the magnitudes of the λ's express the variability of the new gene frequencies; then the previous matrix equality may be written

$$R = \begin{pmatrix} e_{1i} & e_{2i} & e_{3i} \\ e_{1j} & e_{2j} & e_{3j} \\ e_{1k} & e_{2k} & e_{3k} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} \begin{pmatrix} e_{1i} & e_{1j} & e_{1k} \\ e_{2i} & e_{2j} & e_{2k} \\ e_{3i} & e_{3j} & e_{3k} \end{pmatrix}$$

or

$$R = E\Lambda E^T,$$

where Λ is a diagonal matrix with elements λ_1, λ_2, λ_3.

Since the matrix R was written originally as the product of a matrix and its transpose, it is symmetric and positive semidefinite. The matrix E and numbers λ are the eigenvectors and eigenvalues of R. A discussion of the algebra of eigenvectors and eigenvalues as used here may be found in Tatsuoka (1971). Because R is symmetric, the matrix E is orthogonal, that is, $EE^T = I$, or, in our terms, the imaginary gene frequencies are uncorrelated. Further, since R is positive semidefinite, the eigenvalues are all greater than or equal to zero. At least one will be zero, because k populations occupy a space of $k - 1$ or fewer dimensions, and several more may be zero if fewer independent alleles than populations are studied.

If the eigenvalues and their associated vectors are arranged and labeled in descending order of magnitude, the imaginary gene frequencies $e_{1i}\sqrt{\lambda_1}$, $e_{1j}\sqrt{\lambda_1}$, $e_{1k}\sqrt{\lambda_1}$ will array the populations along an axis, a rotation of the original gene frequency axes, along which the dispersion is maximized. The second set of imaginary frequencies, $e_{2i}\sqrt{\lambda_2}$, $e_{2j}\sqrt{\lambda_2}$,

$e_{2k}\sqrt{\lambda_2}$, gives a second axis at right angles to the first, along which dispersion is again maximized, after the variation accounted for by the first axis has been removed. The first two or three axes found in this way then provide the "best" reduced dimension picture of the "distance" relations among the groups.

Several remarks concerning the relation of this procedure to ordinary principal components analysis are in order. In many studies, when the positions of objects on component axes are shown, the eigenvectors themselves are plotted, rather than the eigenvectors multiplied by the square root of the corresponding eigenvalue. Since the eigenvectors are scaled so that the sum of their squared elements is equal to one, the plots appear spherical; this may be misleading if the eigenvalues—the magnitudes of dispersion along the corresponding axes—are very different.

Principal components analysis is widely used to array objects in a new space when these objects have been measured on a number of metric traits. Often the reduced-space representation is very satisfactory, since much of the variation can be accounted for by a few components, usually identifiable as size, robusticity, linearity, and so forth. When this method of analysis is used in this population-genetic context, there is no reason to expect that a small number of new variables will account for the relations among the groups in a satisfactory manner, since under pure drift, no locus should be correlated in any way with any other locus. This method should be valuable, on the other hand, in identifying clusters of related groups, since relatedness should be the only factor introducing correlations among groups. In morphometric studies, items like total size seem to become confounded with relatedness.

In morphometric studies, a decision must be made whether to find components of the correlation matrix, which gives all variables equal weight, or of the covariance matrix, which weights variables by their magnitude. Here, gene frequency covariances are divided by the scaling factor $\bar{p}(1 - \bar{p})$ derived from genetic drift theory. This may not be the most appropriate procedure if the object is discrimination, but it has the advantage that it articulates distance analysis with genetic theory and that it provides a map optimally corresponding to the population structure that determines drift.

Finally, in ordinary components analysis, a covariance or correlation matrix among the variables is computed, the eigenvalues and eigenvectors determined, and the component scores (our imaginary gene frequencies)

determined by multiplying the eigenvectors by the variable values for each object. The procedure described here is a shortcut and offers further advantage of yielding a distance matrix (if desired) and population structure statistics along the way. It is particularly useful, and economical of computing effort, when many more alleles than populations are studied.

PLOTTING ALLELES

It will usually be of considerable interest to go back and look at the relations among the alleles as well as among the populations. Coordinates for alleles along imaginary axes of greatest variation may be obtained from the eigenvectors of the matrix R by simple matrix multiplication; these coordinates will be eigenvectors of the scaled matrix of covariances among allele frequencies, which we write S.

The matrix Z defined on p. 186 of scaled data values will have k rows and p columns, where p is the number of alleles and k is the number of populations under study. Apart from scalar divisors, the k by k matrix R is equal to Z postmultiplied by its transpose, while the p by p matrix S is equal to Z premultiplied by its transpose. If k is much smaller than p, that is, if there are many more alleles than populations, there are, at most, k eigenvectors of R, imaginary gene frequencies along which the populations are arrayed. These will always be sufficient to reconstruct the kinship or distance matrices, since k populations occupy a hyperspace of, at most, $k - 1$ dimensions. For example, in the three-population example, no matter how many alleles are studied, three populations define a two dimensional space, and distances between them may be drawn exactly in two dimensions. In general, the dimension of the space occupied by k populations defined on p allele frequencies is the minimum of $k - 1$ and p, so that either the S or R matrix contains all the information available about kinship and genetic distance. If, as in this study, k is much smaller than p, it is much more convenient to work with the matrix R. If, on the other hand, p is much smaller than k, it is convenient to work with the scaled covariance matrix among alleles S.

The important algebraic result that allows the use of either matrix for this analysis is that S and R have exactly the same set of nonzero eigenvalues. This means that the dimensions of the space in which the populations are located are exactly the same, whether we imagine populations as points on axes corresponding to gene frequencies, or whether we

imagine gene frequencies as points on axes corresponding to populations. Further, the eigenvectors of one are given by multiplying the corresponding eigenvectors of the other by the scaled data matrix Z (Dempster 1969). In matrix notation, let the number of nonzero eigenvalues of S and R be l. Then the l by p matrix of eigenvectors of S is given by the product of the l by k matrix of eigenvectors of R postmultiplied by the k by p matrix of scaled data Z. In general these will have to be rescaled. Then, the p-element eigenvectors of S corresponding to the first few largest eigenvalues give an optimum graphic portrayal of the correlations among the alleles and of their contributions to genetic distances along the corresponding axes.

NUMERICAL EXAMPLE

A numerical example of this procedure using three populations and three gene frequencies may clarify the meaning of the operations. Let populations I, J, and K have gene frequencies p, q, and r as follows:

	I	J	K		
p	0.4	0.8	0.6	$\bar{p} = 0.6$	$\bar{p}(1-\bar{p}) = 0.24$
q	0.5	0.2	0.8	$\bar{q} = 0.5$	$\bar{q}(1-\bar{q}) = 0.25$
r	0.5	0.5	0.8	$\bar{r} = 0.6$	$\bar{r}(1-\bar{r}) = 0.24.$

First, the kinship matrices are formed for each allele, as

$$R_p = \begin{matrix} 0.17 & -0.17 & 0.00 \\ -0.17 & 0.17 & 0.00 \\ 0.00 & 0.00 & 0.00 \end{matrix} \qquad R_q = \begin{matrix} 0.00 & 0.00 & 0.00 \\ 0.00 & 0.36 & -0.36 \\ 0.00 & -0.36 & 0.36 \end{matrix}$$

$$R_r = \begin{matrix} 0.04 & 0.04 & -0.08 \\ 0.04 & 0.04 & -0.08 \\ -0.08 & -0.08 & 0.16, \end{matrix}$$

and averaged:

$$R = \begin{matrix} 0.07 & -0.04 & -0.03 \\ -0.04 & 0.19 & -0.15 \\ -0.03 & -0.15 & 0.18. \end{matrix}$$

Inspection of this matrix shows that the average diagonal element, Wahlund F, is 0.15, which is of the order of those found for comparisons

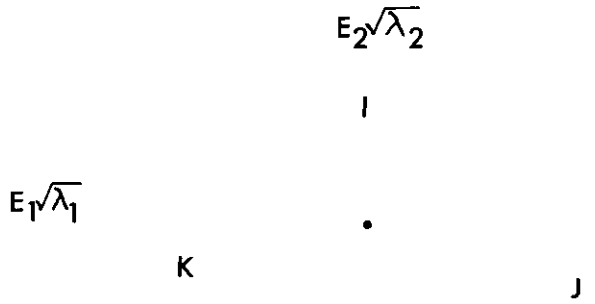
among major racial groups. Second, the inbreeding coefficient of the first population (r_{11}) is small, implying that it is nearest the center of the swarm. Third, the sum of any row or column is zero. In general, it is true that sample random kinship $\bar{r} = \sum_{ij} w_i w_j r_{ij}$, or random kinship of subpopulation j, $\bar{r}_j = \sum_k w_k r_{jk}$, is zero. This corresponds to the intuitive notion that the correlation between two random populations is zero, since correlation is calculated from observed sample mean gene frequencies.

The eigenvectors and eigenvalues of this matrix are:

$\lambda_1 = 0.34$	$e_{1I} = -0.03$	$e_{1J} = 0.72$	$e_{1K} = -0.69$
$\lambda_2 = 0.10$	$e_{2I} = 0.82$	$e_{2J} = -0.38$	$e_{2K} = -0.43$
$\lambda_3 = 0.00$	—	—	—

The sum of the eigenvalues, 0.44, is the sum of the diagonal elements of the original matrix. The eigenvalues measure the dispersion along the new axes; the total dispersion is preserved under the rotation and new representation and given by the sum of the eigenvalues. The third eigenvalue is zero, implying that the three populations occupy a space of only two dimensions, as is obvious. The two eigenvectors corresponding to the two positive eigenvalues are natural coordinate systems for showing the relations among the groups. The squares of the elements of each vector sum to one, so the dispersion along each axis will be the same if the vectors are plotted. More naturally, each vector should be multiplied by the square root of its eigenvalue to show "true" relations (Figure 28).

FIGURE 28. Distance relations among three hypothetical populations.



Now, to obtain graphic representation of the relations among the alleles, we obtain the eigenvectors of S from those of R.

$$Z = \begin{pmatrix} \frac{(0.4-0.6)}{\sqrt{0.24}} & 0 & \frac{(0.5-0.6)}{\sqrt{0.24}} \\ \frac{(0.8-0.6)}{\sqrt{0.24}} & \frac{(0.2-0.5)}{\sqrt{0.25}} & \frac{(0.5-0.6)}{\sqrt{0.24}} \\ 0 & \frac{(0.8-0.5)}{\sqrt{0.25}} & \frac{(0.8-0.6)}{\sqrt{0.24}} \end{pmatrix},$$

$$E^T = \begin{pmatrix} -0.03 & 0.72 & -0.69 \\ 0.82 & -0.38 & -0.43 \\ — & — & — \end{pmatrix},$$

and

$$E^TZ = \begin{matrix} \\ \\ \\ \end{matrix} \begin{matrix} p \\ \\ \\ \end{matrix} \begin{matrix} q \\ \\ \\ \end{matrix} \begin{matrix} r \\ \\ \\ \end{matrix} \begin{pmatrix} 0.30 & -0.85 & -0.42 \\ -0.16 & 0 & -0.24 \\ — & — & — \end{pmatrix},$$

approximately. The rows of E^TZ are eigenvectors corresponding to the two nonzero eigenvalues of the scaled covariance matrix among alleles given by $S = Z^TZ$, apart from some scalar row multipliers to make the sum of squares of each vector equal to unity. These relations are shown in Figure 29. The axes are the same in Figures 28 and 29 and reveal immediately, for example, that allele q contains no information for telling population I from J and K, or that allele q is very low in population j.

FIGURE 29. Relations among alleles.

ANALYSIS OF SOUTHERN AFRICAN DATA

We have applied these techniques to data on 15 marker loci from 18 populations in South Africa (Jenkins, Zoutendyk, and Steinberg 1970, Jenkins et al. 1971, Jenkins and Steinberg, unpublished data). These publications will present the results of the various genetic analyses in detail.

The populations studied fall into four main subdivisions; San, Khoi, Bantu, and Colored. We use the new terminology of San for "Bushman" and Khoi for "Hottentot." The provenance of the samples from the Bantu-speaking peoples has been given in Jenkins, Zoutendyk, and Steinberg (1970), except for the Herero sample, which was collected by Harpending during fieldwork in northwest Botswana with !Kung San. The samples from !Kung, Naron, the three groups from Sesfontein, southwest Africa (Damara, Sesfontein Colored, and Khoi), and the Johannesburg Colored are described in Jenkins et al. (1971) and Harpending and Jenkins (1972b). The Khoi from Keetmanshoop, Southwest Africa, and the Colored from Kuboes are described in Jenkins and Corfield (1972). The 18 populations are shown in Figure 30.

Since we are interested in the distance relations among these populations, we have treated them as if they had equal weight in the calculation of the R matrix given in Table 29. If this were a study of population struc-

FIGURE 30. Location of populations in southern Africa.

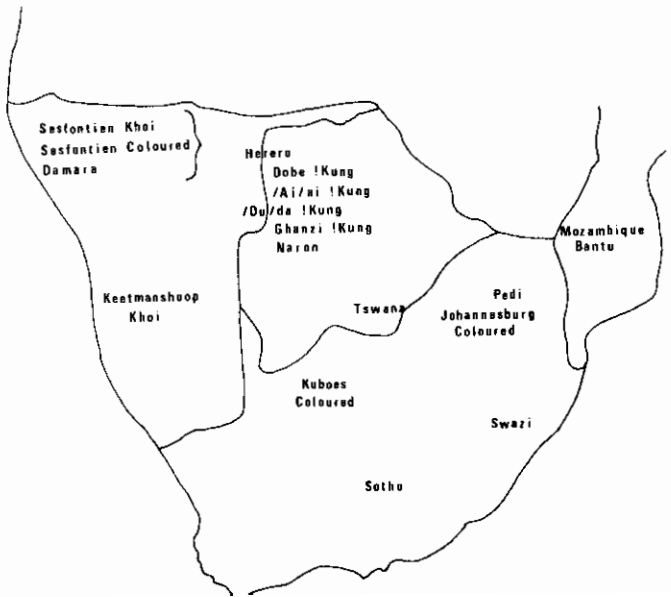


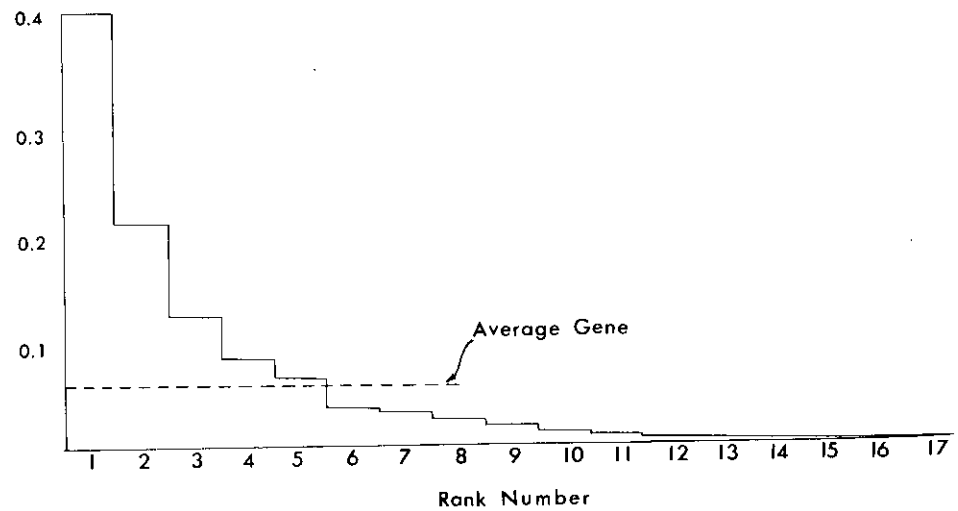
TABLE 29
RELATIONSHIP MATRIX AMONG 18 SOUTHERN AFRICAN POPULATIONS ($\times 10^3$)

1.	056	047	060	038	036	−010	001	−032	−024	−012	−016	−024	−018	−030	−025	−013	001	−037
2.		070	070	037	035	−014	−004	−043	−013	−017	−013	−017	−025	−030	−030	−019	−007	−028
3.			098	057	045	−006	−008	−041	−027	−023	−022	−026	−035	−036	−026	−023	−007	−050
4.				057	038	−008	009	−036	−024	−021	−018	−021	−027	−028	−019	−002	011	−043
5.					040	−009	001	−034	−014	−013	−013	−014	−016	−020	−022	−012	002	−030
6.						041	009	011	−003	−003	−001	000	−014	−004	027	−012	013	−018
7.							060	−018	−015	−007	−012	−006	−016	−015	−003	004	040	−020
8.								088	015	011	011	003	024	009	043	−027	−023	039
9.									309	009	010	013	019	017	006	−020	028	038
10.										021	011	037	019	020	003	−010	−021	016
11.											018	017	014	018	003	008	−017	017
12.												033	012	026	003	−010	−024	017
13.													037	021	−000	−010	−023	038
14.														044	002	000	−019	025
15.															051	−009	−010	006
16.																144	043	−020
17.																	102	−034
18.																		084

ture, attention would have to be given to the varying census sizes of these groups and to the biases in these relationship statistics from the variability in sample sizes. For our purposes here, those considerations are not important, since the genetic distances do not depend (very much) on how mean gene frequencies are defined, or on the niceties of biased versus unbiased estimates of R. Nevertheless, it is of some note that F_{ST}, or Wahlund F, for all these groups considered together is only 0.06, which is on the order of that found within single tribes or small areas among tropical gardeners (Friedlaender 1971a), and that F_{ST} calculated by weighting groups by census size would be even smaller.

The eigenvalues of this matrix are plotted in descending order in Figure 31. Since the average allele has a scaled variance of 0.06, we may, as a rule of thumb, consider a dimension corresponding to an eigenvalue greater than this to be significant. There are only five such eigenvalues, but the first two or three seem clearly larger than the others. Most of the variation will be described by the first two or three axes, but it will be worth-

FIGURE 31. Eigenvalues of R.



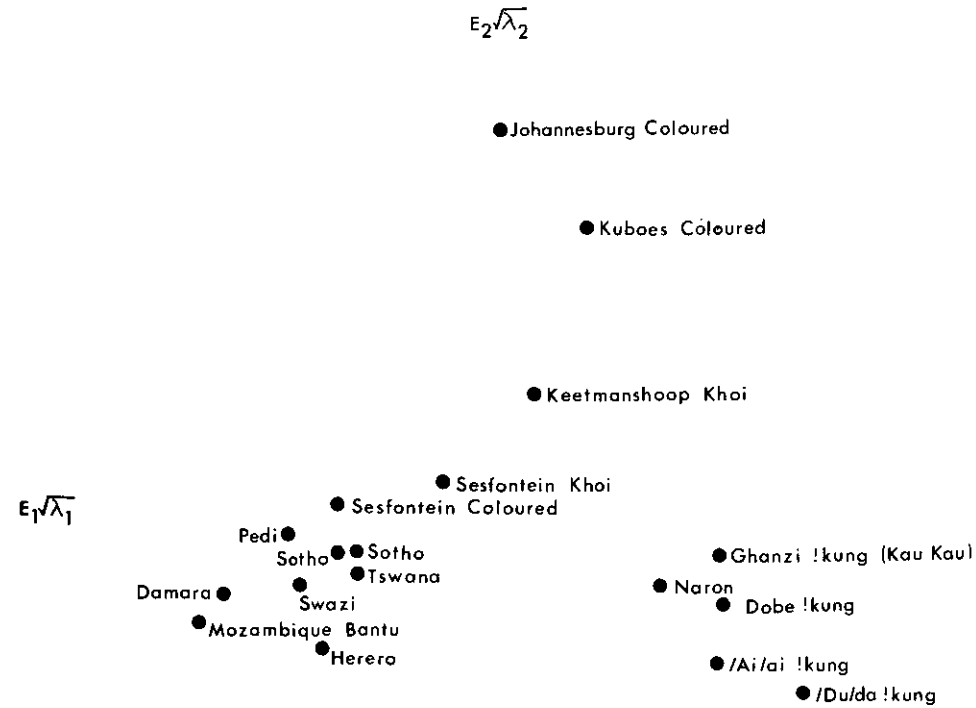
while to examine the others to see if any one group or cluster of groups is being differentiated by this axis.

The 18 groups are plotted on dimensions corresponding to the first two eigenvectors in Figure 32, and the alleles used are plotted on the first two eigenvectors of the S matrix in Figure 33. The first four eigenvectors of R are given in Table 30, along with the corresponding eigenvalues.

RELATIONS AMONG POPULATIONS

The first axis in Figure 32 is clearly differentiating Bantu-speaking and Khoisan-speaking peoples. It is worthy of note that the extreme Khoisan population, the /Du/da !Kung, is the !Kung population most isolated

FIGURE 32. Populations plotted on first two scaled eigenvectors.



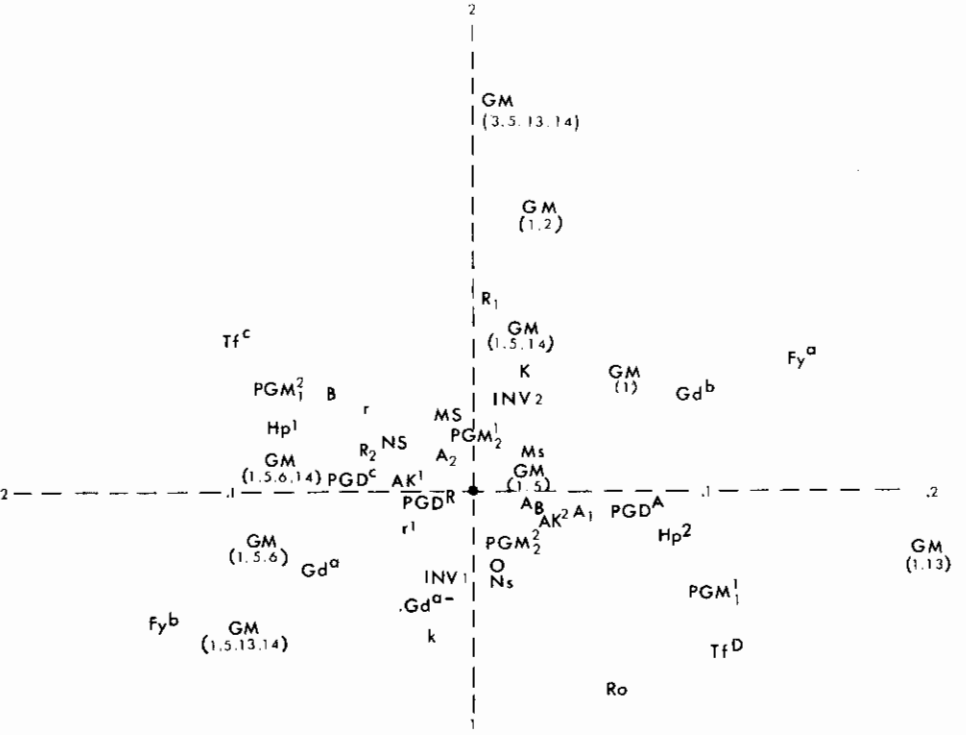


FIGURE 33. Plot of alleles on first two eigenvectors of S (eigenvectors not scaled).

from contact with Tswana and Herero Bantu-speakers. The Mozambique Bantu sample is at the other extreme of the first axis, confirming the low Khoisan admixture into this sample deduced by Jenkins, Zoutendyk, and Steinberg (1970). The ordering of the other Bantu groups on this axis also conforms to the admixture estimates in that publication.

The second axis in Figure 32 is just as clearly revealing non-African admixture. The Johannesburg Colored population is highest on this dimension, followed by that of Kuboes, also from South Africa, the Khoi from Keetmanshoop in southern Southwest Africa, and last by the populations at Sesfontein in the extreme northwest part of Southwest Africa,

TABLE 30
EIGENVECTORS OF KINSHIP MATRIX

Population	Eigenvectors 1	2	3	4
1. Dobe !Kung	0.31	−0.15	−0.03	0.06
2. /Ai/ai !Kung	0.32	−0.23	−0.13	0.00
3. /Du/da !Kung	0.42	−0.26	0.00	0.22
4. Ghanzi !Kung	0.32	−0.05	0.01	0.05
5. Naron	0.25	−0.12	−0.06	−0.05
6. Sesfontein Khoi	−0.03	0.08	0.41	0.06
7. Keetmanshoop Khoi	0.09	0.23	0.26	−0.47
8. Damara	−0.31	−0.10	0.37	0.42
9. Herero	−0.19	−0.15	−0.10	−0.10
10. Tswana	−0.14	−0.07	−0.09	−0.09
11. Sotho I	−0.14	−0.06	−0.09	−0.05
12. Sotho II	−0.16	−0.06	−0.12	−0.18
13. Swazi	−0.21	−0.09	−0.17	−0.07
14. Pedi	−0.22	−0.01	−0.20	−0.15
15. Sesfontein Colored	−0.17	0.03	0.39	0.36
16. Johannesburg Colored	0.04	0.65	−0.49	0.46
17. Kuboes Colored	0.15	0.52	0.25	−0.33
18. Mozambique Bantu	−0.33	−0.16	−0.20	−0.12
Eigenvalue	0.41	0.22	0.13	0.08
Cumulative percent of total kinship	38	58	70	78

furthest from areas of dense European settlement. The samples from Sesfontein and Keetmanshoop labeled Khoi were chosen as people who claimed four Khoi grandparents. In general, this analysis shows that such individuals are not significantly "purer" Khoi and that, indeed, the Khoi and Colored groups are not genetically distinguishable. This point is reinforced by the location in Figure 32 of the Naron, who are hunter-gatherers of the central Kalahari Desert around Ghanzi, Botswana, and who speak a language mutually intelligible with Nama; in other words, they are linguistically Khoi. The results of this analysis suggest that there are no significant differences between the San and Khoi (or "Bushman" and "Hottentot") peoples of South Africa and that some of the ascribed differences are the results of admixture, a different way of life with a different diet, and so forth. The populations in southwest Africa who call themselves Khoi seem fully genetically allied with the Colored populations around them. However, these findings are subject to different possi-

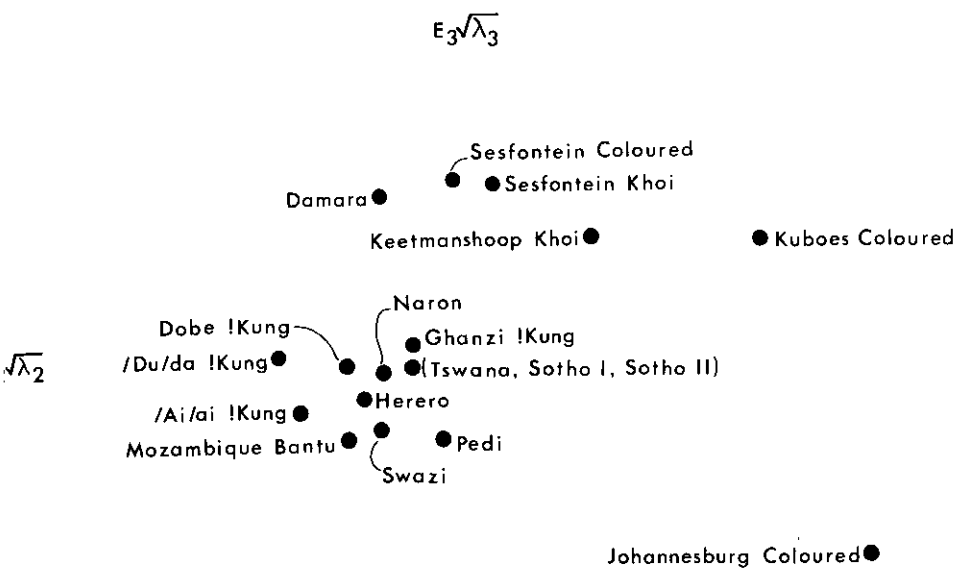


FIGURE 34. Populations plotted on second and third scaled eigenvectors.

blc interpretations, especially when the third dimension is considered. Figure 34 shows these populations arrayed on axes 2 and 3. When looking at this, it should be kept in mind that the San and Khoi-Colored are separated on axis 1 from the Bantu and that the Damara are good Bantu on the first axis.

The San (including the Naron) and the Bantu groups are indifferent on the third axis; on it, the Johannesburg Colored population is alone at one extreme, while the rest of the Damara-Khoi-Colored complex occupies the other. This might be taken to indicate that the composition of the Johannesburg population is basically different from that of the other Colored groups in southern Africa, possibly reflecting Malay admixture. On this axis, the Damara are in the midst of the Khoi-Colored complex, while on the first two axes, they are clearly Bantu. This confirms their puzzling status among these populations and indicates that an extension of this analysis to include populations from the rest of the continent would be useful. See Jenkins et al. (1971) for a discussion of the Damara.

The contribution of the alleles studied to the variation along each of the axes is shown graphically in Figure 33. Such a presentation immediately identifies alleles highly associated with San to be $Gm^{1,13}$, Fy^a, Tf^D, Hp^2, and $PGM_1{}^1$, and alleles indicative of extra-African admixture to be $Gm^{3,5,13,14}$, $Gm^{1,2}$, R_1, Tf^Q, and so forth. This kind of representation seems to us very useful for studies of genetic distance among heterogeneous groups, since allelic variation is comprehensible at a glance. A collection of plots such as this from various areas might provide the simplest way to scan for associations among alleles consistent over different areas.

NOTE

1. Present address: Department of Anthropology, University of New Mexico.