Ann. Hum. Genet., Lond. (1973), **36**, 341
Printed in Great Britain

Genetic studies on hybrid populations

I. Individual estimates of ancestry and their relation to quantitative traits

BY CHARLES J. MACLEAN
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

Any two human populations may be assumed to differ genetically, in the relative frequencies of alleles present in both populations, and, possibly, in the types of alleles represented at a locus. For qualitative traits the differences can be demonstrated simply by a comparison of the frequencies of the alleles at one or more loci. However, for heritable quantitative traits, especially those whose environmental component of variation is in part the result of cultural factors, there appears to be no direct method for interpopulation comparisons. The genes controlling variation in such traits cannot be identified, and the effects of differing physical and cultural environments, the complications of social heredity, and genotype–environment interactions, all make the prospect for such comparisons poor. Thus, as discussed by Thoday (1969), there appears to be no direct answer to the question of whether there are genetic differences between races, either in genes or in gene frequencies, with respect to behavioural traits such as that measured by I.Q. tests.

However, in certain cases, migrants from two populations which we might wish to compare have intermixed to form a distinct hybrid population, e.g. the contemporary Chileans formed by intermixture among Spaniards and Auracanian Indians (Nagel & Soto, 1964) and the Negroes in the United States. If immigration from the ancestral populations occurs over a long time, then the contribution from either ancestral population to the gene pool of an individual in the hybrid population can vary between 0 and 100%.

In this paper we shall present a method for relating such variation in ancestry to individual observations on quantitative traits in order to answer a more restricted question: Is the intrapopulation variation in a quantitative trait related to individual differences in ancestral origins? Thus, the method provides an indirect approach to questions about genetic differences between populations.

The technique requires a knowledge of the gene frequencies in the parental populations and the phenotypes of a sample of hybrid individuals. There are two distinct computational problems involved. First, for each individual we must estimate the proportions of the genome derived from either ancestral population. This is accomplished by a Bayesian inversion of conditional gene frequencies. We calculate both a probability distribution and a point estimate of each individual's proportion of ancestry. Secondly, by a regression method, we relate each individual's ancestry to his quantitative score. Finally, limitations on the applicability of this approach are considered in terms of its potential for studies of quantitative variation in American Negro populations.

THE MODEL

Let Q_0 and Q_1 denote random mating populations at equilibrium, and suppose that a distinct hybrid population, H, has been formed over time by intermixture of randomly drawn migrants from each. Each individual in H has some proportion, θ_i, of his ancestry from Q_1, and the remainder, $1-\theta_i$, from Q_0. We must distinguish here between two interpretations of ancestry. One is the historical and is represented by an individual's pedigree. Another is probabilistic and is represented by an individual's genome. Since humans have a large number of chromosomes, the two interpretators are usually in good agreement; i.e. the proportion of an individual's genes from Q_1 is a good indicator of the proportion of his historical ancestry. However, the two are not formally identical. We shall deal entirely with the probabilistic interpretation. Let G represent observations of the phenotypes at some number of polymorphic loci for this individual. We define a conditional density function, $q_i(\theta|G)$, which gives the probability that the individual has proportion θ of his genes derived from Q_1 given that he has the genetic characterization G, i.e. the probability that $\theta_i = \theta$. The function $q_i(\theta|G)$ is defined for θ from 0 to 1.

We assume that the gene frequencies in the ancestral populations are accurately known, and that only intermixture, not selection or mutation, has occurred in H. We define a breeding function $h(\gamma, \delta)$ which represents the relative frequency of offspring from males of proportion γ and females of proportion δ. For example, if we assume the same number of men as women at every θ, together with panmixia, we have

$$h(\gamma, \delta) \propto g(\gamma)\, g(\delta),$$

where $g(\gamma)$ is the relative frequency of individuals with proportion γ in the hybrid population. The estimation of $g(\gamma)$ is described in detail in MacLean & Workman (1972). Although the assumption of panmixia in most hybrid populations would be unrealistic, numerical studies indicate that h plays a relatively minor role in the distribution of phenotypes and, within limits, accuracy in its estimation is not imperative. The most common deviation from panmixia within hybrids is assortative mating with respect to ancestry which yields a positive correlation between mates. If this assortative mating is extreme, it reduces the dispersion of h. If it is considered important, correlation between mates could be incorporated into $h(\gamma, \delta)$ in several elementary ways.

Under the assumptions of ancestral gene frequencies and breeding structure within H, we can calculate $p(G|\theta)$, the probability that an individual has the phenotype characterized by G given that he has proportion of ancestry θ. Having $p(G|\theta)$, we can calculate the posterior probability of proportion ancestry θ, $q_i(\theta|G)$, from an assumed prior. We take the estimated hybrid population density function $g(\theta)$ (see MacLean & Workman, 1973) as the prior probability density at θ. Then $p(G|\theta)$ can be used to convert $g(\theta)$ into the posterior $q_i(\theta|G)$ by Bayes' law:

$$q_i(\theta|G)\, d\theta = \frac{p(G|\theta)\, g(\theta)\, d\theta}{\int_0^1 p(G|\phi)\, g(\phi)\, d\phi}. \tag{1}$$

We shall consider first the derivation of $p(G|\theta)$ given observations only on a single locus, and extend the procedure to observations on several unlinked loci.

THE CONDITIONAL PHENOTYPE FREQUENCY $p(G|\theta)$

Co-dominant loci

Consider a co-dominant, biallelic locus with alleles A and a. The gene frequency in H conditional upon proportion of ancestry from Q_1, $p(A|\theta)$, 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

$$p(A|\theta) = \theta D + W. \tag{2}$$

However, we must infer individual ancestry not from genes but from phenotypes. The phenotypic frequency is a function not only of the individual's proportion of ancestry, but also a function of the difference between the individual's and his parents' proportions. Now, a mother and father with proportions of ancestry from Q_1 denoted by γ and δ will produce offspring of proportion $(\gamma+\delta)/2$. The same proposition, from the offspring's viewpoint, is that an individual of proportion θ_i has parents equidistant and opposite, i.e. with proportions $\theta_i - \phi$ and $\theta_i + \phi$, for some value of ϕ. Assuming for a moment that we know ϕ (but not the parents' genotypes), we can write the phenotypic frequencies conditional upon the gene frequencies at the parental proportions:

$$\left.\begin{aligned} p(AA|\theta_i) &= p(A|\theta_i-\phi)\,p(A|\theta_i+\phi), \\ p(Aa|\theta_i) &= p(A|\theta_i-\phi)\,p(a|\theta_i+\phi)+p(a|\theta_i-\phi)\,p(A|\theta_i+\phi), \\ p(aa|\theta_i) &= p(a|\theta_i-\phi)\,p(a|\theta_i+\phi). \end{aligned}\right\} \tag{3}$$

Substituting the values from (2) for the parental gene frequencies, we have

$$\left.\begin{aligned} p(AA|\theta_i) &= [(\theta_i-\phi)D+W][(\theta_i+\phi)D+W] \\ &= (\theta_i D+W)^2-\phi^2D^2 \\ &= p(A|\theta_i)^2-\phi^2D^2, \\ p(Aa|\theta_i) &= 2(\theta_i D+W)(1-\theta_i D-W)+2\phi^2D^2 \\ &= 2p(A|\theta_i)\,p(a|\theta_i)+2\phi^2D^2, \\ p(aa|\theta_i) &= (1-\theta_i D-W)^2-\phi^2D^2 \\ &= p(a|\theta_i)^2-\phi^2D^2. \end{aligned}\right\} \tag{4}$$

We see that if parents have different proportions of ancestry ($\phi \neq 0$), then the phenotypes of their offspring are not in Hardy–Weinberg ratio. There is a deviation due to the difference in the parental ancestry.

In order to apply (4) to the entire subset of the hybrid population with proportion θ, we integrate over the matings which produce such offspring. For example,

$$\begin{aligned} p(AA|\theta) &= \int_{-R}^{+R} [(\theta D+W)^2-\phi^2D^2]\,h(\theta-\phi,\theta+\phi)\,d\phi \\ &= (\theta D+W)^2 - D^2\int_{-R}^{+R}\phi^2 h(\theta-\phi,\theta+\phi)\,d\phi, \end{aligned} \tag{5}$$

where the breeding function $h(\theta-\phi,\theta+\phi)$ gives the relative frequency of offspring from males of proportion $\theta-\phi$ and females of proportion $\theta-\phi$. The limits of integration are established by the domain of $h(\theta-\phi,\theta+\phi)$, which is defined over 0 to 1 for both arguments, so that

$$R = \min(\theta, 1-\theta).$$

The integral in (5)

$$\int_{-R}^{+R}\phi^2 h(\theta-\phi,\theta+\phi)\,d\phi = \mathrm{MSD}(\theta) \tag{6}$$

is the mean square difference in ancestry between individuals with proportion θ and their parents. It is a measure of the ancestral dispersion of the parents of individuals of ancestry θ, roughly equivalent to a variance. For example, in an ancestrally homogeneous population, MSD $(\theta) = 0$ for all values of θ.

The effect of large variance is a reduction in homozygotes and an increase in heterozygotes, in proportion to the square of the difference in ancestral gene frequencies, D^2.

Complex phenotypes

Because we must attack each locus as a unit in any case, complicated loci are really no more difficult to analyse than biallelic co-dominant ones.

The case of dominance introduces a slight change in calculations from the co-dominant case, but is equally easy to see.

$$\begin{aligned}\text{Prob (phen} = A-|\theta) &= p(AA|\theta)+p(Aa|\theta)\\ &= p(A|\theta)^2+2p(A|\theta)\,p(a|\theta)+D^2\,\text{MSD}\,(\theta).\end{aligned}$$

A further problem arises out of confounded genotypes. For example, the MNSs phenotype might be either MS/NS or Ms/NS. Since the confounded genotypes are exclusive, their probabilities are additive. Therefore,

$$\text{Prob (phen} = \text{MNSs}|\theta) = 2[p(\text{MS}|\theta)\;p(\text{Ns}|\theta)+p(\text{Ms}|\theta)\;p(\text{NS}|\theta)-(D_{\text{MS}}D_{\text{Ns}}+D_{\text{Ms}}D_{\text{NS}})\,\text{MSD}(\theta)].$$

The same problem arises in the Rh locus and has the same simple solution.

Multiple loci

We have derived the phenotype frequency function $p(G|\theta)$ for only one locus at a time. In order to use the information from many loci, we must consider their joint distribution.

Consider the case of two independently assorting loci and their genes A, a and B, b respectively. We take $p(AB|\theta_i)$ as the probability of individual i passing the pair of genes AB to an offspring. We shall express this gene frequency in terms of the gene frequencies from his parents, who have proportions $\theta_i-\phi$ and $\theta_i+\phi$ respectively. On the hypothesis that at each parental proportion, A and B are statistically independent,

$$\begin{aligned}p(AB|\theta_i) &= [4p(AABB|\theta_i)+2p(AABb|\theta_i)+2p(AaBB|\theta_i)+p(AaBb|\theta_i)]/4\\ &= \{4p(A|\theta_i-\phi)\,p(B|\theta_i-\phi)\,p(A|\theta_i+\phi)\,p(B|\theta_i+\phi)+2p(A|\theta_i-\phi)\,p(A|\theta_i+\phi)\\ &\quad\times[p(B|\theta_i-\phi)\,p(b|\theta_i+\phi)+p(b|\theta_i-\phi)\,p(B|\theta_i+\phi)]+2p(B|\theta_i-\phi)\,p(B|\theta_i+\phi)\\ &\quad\times[p(a|\theta_i-\phi)\,p(A|\theta_i+\phi)+p(A|\theta_i+\phi)\,p(a|\theta_i+\phi)]\\ &\quad+[p(a|\theta_i-\phi)\,p(A|\theta_i+\phi)+p(A|\theta_i-\phi)\,p(a|\theta_i+\phi)]\\ &\quad\times[p(B|\theta_i-\phi)\,p(b|\theta_i+\phi)+p(b|\theta_i-\phi)\,p(B|\theta_i+\phi)]\}/4\\ &= \{[p(A|\theta_i-\phi)+p(A|\theta_i+\phi)]/2\}\{[p(B|\theta_i-\phi)+p(B|\theta_i+\phi)]/2\}.\end{aligned}\tag{7}$$

Replacing the parental frequencies with their values from (2) above, (7) becomes

$$p(AB|\theta_i) = (\theta_i D_A+W_A)\,(\theta_i D_B+W_B).\tag{8}$$

Equation (7) clearly holds for the first hybrid generation where the parents were from the ancestral populations, Q_0 and Q_1, because in both these populations, which we assumed to be in equilibrium, A and B are statistically independent. Therefore from (8), A and B are also statistically independent within any individual in generation 1. The argument can be extended

to the second generation based upon the first, and thence to all subsequent generations. Therefore, within any individual or any subset with identical proportion θ, we find that independently assorting genes are also statistically independent. The same applies to phenotypes. For example,

$$\begin{aligned} p(AABB|\theta_i) &= p(A|\theta_i-\phi)\,p(B|\theta_i-\phi)\,p(A|\theta_i+\phi)\,p(B|\theta_i+\phi) \\ &= p(AA|\theta_i)\,p(BB|\theta_i). \end{aligned}$$

Therefore the phenotypic frequencies from unlinked loci are also statistically independent within the subgroup of individuals with proportion θ. In general then, for n unlinked loci,

$$p(G|\theta) = p(\text{all } G_j|\theta) = \prod_j^n p(G_j|\theta). \tag{9}$$

Calculation of the posterior, $q_i(\theta|G)$

Given the prior $g(\theta)$ and the conditional phenotype frequency $p(G_j|\theta)$ for n unlinked loci together with their conditional independence under θ, we can perform Bayes' inversion (1) as follows:

$$q_i(\theta|\text{ all } G_j)\,d\theta = \frac{g(\theta)\prod_{j=1}^n p(G_j|\theta)\,d\theta}{\int_0^1 g(\phi)\prod_{j=1}^n p(G_j|\phi)\,d\phi}. \tag{10}$$

Clearly, calculation of (10) would be tractable analytically only under very simple assumptions. The numerical computations are easily within the power of computing machinery however. For each individual, equation (10) is evaluated over a set of small increments of θ from $\theta = 0$ to $\theta = 1$, and the function q_i is recorded as a table of these values. Notice that by using the restriction that $\int q_i\,d\theta = 1$, the denominator of (10), which is constant with respect to θ, need not be calculated. Rather the values of q_i are normalized at the end by dividing each by their sum.

The first step in the calculation of the numerator of (10) is the evaluation of $g(\theta)$ for each incremental value of θ. Next an individual's set of phenotypes $[G_j]$ is recorded. Then for each value of θ and for each of the individual's phenotypes, G_j, the value $p(G_j|\theta)$ is calculated from the appropriate equation, e.g. (5). Finally the product

$$g(\theta)\prod_{j=1}^n p(G_j|\theta)$$

is calculated for each value of θ, and recorded in $q_i(\theta)$.

POINT ESTIMATION; x_i AND ITS MEAN SQUARE ERROR

Each individual has some true ancestry, θ_i. The conditional probability function, $q_i(\theta)$ (which we shall from now on write with the condition G implicit), is only a statement about our knowledge, and its variance is a measure of our ignorance. For the argument in the following section it is necessary to summarize our estimate of an individual's ancestry with one point, x_i. Two natural estimators for θ_i are the posterior maximum, i.e. the mode of $q_i(\theta)$, and the posterior average, i.e. the mean of $q_i(\theta)$. The major drawback of the posterior average estimator is its bias. Because the posterior retains the prior in it (see formula (10)) the mean of the posterior is biased toward the mean of the prior. The mode of $q_i(\theta)$, the posterior maximum estimator, is not biased in this way. It is, however, much more subject to random variation. Computer simulation of various theoretical populations indicates that its mean square error is about 10 % larger than that of the posterior average.

Mean square error of x_i

We present two alternative methods of error calculation. One method is very easy to apply, but it relies upon the same assumptions and formulation as the estimate itself and might therefore be subject to the same bias. Another method, requiring additional data, is based upon minimum assumptions. Both calculations apply to any kind of point estimator of θ_i.

The simple estimate of mean square error of estimate is derived directly from the posterior function $q_i(\theta)$. Consider the infinite subset of all possible individuals with total phenotype G_i. If this subset were drawn at random from $g(\theta)$, and if the genetic assumptions were to hold, then the true values, $[\theta_j]$, of this subset would be distributed according to $q_i(\theta)$. Since we estimate all of these true values θ_j by the same point value, x_i, then the mean square of the resulting errors is

$$\int_0^1 (\theta - x_i)^2 q_i(\theta)\, d\theta = \text{MSE}\,(x_i). \tag{11}$$

MSE (x_i) is easily calculated from the numerical form of $q_i(\theta)$. It is the error variance only in the case of an unbiased estimator x_i. For any biased estimator the error variance is less than (11) by the square of the bias. The minimum of (11) occurs for x_i equal to the posterior average.

For the population value,

$$\begin{aligned}\text{MSE} &= \sum_i \int (\theta - x_i)^2 q_i(\theta)\, d\theta\, p(G_i) \\ &= E(\theta^2) - 2E(y, x) + E(x^2),\end{aligned}$$

where y_i is the average of the posterior $q_i(\theta)$. In the case where x_i is also the posterior average

$$\text{var}\,(x) = \text{var}\,(\theta) - \text{MSE}. \tag{12}$$

Var (θ) can be estimated directly by another method (see MacLean & Workman, 1973).

Since the error variance plays an important role in relating individual scores to the ancestry, it is well to have an independent estimate of the error. Suppose we have a subsample of sibling pairs. The members of each sib pair naturally have the same true proportion of ancestry. The estimates from a pair of sibs of true ancestry θ are

$$\left.\begin{aligned} x_1 &= \theta + \epsilon_1 \\ x_2 &= \theta + \epsilon_2. \end{aligned}\right\} \tag{13}$$

Assuming only that the error is distributed the same for both sibs, we have

$$\text{var}_\theta(\epsilon) = \frac{\text{intrapair var (sibs)}}{1 - \text{corr (sibs)}}. \tag{14}$$

The expected value over θ of this conditional variance is the desired var (ϵ). Since x is derived entirely from genetic factors, the sibling correlation is 0·5 except for dominance deviation due to the specific loci used.

If parents and children are available their values are used as follows:

$$\begin{aligned} x &= \theta + \epsilon_1, \\ y &= \tfrac{1}{2}(x_M + x_F) = \theta + \tfrac{1}{2}(\epsilon_2 + \epsilon_3). \end{aligned} \tag{15}$$

In order that $E(x - y) = 0$ we must assume that the bias in error at θ equals the average of the

biases at the parental values, equidistant from θ. Although this is not precisely true in all cases, it is usually very close. If in addition the parents are uncorrelated

$$\text{var}\,(x-y) = \text{var}_{\theta}(\epsilon) + [\text{var}_{\theta-\phi}(\epsilon) + \text{var}_{\theta+\phi}(\epsilon)]/4 - \text{cov}\,(x, x_M + x_F)$$

and taking expected values first over ϕ and then over θ,

$$\text{var}\,(\epsilon) = \frac{2\,\text{var}\,(x-y)}{3\text{–}4\,\text{corr}\,(\text{parent, child})}. \tag{16}$$

THE RELATION BETWEEN OBSERVATIONS ON INDIVIDUALS AND THEIR ESTIMATED PROPORTION OF ANCESTRY

The approach taken in the foregoing sections concentrates entirely upon the individual rather than the population. Ancestral characterization of a population as a whole is better accomplished by another method (see MacLean & Workman, 1973). This emphasis upon the individual is motivated by interest in the relation between ancestry and heritable quantitative traits. In this section we consider how to relate the individual variation in the proportion of ancestry, θ_i, to the individual quantitative score, s_i. The relationship is a function of θ from 0 to 1, such as the expected value $E(s_i \text{ for all } \theta_i = \theta)$. The general shape of the function can be estimated by the weighted scores

$$f(\theta) = \Sigma s_i q_i(\theta)/\Sigma q_i(\theta).$$

This function weights each score in proportion to the probability that it is appropriate. Examination of $f(\theta)$ would indicate whether there were heterotic effects of intermixture, or whether only additive effects of ancestral differences were present.

We shall confine our analysis to the latter case, in which f is a straight line

$$f(\theta) = a + b\theta.$$

More precisely, we hypothesize a linear relationship between each individual's score and his ancestry,

$$s_i = a + b\theta_i + \delta_i, \tag{17}$$

where the error of regression, δ, is distributed independently of s. It is a and b from the unknown (17) that we wish to estimate. The regression equations, when θ is known, are quite familiar (Kendall & Stuart, 1961).

$$b = \frac{\text{cov}\,(s, \theta)}{\text{var}\,(\theta)} \tag{18}$$

and

$$a = E(s) - E(\theta)\,b, \tag{19}$$

where all expected values are from the distribution of ancestry within H, $g(\theta)$.

In our case we do not know θ_i, but rather have an estimate

$$x_i = \theta_i + \epsilon_i, \tag{20}$$

so that

$$s_i = a + b(x_i - \epsilon_i) + \delta_i. \tag{21}$$

The usual procedure is to replace the parameter with its estimator so that

$$b' = \frac{\text{cov}\,(s, x)}{\text{var}\,(x)}. \tag{22}$$

But this procedure in our case incorporates bias into the regression, since we are faced not only with error in both variables, but in addition with the fact that for the independent variable, x, the error of estimate, ϵ, is correlated with the true parameter, θ.

Regardless of the estimator we use, its error distribution will be a function of the true proportion θ. This is inherent in the finite range of θ and also in the Bayesian formulation. It affects the regression equation in both numerator and denominator. Cov (s, x) is a function of the information content of the estimator; in general, the worse the estimate the lower the covariance. Var (x) depends upon the estimation method. The posterior maximum estimator, for example, yields var (x) about equal to or slightly larger than var (θ) except for very low information. Therefore b' calculated from the posterior maximum estimate is too small. The posterior average on the other hand always yields var (x) smaller than var (θ). Therefore b' calculated from the posterior average is too large.

To avoid the bias in (22) we must retain formula (18). In order to use (18) we need var (θ), which is calculated from another method (see MacLean & Workman, 1973), and we must derive cov (s, θ) in terms of quantities we can measure. By using (20) we see that

$$\begin{aligned} \operatorname{cov}(s,\theta) &= \operatorname{cov}(s, x-\epsilon) \\ &= \operatorname{cov}(s,x) - \operatorname{cov}(s,\epsilon). \end{aligned} \tag{23}$$

And by (17)

$$\begin{aligned} \operatorname{cov}(s,\epsilon) &= \operatorname{cov}[a+b\theta+\delta, \epsilon] \\ &= b \operatorname{cov}(\theta,\epsilon). \end{aligned} \tag{24}$$

Substituting (24) into (23), then (23) into (18),

$$b = \frac{\operatorname{cov}(s,x)}{\operatorname{var}(\theta)+\operatorname{cov}(\theta,\epsilon)}. \tag{25}$$

We see by calculating the variance of both sides of (20) that

$$\operatorname{cov}(\theta,\epsilon) = [\operatorname{var}(x) - \operatorname{var}(\theta) - \operatorname{var}(\epsilon)]/2. \tag{26}$$

Therefore, finally, substituting (26) into (25) we have

$$b = \frac{2\operatorname{cov}(s,x)}{\operatorname{var}(\theta)+\operatorname{var}(x)-\operatorname{var}(\epsilon)}. \tag{27}$$

Formula (19) remains unaltered,

$$a = E(s) - E(\theta)\, b.$$

We measure var (x), cov (s, x) and $E(s)$ as usual from our sample. We have an independent method of estimating var (θ) and $E(\theta)$ (MacLean & Workman, 1973), and estimation of var (ϵ) is discussed in the previous section.

The errors of estimation are quite analogous to the standard case. The residuals have variance

$$\begin{aligned} \operatorname{var}(\text{res}) &= \operatorname{var}(s) - b \operatorname{cov}(s,\theta) \\ &= \operatorname{var}(s) - b \operatorname{cov}(s,x) + \tfrac{1}{2}b^2[\operatorname{var}(x) - \operatorname{var}(\theta) - \operatorname{var}(\epsilon)]. \end{aligned}$$

The coefficients have error variances

$$\operatorname{var}(b) = \operatorname{var}(\text{res})/[n \operatorname{var}(\theta)],$$

$$\operatorname{var}(a) = E(\theta^2) \operatorname{var}(b)$$

and

$$\operatorname{cov}(a,b) = -E(\theta) \operatorname{var}(b).$$

We therefore have a method of regression when the independent variable is not known, but rather is estimated with bias.

DISCUSSION

If complete genealogies were available we could ascertain the true historical proportion of ancestry of an individual in a dihybrid population, but the chance of ever obtaining such data is small. However, given a genetic characterization of hybrid individuals and the gene frequencies in the ancestral population, the present method provides an estimate of the probability distribution, $q_i(\theta)$, of each individual's genetical proportion of ancestry. From $q_i(\theta)$ we can obtain a point estimate of an individual's ancestry and then relate that value to individual quantitative scores. Whether or not the method can be meaningfully applied to any real dihybrid population depends on the extent of the genetic differences between the parental populations, the reliability of the data to be used, and certain assumptions about the hybrid population structure. The problems inherent in the use of this approach will be discussed briefly in terms of a possible application in the study of intrapopulation variation in United States Negro populations.

The African origins of the American Negroes are known only in terms of the general areas from which the slaves were obtained and the Caucasian contributors to the Negro gene pool can only be described as being primarily western European. Thus despite the advantage of considerable genetic distance between the parental populations, they are far from the idealized panmictic populations of the model. In addition, the African gene frequencies are not generally based on extensive sampling of the appropriate geographical areas (see Reed, 1969). In order to diminish the effect of such systematic errors in obtaining the African data there must be careful selection of the genes used to calculate $q_i(\theta)$. Genes whose African frequencies show considerable heterogeneity or which are derived from a very small number of observations should not be used.

We must also assume that contemporary gene frequencies reflect those in the ancestral populations, i.e. that they have not been affected by selection or drift. For Caucasian estimates, since U.S. White and European frequencies show great similarity, such an assumption seems warranted (Workman, Blumberg & Cooper, 1963). However, previous studies have demonstrated that some genes, such as Hb^s and Hp^1, yield highly discordant estimates of the mean amount of the Caucasian contribution to U.S. Negro gene pools (Workman *et al.* 1963; Workman, 1968). Genes which yield discordant estimates of the mean amount of intermixture, whether because of selection or sampling error, can be located directly by the use of Bernstein's (1931) formula and omitted from the computation of $q_i(\theta)$.

Finally, genes with very similar frequencies in Africans and Caucasians provide no information about the ancestral origins and these too should not be used to determine $q_i(\theta)$.

After the elimination of inappropriate polymorphic markers there remain only a small number of loci which can be considered acceptable for estimating the distribution of the proportion of ancestry in an American Negro population. These include the Rhesus (Rh) and Duffy (Fy) loci, both of which possess alleles or antigenic factors which are in very high frequency in one parental population and very low frequency or absent in the other. Such factors provide the greatest discriminatory power in the determination of ancestry and are likely to involve the least error (see Reed, 1969). Loci such as ABO, Kidd (Jk), Lewis, S, Gm, Inv, and Gc can also be used. By themselves these loci may not provide distributions of $q_i(\theta)$ or point estimates with a satisfactorily low variance. However, in Charleston Negroes, estimates of the biological distances among parental and hybrid populations based on gene frequencies were quite similar to those based on anthropometric characters (Pollitzer, 1958). Thus, despite the difficulties in obtaining reliable

gene frequencies in the African populations, the incorporation of anthropometric data into the calculation might provide estimates of $q_i(\theta)$ which have sufficiently small error variance. Of course, as knowledge of the African gene frequencies increases we should obtain increasingly reliable estimates of the ancestry.

The derivation of the proportion of ancestry also requires a function describing the breeding structure of the population. Although the model incorporates a general breeding function, h, in the absence of other information, we assume that both mating and distribution of family sizes be independent of θ. Although this is probably not the case in real populations, computer simulations, to be presented elsewhere, show that moderate deviations from these conditions have little effect on the results. Regular deviations from panmixia could be incorporated into the model but there is no evidence suggesting any particular alternative form of h.

In order to relate variation in ancestry to variation in quantitative scores, the hybrid population should be one in which there is considerable variation in the proportion of ancestry. Negro populations in the southern United States with relatively low proportions of Caucasian genes and a recent history of social isolation are probably not suitable. On the other hand, northern urban Negro populations which have higher proportions of Caucasian genes, approximately 20–30 % (Glass & Li, 1953; Workman, 1968; Reed, 1969), may well have individuals with proportions of Negro ancestry varying between 5 and 60 %. In order to maximize the within-population heritability of the traits whose variation is under study one should also choose a hybrid population with minimal variation in both physical and cultural environmental factors, e.g. unrelated children from the same school or neighbourhood. In addition, in the regression of scores on ancestry one should hold constant socio-economic factors and any anthropometric character such as skin colour which might, for cultural reasons, interact with performance on the trait. That is, traits giving a visual indication of ancestral origins cannot be used to estimate $q_i(\theta)$ under such circumstances. Hence, if adequate parental gene frequencies are not available, the method is probably least suitable for studying behavioural variables such as I.Q. scores.

Finally, suppose that reliable estimates of the ancestry have been obtained in a study on a random sample from a Negro population and that some quantitative scores, such as physiological variables related to hypertension, were regressed on point estimates of ancestry. The linear regression model assumes that if there are genetic differences between the parental populations, they are additive. We must also assume the absence of specific genotype–environment interactions and that the environmental factors have been randomized with respect to the true value of the proportion of ancestry. Under these assumptions, the slope of the regression line should indicate the extent to which, in the environment of the Negro population, differences in scores are related to additive mean differences in the genetic contributions from the parental populations. Further, the extrapolation of the regression line to $\theta = 0$, i.e. pure Caucasian, would indicate the expected performance of Caucasians in that environment.

In no way can this approach prove anything about genetic differences between the parental populations. However, direct comparisons of Caucasian and Negro traits, whether behavioural, physiological or anatomical, do so without correcting for any differences in the physical or cultural environments of the groups being compared (e.g. Jensen, 1969). The present approach is an attempt to correct, in so far as is possible, for such environmental differences and, notwithstanding the complications in obtaining the genetic data and the assumptions about the model and the regression analysis, it should provide a much more useful basis for interpopulation comparisons.

SUMMARY

An individual i in a hybrid population H descended from two ancestral populations Q_0, Q_1 will have a certain proportion θ_i of his genes descended from Q_1. Methods are presented for estimating the value of θ_i for a given individual on the basis of his (or her) phenotype. Both the posterior distribution of θ_i and point estimates are obtained. Sources of error or inaccuracy are considered.

REFERENCES

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

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.

JENSEN, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review* **39** (1), 1.

MACLEAN, C. J. & WORKMAN, P. L. (1973). Genetic studies of hybrid populations. II. The distribution of ancestry. *Annals of Human Genetics* (in the press.)

NAGEL, R. & SOTO, O. (1964). Haptoglobin types in native Chilians: 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 type, serology and morphology. *American Journal of Physical Anthropology* **16**, 241–63.

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

THODAY, J. M. (1969). Limitations to genetic comparison of populations. *Journal of Biosocial Science* Suppl. 1, p. 3.

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.