Ann. Hum. Genet., Lond. (1973), **36**, 459
Printed in Great Britain

Genetic studies on hybrid populations

II. Estimation of the distribution of ancestry

BY CHARLES J. MACLEAN

Population Genetics Section, Human Genetics Branch,
National Institute of Dental Research,
National Institutes of Health, Bethesda, Maryland 20014

AND PETER L. WORKMAN

Division of Medical Genetics, Mt. Sinai School of Medicine,
New York, New York 10029

INTRODUCTION

The analysis of intermixture in a dihybrid population has generally entailed estimation of the proportions of the hybrid gene pool derived from each parental population. Given the frequencies of a gene in the hybrid and parental populations, Bernstein's (1931) formula has been used to estimate the relative ancestral contributions to numerous human dihybrid populations: United States Negroes (Glass & Li, 1953; Pollitzer, 1958; Workman, Blumberg & Cooper, 1963; Reed, 1969); Brazilian Negroes (Salzano, 1963; Salzano & Hirschfeld, 1965); Chileans (Saldanha & Nacrui, 1963; Nagel & Soto, 1964); Mexican mestizos (Rodriguez *et al.* 1963), etc. The estimate from each gene may describe the effects of a single generation of intermixture or the cumulative effect of several generations of intermixture in varying amounts. Genetics observations on different loci can be combined, with certain restrictions, to provide a single joint estimate of the ancestry of the dihybrid population by the method of maximum-likelihood (Krieger *et al.* 1965) or by least-squares analyses (see discussion in Elston, 1971).

If the hybrid population is in genetic equilibrium then all individuals will have the same expected degree of ancestry – that is, the same proportion of their genes derived from the parental populations. Of course, even for full sibs there can be some difference due to sampling effects in the formation of gametes. For a single locus, a single generation of random mating, without associated intermixture, will result in an equilibrium (Hardy–Weinberg) distribution, but, for n loci, the rate of approach to equilibrium will depend both upon n and the pattern of linkage among the loci. Assortative mating related to ancestral origins will retard the approach to equilibrium and continuous immigration from one or both parental populations will result in a continual state of genetic disequilibrium. In general, the contribution from each parental population to an individual in the hybrid population can vary between 0 and 100 %, and in some populations the variation in ancestry among individuals may be considerable.

Since complete pedigrees showing ancestral origins are rarely, if ever, available, it is generally not possible to determine the true proportion of genes which an individual derives from each ancestral population. However, using estimates of the gene frequencies in the parental populations and a characterization of the phenotypes of individuals in the hybrid population for an arbitrary number of polymorphic loci, MacLean & Workman (1972) provide a method for estimating the probability distribution of proportion of ancestry for each individual. In this

paper, assuming similar data, we provide a method for estimating the relative frequency of individuals of every proportion of ancestry. Knowledge of the form of this frequency distribution provides insight into the history of intermixture in the population. In addition, if a hybrid population is shown to be heterogeneous with respect to ancestry, then an analysis of quantitative variation in relation to variation in individual ancestry can be performed as described by MacLean & Workman (1972).

THE MODEL

Let Q_0 and Q_1 denote random mating populations at equilibrium, and suppose that a distinct hybrid population has been formed over time by intermixture of randomly drawn migrants from each. We confine our analysis to two ancestral populations; the extension to more, although complicated, is trivial conceptually. Let θ denote an individual's proportion of ancestry from Q_1. The relative frequency of individuals of proportion θ in the hybrid population is specified as a probability density, $g(\theta)$, over the domain [0, 1].

The method we shall describe employs the relationships among the gene frequencies of independently assorting loci in order to estimate the moments of $g(\theta)$, and subsequently employs the moments to estimate the form of $g(\theta)$. The analysis rests upon the assumptions that for a certain number of polymorphisms (blood types, serum proteins, etc.) the gene frequencies in the ancestral populations are accurately known, and that only intermixture, not selection or mutation, has occurred in the hybrid population. Because of redundancy in the estimation process, it will be possible for us to check these assumptions for each locus against the rest.

Since the proportion of phenotypes of a locus depend upon the breeding structure of the population with respect to θ (see MacLean & Workman, 1972), we must deal with gene frequencies rather than phenotype frequencies. In co-dominant loci we can estimate gene frequencies without regard to panmixia, etc. However, in loci at which there is dominance, or in closely linked loci (e.g. Rh), there are several complications. The problem of estimating such gene frequencies is treated by Fisher (1940), Cotterman (1947), Ceppellini *et al.* (1955), Smith (1957, 1967) and others.

In a locus with k alleles, the gene frequencies of only $k-1$ of these contain information; it is irrelevant which ones are used. Suppose that we measure m' loci with a total of m'' genes. Then we can use

$$m = m'' - m'$$

genes in the analysis.

ESTIMATION OF MOMENTS OF $g(\theta)$

The basis of estimation of $g(\theta)$ is the conditional gene frequency, $p(A/\theta)$, the frequency of gene A in hybrid individuals with proportion of ancestry θ, the form of which is well known (Bernstein, 1931). If the frequencies of A in Q_1 and Q_0 are denoted by V and W, and $D = V - W$, then

$$p(A/\theta) = \theta D + W. \tag{1}$$

Estimation of the mean proportion of ancestry is usually derived from a consideration of the total hybrid gene pool, but can as easily be derived from $g(\theta)$. We relate $p(A)$, the observed proportion of gene A in the hybrid sample, to (1) by integration over $g(\theta)$. That is,

$$\begin{aligned} p(A) &= \int p(A/\theta)\, g(\theta)\, d(\theta) \\ &= D \int \theta g(\theta) d\theta + W \int g(\theta)\, d\theta \\ &= DE(\theta) + W. \end{aligned} \tag{2}$$

Each gene yields an estimate of $E(\theta)$ in this way. We shall discuss reconciling these estimates in the next section.

Estimation of higher moments of $g(\theta)$ rests upon the frequency of combinations of genes from unlinked loci. This requires that we derive the frequency of sets of genes in a gamete, when what we directly observe are individuals. For example, in the case of two unlinked loci and their genes A, a and B, b respectively, each $AABb$ individual yields one AB gamete and one Ab gamete. Dominance causes complications, and linked genes cannot be used at all.

We have shown in another paper (MacLean & Workman, 1972) that within a subset of individuals of proportion θ, independently assorting genes are also statistically independent, but that they are not independent over the population. That is, at every value θ we have

$$p(AB/\theta) = p(A/\theta)\, p(B/\theta).$$

The population frequency of AB, observed in our sample, again arises from that of the members through integration over $g(\theta)$, so that by replacing $p(A/\theta)$ and $p(B/\theta)$ with appropriate values from (1), we have

$$\begin{aligned} p(AB) &= \int (\theta D_A + W_A)(\theta D_B + W_B)\, g(\theta)\, d\theta \\ &= D_A D_B\, E(\theta^2) + (W_A D_B + D_A W_B)\, E(\theta) + W_A W_B. \end{aligned} \tag{3}$$

The relationship between $p(AB)$ derived from (3) and the product $p(A)\, p(B)$ from (2) yields information about the shape of $g(\theta)$, specifically about its variance. Since

$$p(A)\, p(B) = (D_A E(\theta) + W_A)\, (D_B E(\theta) + W_A)$$

we see from (3) together with (2) that

$$p(AB) - p(A)\, p(B) = D_A D_B \text{ var } (\theta).$$

The technique is easily extended to the nth case.

$$\begin{aligned} p \text{ (all } A_i) &= \int \prod_{i=1}^{n} (\theta D_i + W_i)\, g(\theta)\, d\theta \\ &= \sum_{k=0}^{n} C_k\, E(\theta^k), \end{aligned} \tag{4}$$

where C_k is the sum of products of all combinations of kD's and $(n-k)$ W's.

Equation (4) is a representative of a system of equations between observed gene frequencies and moments of $g(\theta)$. We count the joint frequency of every possible combination of n genes from the total of m genes under consideration. Each such combination yields one observation upon the moments of $g(\theta)$ up to $E(\theta^n)$. Each gene of course appears in several different combinations. The total number of observations is $\binom{m}{n}$; the number of combinations of m things taken n at a time. Although this may be a large number (for example, $\binom{10}{4} = 210$), the gene counting is so systematic that it can easily be accomplished by computing machine.

Naturally the observations upon the moments derived from (4) will to some extent contradict one another, partly from failure of our assumptions and partly from random variation. We treat these two sources separately.

RECONCILING ESTIMATES OF THE MOMENTS

In order to disclose failure in our assumptions about the data, we compare the m estimates of $E(\theta)$ which we obtain from (2). In those loci which yield estimates of $E(\theta)$ far outside the range of the others, either the ancestral gene frequencies are very different from those we have assumed, or else selection or mutation has occurred. The problem of determining which loci are deviant is treated by several authors (see Workman, 1968; Reed, 1969).

After elimination of the anomalous loci, we wish to average the estimates of the moments from those remaining. We assume that all the equations of system (4) contain random errors of observation.

$$p_j = \sum_{k=0}^{n} c_{jk} E(\theta^k) + \epsilon_j \quad j = 1, 2, \ldots, \binom{m}{n}. \tag{4a}$$

We use a standard least squares method for estimating all $E(\theta^k)$ from these equations. The procedure is fully described in Kendall & Stuart (1961, p. 87). In our case, let $\mathbf{p}$ be the $\binom{m}{n}$ dimensional vector of observed gene frequencies. Let $\mathbf{e}$ be the $(n+1)$ dimensional vector of unknown moments.

$$\mathbf{e} = \begin{pmatrix} 1 \\ E(\theta) \\ \vdots \\ E(\theta^n) \end{pmatrix}.$$

Let $\mathbf{C} = [c_{ik}]$ be the $\binom{m}{n} \times (n+1)$ matrix of sums of products of D's and W's. Let $\boldsymbol{\epsilon}$ be the $\binom{m}{n}$ dimensional vector of observation errors in the corresponding elements of $\mathbf{p}$, and finally let $\mathbf{V} = [\mathrm{cov}\ (\epsilon_i, \epsilon_j)]$ be the $\binom{m}{n}$ square error covariance matrix.

We shall assume that the correlation of the errors associated with two combinations of genes is proportional to the number of genes the two have in common. Although the relationships in some cases are much more complex, this simplified error model has yielded approximately optimal weightings in many numerical experiments. Our assumption is written

$$\mathbf{V} \propto [k_{ij}],$$

where k_{ij} is the number of genes which p_i and p_j have in common. The constant of proportionality of $\mathbf{V}$ is not needed.

In this notation, equation (4a) can be written

$$\mathbf{p} = \mathbf{Ce} + \boldsymbol{\epsilon},$$

whose least squares solution is given by

$$\hat{\mathbf{e}} = (\mathbf{C}' \mathbf{V}^{-1} \mathbf{C})^{-1} \mathbf{C}' \mathbf{V}^{-1} \mathbf{p}.$$

Judging from numerical experiments, accuracy of estimation drops off rapidly for higher moments. Estimating from 10 genes, with sample sizes on the order of 500, about 4 or 5 is the maximum to give reasonable results. None the less, quite a lot concerning the shape of $g(\theta)$ can be inferred from these moments.

ESTIMATION OF $g(\theta)$

There are several elegant methods for approximating a density function from its moments. See, for example, von Mises (1964), Kendall & Stuart (1958) or Cramér (1946). However, the task of estimating $g(\theta)$ can be greatly simplified by the finite range of θ. To take advantage of this property, we follow the approach of Jackson (1930) in fitting $g(\theta)$ with the best polynomial curve over the range, in the mean square sense. That is, we find the minimum of

$$M = \int_0^1 \left(g(\theta) - \sum_{j=0}^{n} q_j \theta^j\right)^2 d\theta \tag{5}$$

for values $q_j, j = 0, 1, \ldots n$, where we have calculated n moments of $g(\theta)$. The $n+1$ minimization criteria, $\partial M/\partial q_k = 0$, yield equations

$$E(\theta^k) = \sum_{j=0}^{n} q_j/(j+k+1)$$

for $k = 0, \ldots, n$.

If we call $\mathbf{q}$ the $n+1$ dimensional vector of unknown coefficients, $\mathbf{e}$ the vector of observed moments, and $\mathbf{B}$ the $(n+1)$ square matrix

$$\mathbf{B} = \begin{pmatrix} 1 & \frac{1}{2} & \cdots & 1/(n+1) \\ \frac{1}{2} & \frac{1}{3} & \cdots & \\ \vdots & & & \vdots \\ 1/(n+1) & \cdots & & 1/(2n+1) \end{pmatrix},$$

then

$$\mathbf{q} = \mathbf{B}^{-1}\,\mathbf{e}.$$

The function $\sum_{j=0}^{n} q_j \theta^j$ with the values q_j estimated in this way is the least squares estimator of the function $g(\theta)$.

For many cases of $g(\theta)$ the polynomial method is not optimal. It is most appropriate to functions that are near to uniform over the range of θ. If the distribution is concentrated, a transformation of this method using an exponential polynomial model for $g(\theta)$ (see MacLean, 1972) yields better results.

ERROR OF ESTIMATION OF $g(\theta)$

The error in estimation of $g(\theta)$ is rather difficult to calculate. We must go round-about and calculate the error in estimation of θ, which we are not trying to estimate, and thence by an approximate method translate to the error in frequency. Moreover, we must settle for the minimum variance bound derived from the information statistic.

For estimating θ from observations of a locus, L, with alleles $[A_i]$, the information is

$$\begin{aligned} I_\theta(L) &= -E\,\frac{\partial^2 \log p(A_i/\theta)}{\partial\theta^2} \\ &= -\sum \frac{\partial^2 \log (\theta D_i + W_i)}{\partial\theta^2}\,\theta D_i + W_i \\ &= \sum_i \frac{D_i^2}{\theta D_i + W_i}. \end{aligned}$$

This is the information in locus L for the subset of individuals whose true proportion is θ.

Because independently assorting loci are statistically independent within proportion θ, the information from such loci is additive.

$$I_\theta(\text{all } L_j) = \sum_j I_\theta(L).$$

The minimum possible variance of any estimator of θ based upon these loci is the reciprocal of I_θ.

By treating $q(\theta)$ as an ordinary function of θ we can use the standard transformation of error approximation to estimate the variance of $q(\theta)$ in terms of error in θ (Kendall & Stuart, 1958, p. 232). For every value of θ,

$$\text{var } (q(\theta)) \cong \left(\frac{\partial q(\theta)}{\partial \theta}\right)^2 \text{var } (\theta),$$

so that the minimum variance bound for $q(\theta)$ is approximately

$$MVB(q(\theta)) \cong \left(\frac{\partial q(\theta)}{\partial \theta}\right)^2 I_\theta^{-1}. \tag{6}$$

We can use (6) to place confidence limits around $q(\theta)$, along θ from 1 to 1, or we can integrate (6) to get a summary minimum mean square error corresponding to (5).

SUMMARY

In a dihybrid population, the contributions from an ancestral population to the gene pool of an individual may vary between 0 and 100 %. A method is given for estimating the relative frequency of individuals of different ancestry, given data on gene frequencies in the parental populations and the phenotypes of a sample of hybrid individuals for an arbitrary number of polymorphic loci.

REFERENCES

Bernstein, F. (1931). Die geographische Verteilung der Blutgruppen und ihre anthropologische Bedeutung. *Comitato Italiano per lo Studio dei Problemi della Popolazione*, pp. 227–43. Rome: Instituto Poligrafico dello Stato.

Ceppellini, R., Siniscalco, M. & Smith, C. A. B. (1955). Estimation of gene frequencies in a random-mating population. *Annals of Human Genetics* **20**, 97–115.

Cotterman, C. W. (1947). A weighting system for the estimation of gene frequencies from family records. *Contr. Lab. Vert. Biol.*, University of Michigan, no. **33**, 1.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Elston, R. (1971). The estimation of admixture in racial hybrids. *Annals of Human Genetics* **35**, 9–17.

Fisher, R. A. (1940). The estimation of the proportion of recessives from tests carried out on a sample not wholly unrelated. *Ann. Eugen.* **10**, 160.

Glass, B. & Li, C. C. (1953). The dynamics of racial intermixture–an analysis based on the American Negro. *American Journal of Human Genetics* **5**, 1–20.

Jackson, D. (1930). *The Theory of Approximation*. American Mathematical Society, Colloquium Publications, vol. xi.

Kendall, M. G. & Stuart, A. (1958). *The Advanced Theory of Statistics*. Vol. 1. *Distribution Theory*. London: Charles Griffin.

Kendall, M. G. & Stuart, A. (1961). *The Advanced Theory of Statistics*. Vol. 2. *Inference and Relationship*. London: Charles Griffin.

Krieger, H., Morton, N. E., Mi, M. P., Azevedo, E., Freire-Maia, A. & Yasuda, N. (1965). Racial admixture in North-eastern Brazil. *Annals of Human Genetics* **28**, 113–25.

MacLean, C. J. (1973). Estimation and testing of the rat function within the non-stationary Poisson process. *Biometrika* (in the Press).

MacLean, C. J. & Workman, P. L. (1972). Genetic studies of hybrid populations. I. Individual estimates of ancestry and their relation to observations on quantitative traits. *Annals of Human Genetics* **36**, 341.

Nagel, R. & Soto, O. (1964). Haptoglobin types in native Chileans: a hybrid population. *American Journal of Physical Anthropology* **22**, 335–8.

POLLITZER, W. S. (1958). The Negroes of Charleston (S.C.); a study of hemoglobin types, serology and morphology. *American Journal of Physical Anthropology* **16**, 241–63.

REED, T. E. (1969). Caucasian genes in American Negroes. *Science* **165**, 762–8.

RODRIGUEZ, H., DE RODRIGUEZ, E., LORIA, A. & LISKER, R. (1963). Studies on several genetic hematological traits of the Mexican population. V. Distribution of blood group antigens in Nahuas, Yaquis, Tarahumaras, Tarascos and Mixtecos. *Human Biology* **35**, 350–60.

SALDANHA, P. H. & NACRUI, J. (1963). Taste thresholds for phenylthiourea among Chileans. *American Journal of Physical Anthropology* **21**, 113–19.

SALZANO, F. M. (1963). Blood groups and gene flow in Negroes from Southern Brazil. *Acta Genetica* **13**, 9–20.

SALZANO, F. M. & HIRSCHFELD, J. (1965). The dyamics of the Gc polymorphism in a Brazilian population. *Acta Genetica* **15**, 116–25.

SMITH, C. A. B. (1957). Counting methods in genetical statistics. *Annals of Human Genetics* **21**, 254–76.

SMITH, C. A. B. (1967). Notes on gene frequency estimation with multiple alleles. *Annals of Human Genetics* **31**, 99–107.

VON MISES, R. (1964). *Mathematical Theory of Probability and Statistics*. New York: Academic Press.

WORKMAN, P. L. (1968). Gene flow and the search for natural selection in man. *Human Biology* **40** (2), 260–79.

WORKMAN, P. L., BLUMBERG, B. S. & COOPER, A. J. (1963). Selection, gene migration and polymorphic stability in a U.S. White and Negro population. *American Journal of Human Genetics* **15**, 429–37.