The Genetic Structure of Admixed Populations

Jeffrey C. Long

Department of Anthropology, University of New Mexico, Albuquerque, New Mexico 87131

Manuscript received June 5, 1990
Accepted for publication October 15, 1990

ABSTRACT

A method for simultaneously estimating the admixture proportions of a hybrid population and Wright's fixation index, F_{ST}, for that hybrid is presented. It is shown that the variance of admixture estimates can be partitioned into two components: (1) due to sample size, and (2) due to evolutionary variance (*i.e.*, genetic drift). A chi-square test used to detect heterogeneity of admixture estimates from different alleles, or loci, can now be corrected for both sources of random errors. Hence, its value for the detection of natural selection from heterogeneous admixture estimates is improved. The estimation and testing procedures described above are independent of the dynamics of the admixture process. However, when the admixture dynamics can be specified, F_{ST} can be predicted from genetic principles. Two admixture models are considered here, gene flow and intermixture. These models are of value because they lead to very different predictions regarding the accumulation of genes from the parental populations and the accumulation of variance due to genetic drift. When there is not evidence for natural selection, and it is appropriate to apply these models to data, the variance effective size (N_e) of the hybrid population can be estimated. Applications are made to three human populations: two of these are Afro-American populations and one is a Yanomamö Indian village. Natural selection could not be detected using the chi-square test in any of these populations. However, estimates of effective population sizes do lead to a richer description of the genetic structure of these populations.

ESTIMATION of the proportional contributions of ancestral populations to their hybrids has been important in genetic analyses of many admixed human populations (*cf.* CHAKRABORTY 1986). Such admixture estimates help to clarify the historical background of admixture and they are becoming useful in genetic epidemiological investigations (CHAKRABORTY and WEISS 1986, 1988). Admixed populations have allele frequencies that are linear combinations of the allele frequencies in their parental populations, and since admixture affects all loci equally, the same set of admixture proportions are expected to apply to all alleles at all loci. However, estimated ancestral contributions vary from allele to allele and from locus to locus for a number of reasons; sampling error in the estimation of parental and hybrid population allele frequencies and genetic drift are prominent sources of random error, and systematic biases are potentially introduced by natural selection for or against some alleles.

It has been argued in the past that natural selection can be detected if admixture estimates from different alleles are highly heterogeneous, or if some small group of alleles yield admixture estimates that deviate in the extreme from admixture estimates at most other alleles (WORKMAN, BLUMBERG and COOPER 1963; WORKMAN 1968; REED 1969; CAVALLI-SFORZA and BODMER 1971). A biased admixture estimate should reflect the cumulative effect of natural selection over several generations, and application of this approach was thought to be more powerful than single generational studies (REED 1969). Unfortunately, the results of heterogeneity searches have been equivocal because of the variation in admixture estimates from other causes (ADAMS and WARD 1973). From an evolutionary perspective, an important source of variation in admixture estimates, other than that caused by natural selection, is caused by genetic drift.

Genetic drift, like admixture, affects all loci simultaneously. Its impact is typically evaluated from the parameter F_{ST} (WRIGHT 1965, 1969). It is shown in this paper that F_{ST} can be used to measure the effect of genetic drift on admixture estimates. It is also demonstrated that F_{ST} is informative about other aspects of the hybrid population's evolutionary structure. For example, F_{ST} is closely related to the variance effective size of the hybrid population.

A method to estimate simultaneously the ancestral contributions to a hybrid population and F_{ST} is presented. This method allows a partition of the variance of the estimated admixture proportions into two components; the first component measures the error inherent in statistical sampling of existing populations, and the second component measures the error that accumulates because of evolutionary change. In the absence of natural selection, this component is a measure of genetic drift. Hereinafter, the first source of error is referred to as *sampling error* and the second

source of error is referred to as *evolutionary error*. The estimator for F_{ST} presented here is distinct from other estimators of F_{ST} (*cf.* COCKERHAM 1969, 1973; WEIR and COCKERHAM 1984; LONG 1986). It is obtained from the variance of allele frequencies with respect to their expectations based on the admixture model. In order to obtain this estimate of F_{ST}, multiple locus samples are required from the hybrid population and from all contributing parental populations. F_{ST} cannot be computed for the hybrid population in the absence of information on the parental populations.

A specific model of admixture dynamics is not required by the basic procedures presented here, but if the dynamics of admixture can be specified, then F_{ST} is predictable from population genetic principles and it is possible to estimate the variance effective size (N_e) of the hybrid population. Two such models of admixture dynamics and methods for estimating N_e are explored in later sections of this paper. One of the models evaluated, gene flow, although ubiquitous in the admixture literature (*cf.* GLASS and LI 1953), is found to be unusual with respect to genetic drift because there will *not* be a single value of F_{ST} common to all alleles at all loci. When admixture takes the form of gene flow, the procedures developed here estimate an average F_{ST} pertaining to all loci sampled.

The variance formula for admixture estimates derived here is used to improve a chi-square statistic that has been designed to detect natural selection from heterogeneity among admixture estimates obtained from different alleles (CAVALLI-SFORZA and BODMER 1971). There are two major problems with the chi-square test (in its original form) that have rendered its results controversial. First, the test confounds heterogeneity due to natural selection and genetic drift (CAVALLI-SFORZA and BODMER 1971; ADAMS and WARD 1973). Second, the method treats different alleles that segregate within the same locus as if they are statistically independent. Thus, notions of random sampling are violated. The methods derived here rectify these problems and the interpretability of a significant chi-square is clarified.

The new methods and approaches for admixture analysis described above are applied to three admixed human populations in the later sections of this paper. The populations analyzed are of historical interest and they are useful for demonstrating the distinctions between the two theoretical models of admixture examined here.

A BASIC ADMIXTURE/DRIFT MODEL AND ESTIMATION PROCEDURES

Suppose that one codominant allele from each of several unlinked loci is chosen for analysis (multiple alleles and dominance are dealt with in a later section of this paper). Let A_i designate the chosen allele from the ith locus ($i = 1, \cdots, I$), and P_{hi}, P_{1i}, and P_{2i} be the values of its frequencies in the hybrid population, the first parental population, and the second parental population, respectively. If admixture and genetic drift are the only evolutionary process that have affected the hybrid gene pool, then

$$\begin{aligned} P_{hi} &= \mu \cdot P_{1i} + (1 - \mu) \cdot P_{2i} + \epsilon_{ei} \\ &= P_{2i} + \mu \cdot (P_{1i} - P_{2i}) + \epsilon_{ei} \end{aligned} \tag{1}$$

where μ is the proportionate contribution of the 1st parental population, $(1 - \mu)$ is the proportionate contribution of the 2nd parental population, and ϵ_{ei} is the error due to genetic drift. The expectation of P_{hi} is $\mu \cdot P_{1i} + (1 - \mu) \cdot P_{2i}$; for convenience, this expectation will be denoted $E[P_{hi}] = \pi_i$. The expectation of ϵ_{ei} is zero, but the expectation of its square, and hence its variance, is not. The variance of the hybrid allele frequency caused by genetic drift is $\sigma_e^2(p_{hi}) = E[P_{hi} - \pi_i]^2$.

If P_{1i} and P_{2i} are known parameters, the method of weighted least squares (WLS) can be used to estimate μ from a sample of hybrid population genotypes (*cf.* ELSTON 1971; LONG and SMOUSE 1983). For these purposes, let $\tilde{P}_{hi}$ be the maximum likelihood *estimate* of the allele frequency in the hybrid population based on a sample of N_s individuals. The expectation of $\tilde{P}_{hi}$ is π_i, however $\tilde{P}_{hi}$ contains a component of random error, ϵ_{si}, due to statistical sampling. The expectation of ϵ_{si} is zero, assuming that the population is large and random mating, and its variance, $\sigma_s^2(p_{hi})$, is $E[\tilde{P}_{hi} - P_{hi}]^2 = P_{hi} \cdot (1 - P_{hi})/2N_s$. The two error terms, ϵ_{si} and ϵ_{ei}, are assumed to be independent, and accordingly, the variance of $\tilde{P}_{hi}$, with respect to both genetic drift and statistical sampling, is $\sigma_s^2(p_{hi}) + \sigma_e^2(p_{hi})$. These quantities are illustrated in Figure 1. Table 1 provides a list of definitions and symbols that will be used throughout this paper.

To obtain the WLS estimate of μ, Equation 1 is rewritten as

$$(\tilde{P}_{hi} - P_{2i}) = \mu \cdot (P_{1i} - P_{2i}) + \epsilon_i \tag{2}$$

where the term ϵ_i is the sum of the two random components, *i.e.*, $\epsilon_i = \epsilon_{ei} + \epsilon_{si}$. The admixture parameter, μ, is then estimated by defining a vector $\mathbf{X} = [(P_{11} - P_{21}), (P_{12} - P_{22}), \cdots, (P_{1I} - P_{2I})]^T$, a diagonal matrix $\mathbf{V}$ with elements $\{V_{ii}\} = E[P_{hi}](1 - E[P_{hi}])$, and a vector $\mathbf{y} = [(\tilde{P}_{h1} - P_{21}), (\tilde{P}_{h2} - P_{22}), \cdots, (\tilde{P}_{hI} - P_{2I})]^T$. The estimate of μ is given by

$$M = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}. \tag{3}$$

The solution to Equation 3 is obtained iteratively (LONG and SMOUSE 1983). At each iteration, the expected hybrid allele frequencies, $E[P_{hi}] = \pi_i$, are approximated by $\tilde{\pi}_i = P_{2i} + M \cdot (P_{1i} - P_{2i})$. Following the usual steps (*cf.* NETER and WASSERMAN 1974), an estimate of the variance of the admixture estimate is

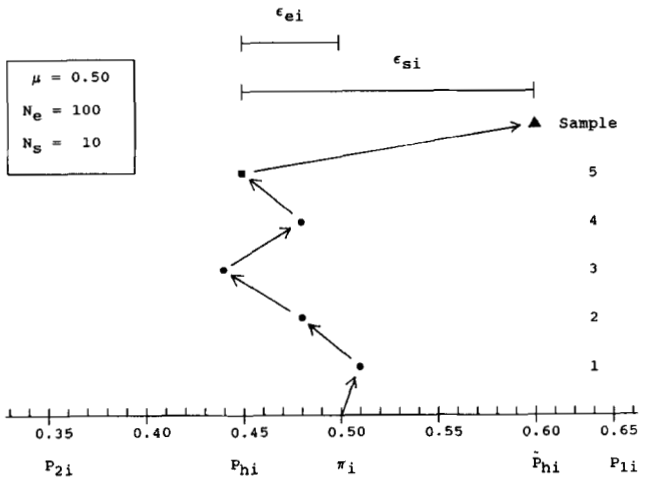

FIGURE 1.—The quantities of interest to this paper are illustrated using a simple simulation. An admixed population (1 locus) was formed by simple Intermixture between two parental populations with parameters, μ, P_{1i} and P_{2i}, as shown above. Following population formation, the admixed population was assigned an effective population size $N_e = 100$, and allowed to drift for five generations. Following the five generations of drift, a sample of $N_s = 10$ individuals was drawn. The allele frequencies for the first four generations of drift are indicated by (●) and the allele frequency, P_{hi}, at the fifth generation is denoted (■). The estimated allele frequency, $\tilde{P}_{hi}$, is marked with (▲). Note that $\tilde{P}_{hi} = \pi_i + \epsilon_{ei} + \epsilon_{si}$. The Intermixture drift model is explained in greater detail in a later section of this paper.

TABLE 1

Summary of symbols and notation used

P_{hi}	=	frequency of A_i in the hybrid population
P_{1i}	=	frequency of A_i in the 1st parental population
P_{2i}	=	frequency of A_i in the 2nd parental population
μ	=	contribution of the 1st parental population to the hybrid
π_i	=	$E[P_{hi}] = \mu \cdot P_{1i} + (1 - \mu) \cdot P_{2i}$.
ϵ_{ei}	=	drift deviation of the ith allele, *i.e.*, $P_{hi} - \pi_i$
$\sigma_e^2(p_{hi})$	=	$E[P_{hi} - \pi_i]^2$ = drift variance of the ith allele
$\tilde{P}_{hi}$	=	maximum likelihood estimate of P_{hi}
N_s	=	sample size drawn from the hybrid population
ϵ_{si}	=	sampling deviation of the ith allele, *i.e.*, $\tilde{P}_{hi} - P_{hi}$
$\sigma_s^2(p_{hi})$	=	$E[\tilde{P}_{hi} - P_{hi}]^2$ = sampling variance of the ith allele
ϵ_i	=	$\epsilon_{ei} + \epsilon_{si}$
M	=	weighted least square (WLS) estimate of μ
$\tilde{\pi}_i$	=	$P_{2i} + M \cdot (P_{1i} - P_{2i})$ approximation of π_i
$s^2(M)$	=	expected variance of the admixture estimate
MSE	=	mean squared error of the admixture model
F_{ST}	=	$\sigma_e^2(p_{hi})/[\pi_i(1 - \pi_i)]$
F_{ST}^*	=	estimate of F_{ST}
$s_s^2(M)$	=	component of $s^2(M)$ due to statistical sampling
$s_e^2(M)$	=	component of $s^2(M)$ due to genetic drift
N_e	=	variance effective size of the hybrid population
N_e^*	=	estimate of N_e
α	=	generational contribution of the donor population to the recipient population (gene flow model)
$\hat{\alpha}$	=	estimate of α

given by

$$s^2(M) = \text{MSE} \cdot (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{4}$$

where the mean squared error (MSE) of the admixture model is given by

$$\text{MSE} = (\mathbf{y} - \mathbf{X} \cdot M)^T \cdot \mathbf{V}^{-1} \cdot (\mathbf{y} - \mathbf{X} \cdot M)/(I - 1). \tag{5}$$

The mean squared error (MSE) estimates $E[(\tilde{P}_{hi} - \pi_i)^2\pi_i \cdot (1 - \pi_i)]$, which is the standardized variance of estimated hybrid population allele frequencies. This quantity is very informative about the breeding structure of the hybrid population because it is closely related to WRIGHT's (1965, 1969) fixation index, F_{ST}. WRIGHT (1965, p. 402; 1969, p. 295, and elsewhere) defines the quantity $F_{ST} = \sigma_{q(ST)}^2/[q_T(1 - q_T)]$, as the "ratio of the actual allele frequency variance of the subdivisions to its maximum possible value $q_T(1 - q_T)$ that is expected if the subdivisions are completely isolated and each completely fixed" (q denotes the allele frequency). Wright used the variance of subdivisions, $\sigma_{q(ST)}^2$, to pertain to: (1) the aggregate of an essentially infinite number of existing subdivisions (*e.g.*, WRIGHT 1969, p. 294), or equivalently, (2) an infinitely large number of isolated lines which might, hypothetically, have been drawn from a founding stock (WRIGHT 1965, p. 407; 1951, p. 328). The latter interpretation is necessary for analysis of an isolated population, or line, such as the hybrid population of the model presented here. The parameters of $\sigma_e^2(p_{hi})$ and π_i defined in this paper correspond to WRIGHT's parameters $\sigma_{q(ST)}^2$ and q_T, respectively.

The relationship between F_{ST} and the MSE of the admixture model is demonstrated by considering the expected variance of $\tilde{P}_{hi}$ about π_i

$$\begin{aligned} E[\tilde{P}_{hi} - \pi_i]^2 &= \sigma_s^2(p_{hi}) + \sigma_e^2(p_{hi}) \\ &= E[\tilde{P}_{hi} - P_{hi}]^2 + E[P_{hi} - \pi_i]^2 \\ &= \frac{E[P_{hi} \cdot (1 - P_{hi})]}{2 \cdot N_s} + \sigma_e^2(p_{hi}) \\ &= \frac{1}{2N_s}[E[P_{hi}] - E[P_{hi}^2]] + \sigma_e^2(p_{hi}) \\ &= \frac{1}{2N_s}[E[P_{hi}] - E[P_{hi}]^2 - \sigma_e^2(p_{hi})] \\ &\quad + \sigma_e^2(p_{hi}). \end{aligned} \tag{6}$$

After simplification, and using the notation $E[P_{hi}] = \pi_i$,

$$E[\tilde{P}_{hi} - \pi_i]^2 = \frac{\pi_i \cdot (1 - \pi_i)}{2N_s} + \sigma_e^2(p_{hi})\left[1 - \frac{1}{2N_s}\right]. \tag{7}$$

Upon standardization by $\pi_i \cdot (1 - \pi_i)$,

$$\frac{E[\tilde{P}_{hi} - \pi_i]^2}{\pi_i \cdot (1 - \pi_i)} = \frac{1}{2N_s} + F_{ST} \cdot \left[1 - \frac{1}{2N_s}\right] \tag{8}$$

(recall that $F_{ST} = \sigma_e^2(p_{hi})/[\pi_i(1 - \pi_i)]$). Thus,

$$E[\text{MSE}] \approx \frac{1}{2N_s} + F_{ST} \cdot \left[1 - \frac{1}{2N_s}\right] \tag{9}$$

and

$$F_{ST}^{*} = 2N_s/(2N_s - 1)[MSE - 1/(2N_s)] \quad (10)$$

can be used to estimate F_{ST}. Since the terms in MSE are standardized by the estimate, $\tilde{\pi} \cdot (1 - \tilde{\pi})$, rather than the parameter $\pi \cdot (1 - \pi)$, Equations 9 and 10 are approximations. Nevertheless they are nearly unbiased estimators when F_{ST} is reasonably low (say $F_{ST} < 0.10$) or, many loci have been assayed (J. SOBUS and J. C. LONG, unpublished computer simulations).

With these points in mind, the estimated variance of the admixture estimate, $s^2(M)$, can be partitioned into two additive components

$$s^2(M) = s_s^2(M) + s_e^2(M) \quad (11a)$$

where,

$$s_s^2(M) = \frac{1}{2N_s} \cdot (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \quad (11b)$$

and

$$s_e^2(M) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \cdot F_{ST}^{*}[1 - 1/(2N_s)]. \quad (11c)$$

It is instructive to review several points before proceeding. First, the variance of the admixture estimate does *not* tend to zero by increasing the number of individuals sampled; it tends to zero as the number of loci assayed is increased. Second, ELSTON (1971) introduced weighted least squares and maximum likelihood estimators for μ under the assumption that the hybrid population has not evolved by genetic drift. It can be shown that $s_s^2(M)$ is the estimated variance of both of these estimators. Third, although Wright's standardized variance definition of F_{ST} has been used so far, the correlational definitions of F_{ST} (WRIGHT 1965; COCKERHAM 1969, 1973; WEIR and COCKERHAM 1984) apply equally well. As points of reference, the parameter $\pi_i \cdot (1 - \pi_i)$ is equivalent to the parameter α_1 defined by REYNOLDS, WEIR and COCKERHAM (1983) and implicit in WEIR and COCKERHAM (1984), and $\sigma_e^2(p_{hi})$ is equivalent to their $\alpha_1\theta$.

EXTENSION TO MULTIPLE ALLELES, DOMINANCE AND UNCERTAIN PARENTAL POPULATIONS

Neither multiple alleles nor dominance presents a serious barrier to the methods described above. Multiple alleles can be accommodated as follows. Expand $\mathbf{X}$ and $\mathbf{y}$ so that P_{ki} is the frequency of the ith allele the kth population. The alleles may be at the same locus, or at separate loci, but one allele from each locus must be discarded from analysis to eliminate redundancy of information. It does not matter which allele from a locus is discarded from analysis (APPENDIX). For the matrix $\mathbf{V}$, retain the diagonal elements as they have been defined, but in each off-diagonal position use $V_{ij} = -E[P_{hi}] \cdot E[P_{hj}]$, where the alleles i and j segregate at the same locus. The expectation of MSE remains $[1/2N + F_{ST} \cdot (1 - 1/2N_s)]$, because an estimate of F_{ST} can be obtained from all pairs of alleles at a locus (NEI 1965), *i.e.*,

$$-E\left[\frac{(P_{hi} - \pi_i) \cdot (P_{hj} - \pi_j)}{\pi_i\pi_j}\right] = F_{ST}. \quad (12)$$

Dominance among alleles at a locus will not affect $s_e^2(M)$, but the sampling portion of the variance, $s_s^2(M)$, is altered. Eq. 4 still estimates $s^2(M)$, but its partition into $s_s^2(M)$ and $s_e^2(M)$ requires special treatment. To obtain $s_s^2(M)$, $\mathbf{V}$ is replaced with a new matrix, $\mathbf{V}^*$. The elements of $\mathbf{V}^{*-1}$ are the sampling variances of hybrid population allele frequencies obtained using maximum likelihood estimation procedures. Following LI (1976), we can obtain these elements from the equation

$$(V^{*-1})_{ij} = \frac{1}{2} \cdot \sum_{c=1}^{C}\left[\frac{1}{PH_c} \cdot \frac{\partial PH_c}{\partial\theta_i} \cdot \frac{\partial PH_c}{\partial\theta_j}\right] \quad (13)$$

where PH_c is the probability of the cth phenotypic class ($c = 1, \cdots, C$) in the hybrid population, and $\partial PH_c/\partial\theta_i$ is the partial derivative of the phenotypic class with the parameter of interest (*i.e.*, P_{hi}). Making this substitution,

$$s_s^2(M) = \frac{1}{2N_s} \cdot (\mathbf{X}^T\mathbf{V}^{*-1}\mathbf{X})^{-1} \quad (14)$$

and $s_e^2(M)$ should be estimated,

$$s_e^2(M) = s^2(M) - s_s^2(M). \quad (15)$$

If the alleles are codominant, $\mathbf{V}^*$ reduces to $\mathbf{V}$ and all estimation procedures remain the same.

The assumption that parental population allele frequencies are known without error is the final point of concern. While this assumption can never be fully met, it is generally assumed (REED 1969; ADAMS and WARD 1973) that allele frequencies obtained as unbiased estimates, with small standard errors, from the modern descendants of the correct ancestral populations will serve the purpose. This should be true for the method described above. As long as there is no systematic error in the estimation of ancestral frequencies, the main impact of uncertain ancestral frequencies should be the amplification of MSE. Hence, our estimate of $s^2(M)$ should be appropriate, as should our estimate of $s_s^2(M)$; however, the estimate of $s_e^2(M)$ will include error in estimation of ancestral allele frequencies as well as genetic drift. If parental population allele frequency estimates are obtained from very small samples, then following CAVALLI-SFORZA and BODMER (1971) and LONG and SMOUSE (1983), the matrix

$$\mathbf{U} = [\mathbf{V}/(2N_s) + M^2[\mathbf{V}_1/(2N_1)] + (1 - M)^2[\mathbf{V}_2/(2N_2)]] \quad (16)$$

should replace the matrix $\mathbf{V}$ in Equation 3–5. The

matrices $\mathbf{V}_1$ and $\mathbf{V}_2$ are the variance-covariance matrices of allele frequencies in the parental populations, and N_1 and N_2 are their sample sizes.

HETEROGENEITY OF ADMIXTURE ESTIMATES

CAVALLI-SFORZA and BODMER (1971) presented the heterogeneity chi-square statistic,

$$\chi^2_{(I-1)} = \Sigma_i(M_i - M)^2/V(M_i) \qquad (17)$$

where M is the weighted average of admixture proportions computed from all alleles, M_i is the estimate from the ith allele, and $V(M_i)$ is the variance of M_i. If a total of "I" independent alleles are assayed, $\chi^2(I-1)$ follows a chi-square distribution with $I-1$ degrees of freedom. Our WLS estimate differs very slightly from the weighted average that they use, and our estimates of $V(M_i)$ will be identical to theirs if the weight matrix U, and the variance formula (Equation 11b) are employed. Since Equation 11b assumes no evolutionary error, this test will detect heterogeneity among admixture estimates regardless of whether it is due to natural selection or genetic drift. Thus, the two most (evolutionarily) interesting sources of error in admixture estimates are confounded.

The test can easily be improved by letting M_i and $V(M_i)$ pertain to the ith locus instead of the ith allele, thereby eliminating redundancy of information, and using the variance formula

$$V(M_i) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \cdot \text{MSE} \qquad (18)$$

where the vector $\mathbf{X}$ is reduced to only those alleles at the ith locus, but MSE pooled from the analysis of all loci. Thereby, the variance of the admixture estimates accounts for the error introduced by sampling the hybrid population and for the error resulting from genetic drift (including error in parental population allele frequencies) in the hybrid population.

MODELS OF ADMIXTURE AND ESTIMATION OF N_e

The expectations of F_{ST}, MSE and $s^2(M)$ can be obtained from first principles of population genetics when the dynamics of admixture and drift can be specified. Two such models will be evaluated here. In the first model, the hybrid population is formed by a single event of intermixture between two parental populations. Hereinafter this model is referred to as *intermixture*. In the second model, there is gene diffusion from a donor population (parent) into a recipient (hybrid) population, but not vice versa. This process is classically referred to as *Gene Flow* and it is thought to apply to major racial groups such as Afro-Americans.

Intermixture: The admixture effects of this process are diagrammed in Figure 2A. The admixture proportions are established at the inception of the process

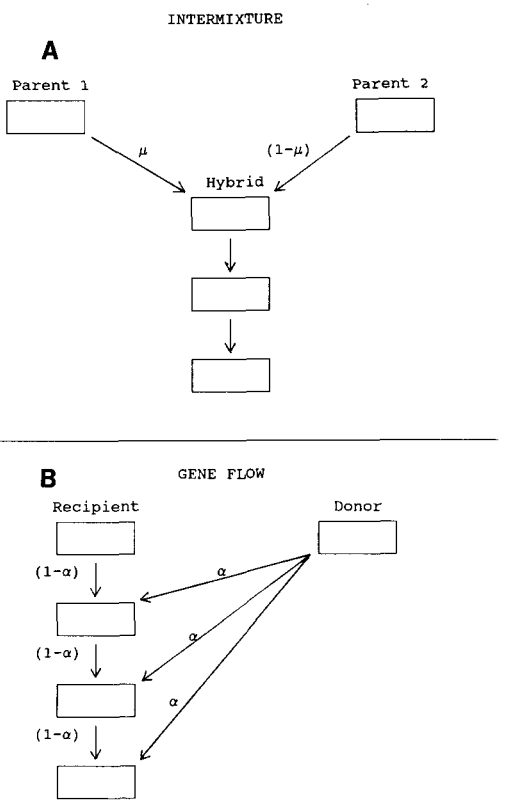

FIGURE 2.—Admixture models. The dynamics of population admixture for the intermixture and gene flow models are diagrammed above.

and all subsequent evolution is by genetic drift. Therefore, a single μ will apply to all generations while F_{ST}, MSE and $s^2(M)$ progressively increase with time. It is well established (*cf.* WRIGHT 1969; p. 345) that, F_{ST}, after t generations of drift will be

$$\begin{aligned} {}^{(t)}F_{ST} &= 1 - \left[1 - \frac{1}{2N_e}\right]^t \\ &\approx 1 - \exp(-t/2N_e). \end{aligned} \qquad (19)$$

Accordingly, we can find the expected MSE, at any time, t, for any sample size

$$\begin{aligned} E[\text{MSE}] &= \frac{1}{2N_s} + {}^{(t)}F_{ST}\left[1 - \frac{1}{2N_s}\right] \\ &\approx \frac{1}{2N_s} + [1 - \exp(-t/2N_e)] \cdot \left[1 - \frac{1}{2N_s}\right]. \end{aligned} \qquad (20)$$

With this relationship established, it is possible to estimate the variance effective size of the hybrid population, using

$$N_e^* = -t/[2 \cdot \ln\{1 - F_{ST}^*\}] \qquad (21)$$

where F_{ST}^* is the estimate of F_{ST} obtained from Equation 10. Application of this formula requires a good estimate of the length of time since the formation of the hybrid population. Such estimates are readily available for many of the populations that are of most interest.

Gene flow: As shown in Figure 2B, the recipient (hybrid) population receives a constant proportion, α, of its genes from a donor population every generation. The cumulative portion of genes from the donor population, μ, is not constant; after t generations of admixture, it is given by $\mu = 1 - (1 - \alpha)^t$. The equation to estimate α follows from (GLASS and LI 1953)

$$\hat{\alpha} = -\ln(1 - M)/t \tag{22}$$

where M is the estimate of μ obtained from Equation 3.

The admixture dynamics of this model are well established (GLASS and LI 1953), but the effects of genetic drift in the hybrid have not been investigated heretofore. Recall that ${}^{(t)}F_{ST}$ (defined for an ith allele) is

$${}^{(t)}F_{ST} = \frac{{}^{(t)}\sigma_e^2(p_{hi})}{{}^{(t)}\pi_i[1 - {}^{(t)}\pi_i]} \tag{23}$$

where ${}^{(t)}\sigma_e^2(p_{hi})$ is the evolutionary variance of the ith allele frequency (in the hybrid population) in the t-th generation, and ${}^{(t)}\pi_i$ is its expected allele frequency. ${}^{(t)}\pi_i$ is obtained simply from $(1 - {}^{(t)}\mu) \cdot P_{1i} + {}^{(t)}\mu \cdot P_{2i}$. It will change every generation until it equals the allele frequency in the donor population. By extension of the principles of variance of gene frequencies in finite populations (WRIGHT 1969, p. 346) to account for migrants, the evolutionary variance of the allele frequency under gene flow is obtained from

$${}^{(t)}\sigma_e^2(p_{hi}) = \sum_{j=1}^{t} [[1 - 1/(2N_e)]^{j-1}][(1 - \alpha)^2]^j \frac{{}^{t}\pi_i(1 - {}^{t}\pi_i)}{2N_e}. \tag{24}$$

${}^{(t)}\sigma_e^2(p_{hi})$ is combined with ${}^{(t)}\pi_i$ to yield an estimate of ${}^{(t)}F_{ST}$ (Equation 23). The expected MSE assuming gene flow remains (Equation 23 is used for ${}^{(t)}F_{ST}$).

$$E[\text{MSE}] = \frac{1}{2N_s} + {}^{(t)}F_{ST} \cdot \left[1 - \frac{1}{2N_s}\right].$$

The conditions of the gene flow model are unlike those of the intermixture model because ${}^{(t)}F_{ST}$ will approach an equilibrium value less than unity. The equilibrium point is

$$\hat{F}_{ST} = \frac{(1 - \alpha)^2}{2N_e - (2N_e - 1)(1 - \alpha)^2} \tag{25}$$

which is identical to the equilibrium F_{ST} under WRIGHT's (1969, p. 291) island model of population structure. If F_{ST} is close to its equilibrium value, and α is small, say $\alpha \leq 0.02$, the formula

$$N_e^* = \frac{1 - F_{ST}^*}{4\hat{\alpha}F_{ST}^*} \tag{26}$$

can be used to estimate the variance effective size of the hybrid population. However, these are very restrictive conditions and it will almost always be preferable to manipulate Equation 24 numerically to estimate the effective population size.

Gene flow can be profitably considered a special case of the island model, but some new considerations regarding the model dynamics must be recognized. While WRIGHT was concerned with the equilibrium of the model, the dynamics have been explored as a special case of the migration matrix (BODMER and CAVALLI-SFORZA 1968; SMITH 1969; MORTON 1969; IMAIZUMI, MORTON and HARRIS 1970). Here it is assumed that the migration/drift process begins with a set of demes drawn independently from a panmictic population. Differentiation among the demes then proceeds up to the equilibrium point. An important feature of this assumption is that the *expected* allele frequencies at a locus do not change from generation to generation, nor do the *expected* allele frequencies vary across demes. Under these conditions there is a single value of F_{ST} that pertains to all alleles at all loci. With gene flow, each *expected* allele frequency changes until convergence with the donor population is realized. Until convergence, a single F_{ST} will not apply to all alleles at all loci. Equation 10 provides an average F_{ST}^* over all alleles analyzed. Although F_{ST} does not vary greatly among alleles, estimates of the effective size of the population should be computed separately for each allele and then averaged. Since it is the reciprocal of N_e that enters into Eq. 24, the harmonic average of the estimated effective population size from each allele should be employed.

Distribution of N_e^*: The usefulness of effective population size estimates depends heavily on their precision. There are two major barriers to the adoption of usual statistical procedures for evaluating this. First, the estimation methods for N_e render a standard error formula for N_e^* algebraically intractable. Second, even if it were possible to derive a standard error for N_e^*, its distribution is heavily right skewed and asymmetric confidence intervals must be used. With these considerations in mind, the bootstrap procedure (EFRON 1979; EFRON and GONG 1983; DIACONIS and EFFRON 1983) is an appropriate tool for approximating the distribution of N_e^*.

Briefly, the steps of the bootstrap procedure are as follows. (1) An estimate of F_{ST} is obtained for each allele analyzed, following the expectations established in Equation 9. Occasionally these estimates will be negative, in which case zero should be substituted for the negative value. (2) An estimate of N_e should be obtained from F_{ST}^* for each allele using Equation 21, or Equation 24, depending on the appropriate admixture model. If F_{ST}^* is negative, then the only realistic estimate of N_e is infinity. (3) The observed frequency

distribution of N_e estimates is resampled with replacement a large number of times (say 10,000). The number of variates in each bootstrap sample is the same as the number of alleles analyzed in the original data set and the sampling is done with replacement. (4) For each bootstrap sample, an average effective population size is computed as the harmonic average of the bootstrapped variates. The approximate confidence limits on N_e^* are obtained by using the appropriate upper and lower percentiles of the bootstrap distribution.

HUMAN POPULATION EXAMPLES

The utility of the above theory for the analysis of actual populations will now be demonstrated on three human population examples. In each example, ancestral contributions to the hybrid will be estimated, a chi-square test for heterogeneity of admixture estimates among alleles will be performed, and additional insights into the genetic structure of the hybrid population will be gleaned from F_{ST}^*. The first two examples are Afro-American populations and the gene flow model will be more appropriate. The third example is a Yanomamö Indian village that became admixed with neighboring Ye'cuana Indians. The Intermixture model will be more appropriate for this population.

Claxton Georgia (Afro-American): The Claxton Afro-American population was sampled in the early 1960s (COOPER *et al.* 1963); its census size was 1287 individuals at this time. Genetic data consisting of 15 alleles at 12 informative loci {ABO, Duffy, G6PD, Hb, Hp, Js, Kidd, M, P, Rh, S, T} are suitable for admixture analysis. The sample sizes varied between 133 and 304 individuals per locus. Like other Afro-American populations, it is assumed that the Claxton population has received Caucasian genes for about 12 generations.

Originally, WORKMAN and co-workers (1963) concluded that admixture estimates obtained from their markers fell into two categories. In the first category, the marker alleles attributed about 10% of the Claxton gene pool to Caucasian origin and about 90% to African origin. Most genetic markers fell into this category and the admixture proportions are similar to expectations based on ethnohistorical considerations. Appreciably higher estimates of the Caucasian contribution to the Claxton Afro-American population were obtained from alleles in the second category. The alleles falling into this category were $G6PD^-$, Hp^1, Tfd and Hb^S. This group was identified as selectively biased because of the presence of G6PD and Hb^S. Significant heterogeneity chi-squares for these data have been observed in subsequent analyses (CAVALLI-SFORZA and BODMER 1971; ADAMS and WARD 1973) but the overall impression is that genetic drift is likely to be a major contributor. The significance of the heterogeneity chi-squares will be reexamined here using the method developed in this paper. Estimates of ancestral allele frequencies for the present analyses are those of ADAMS and WARD (1973).

Sapelo Island (Afro-American): Later in the 1960s, the Afro-American population living on Sapelo Island, Georgia, was sampled for the same genetic loci (BLUMBERG and HESSER 1971) as the Claxton population. There were only 211 Afro-American residents on Sapelo Island and sample sizes ranged from 141 to 191 individuals per locus. Although gene flow occurred over the same 12-generation period, the Sapelo Island Afro-Americans were thought to be less admixed, and to have been somewhat isolated from other Afro-American populations as well. A dichotomous grouping of admixture estimates was not observed for Sapelo Afro-Americans, but there was a significant correlation of admixture estimates between Sapelo and Claxton Afro-Americans. This correlation of admixture estimates was interpreted as good evidence for natural selection (BLUMBERG and HESSER 1971). Later, ADAMS and WARD (1973) found a significant heterogeneity chi-square for Sapelo Island; however, they cautioned that genetic drift as well as natural selection could have produced this result. The parental population allele frequencies compiled by Adams and Ward (1973) will also be used for reanalyzing Sapelo Island.

Borabuk (Yanomamö Indian): Most of the Yanomamö Indians of Northwestern Brazil and Southern Venezuela have been isolated from both Caucasian and other Amerindian populations until quite recently. An exception to this isolation is provided by the Yanomamö village Borabuk. Historical accounts (CHAGNON *et al.* 1970) place this village in close proximity with a Ye'cuana Indian village, Huduaduña, during the later part of the nineteenth century. Genetic exchanges took place during this time and a few Ye'cuana men and women are known to have had an extraordinary impact on the Borabuk gene pool (CHAGNON *et al.* 1970). During the twentieth century the Borabuk population has remained relatively isolated from the Ye'cuana and other Yanomamö villages, thus the Intermixture model is more appropriate to this population. Approximately 6 generations passed between the influx of Ye'cuana genes into Borabuk and the time of sampling. Polygyny and other aspects of Yanomamö social organization (CHAGNON 1968), and the fact that this village was nearly decimated by a disease epidemic early in this century, indicate that Borabuk's effective population size is quite small. Borabuk, like other Yanomamö villages fluctuates in actual size, the seventy-five indi-

TABLE 2

Admixture estimates and variances

Hybrid population	First ancestor	Second ancestor	Standard error	$s^2(M)$	$s_s^2(M)$	$s_i^2(M)$
Claxton	African 0.864	Caucasian 0.136 ($\hat{\alpha} = 0.012$)	±0.051	0.0026	0.0019	0.0007
Sapelo Island	African 0.932	Caucasian 0.068 ($\hat{\alpha} = 0.006$)	±0.055	0.0300	0.0298	0.0002
Borabuk	Yanomamö 0.640	Ye'cuana 0.360	+0.194	0.0377	0.0318	0.0059

TABLE 3

Heterogeneity analysis

Claxton				Sapelo				Borabuk			
Locus	M_i	$V(M_i)$	χ_i^2	Locus	M_i	$V(M_i)$	χ_i^2	Locus	M_i	$V(M_i)$	χ_i^2
ABO	0.15	0.01	0.01	ABO	−0.01	0.01	0.47	P	1.00	24.68	0.01
Rh	0.11	0.14	0.00	Rh	0.44	0.25	0.56	Fy	−6.00	21.99	2.01
M	−0.15	10.69	0.01	M	−0.85	18.83	0.04	Jk	1.23	2.31	0.15
S	−0.30	0.31	0.61	S	−0.05	0.53	0.03	Le	0.92	0.29	0.27
Fy	0.11	0.01	0.16	Fy	0.07	0.00	0.00	Di	−1.00	1.40	1.92
P	−0.23	0.10	1.31	P	0.80	0.18	3.02	MNS	1.18	0.39	0.74
Jk	−0.59	0.23	2.30	Jk	−0.12	0.40	0.08	Rh	0.46	0.09	0.38
Js	−0.05	0.12	0.30	Js	−0.70	0.22	2.69	Hp	1.02	0.15	0.98
T	−0.38	0.37	0.70	T	−0.41	0.65	0.35	Gc	0.00	2.95	0.14
Hp	0.63	0.05	5.06	Hp	0.23	0.08	0.33	PGM	0.50	6.06	0.00
G6PD	0.34	0.07	0.58	G6PD	0.29	0.13	0.36	ACP	−1.00	1.12	2.41
Hb	0.53	0.15	1.03	Hb	0.44	0.28	0.49				
Total χ^2			12.07				8.43				9.00
d.f.			(11)				(11)				(10)

Locus abbreviations: blood groups: (ABO) ABO, (Rh) rhesus, (M, S and MNS) MNSs, (Fy) Duffy, (P) P, (Jk) Kidd, (Js) Sutter, (Di) Diego. Erythrocyte and serum protein loci: (Hp) haptoglobin, (Gc) group specific component, (PGM) phosphoglucomutase, (ACP) red cell acid phosphatase, (G6PD) glucose 6-phosphate dehydrogenase, (Hb) hemoglobin β-chain, (T) phenylthiocarbamide tasting.

viduals sampled represented almost all people present during the visit, the *de facto* size was probably less than 100 individuals. Allele frequencies at eleven loci (ACP, Diego, Duffy, Gc, Hp, Kidd, Le, MNSs, P, PGM, Rh) for Borabuk, Huduaduña, and the Shamatari (Yanomamö ancestor) are given in an earlier paper (LONG and SMOUSE 1983).

RESULTS

Estimates of the ancestral contributions to each of these three populations are presented in Table 2. Admixture estimates for the Afro-American populations are similar to the previously published estimates for these populations (ADAMS and WARD 1973). The estimate of Yanomamö ancestry for Borabuk presented here is lower than, but not significantly different from, the value published earlier (LONG and SMOUSE 1983). The difference between the two estimates is due to the fact that in the 1983 study an admixture estimate was calculated for each locus individually, and then a weighted average across loci was obtained; in the present study a single estimate was iterated over the entire set of loci.

The standard errors of admixture estimates are quite high. More interesting are the variances of the admixture estimates, and their partition into sampling and evolutionary components. In all three populations, the vast majority of the variance of the admixture estimates is due to evolutionary, rather than sampling error. Therefore, the variance of estimated admixture proportions can be seriously affected by genetic drift, despite earlier claims for the Afro-American populations (WORKMAN, BLUMBERG and COOPER 1963; BLUMBERG and HESSER 1971).

Admixture estimates computed by locus, M_i, their variances, $V(M_i)$, and contributions, χ_i^2, to the heterogeneity chi-square statistic are presented for the three populations in Table 3. Although the admixture estimates appear heterogeneous across loci, it is clear that each estimate has an enormous variance associated with it, and the chi-squares computed over all loci within these three populations do not approach

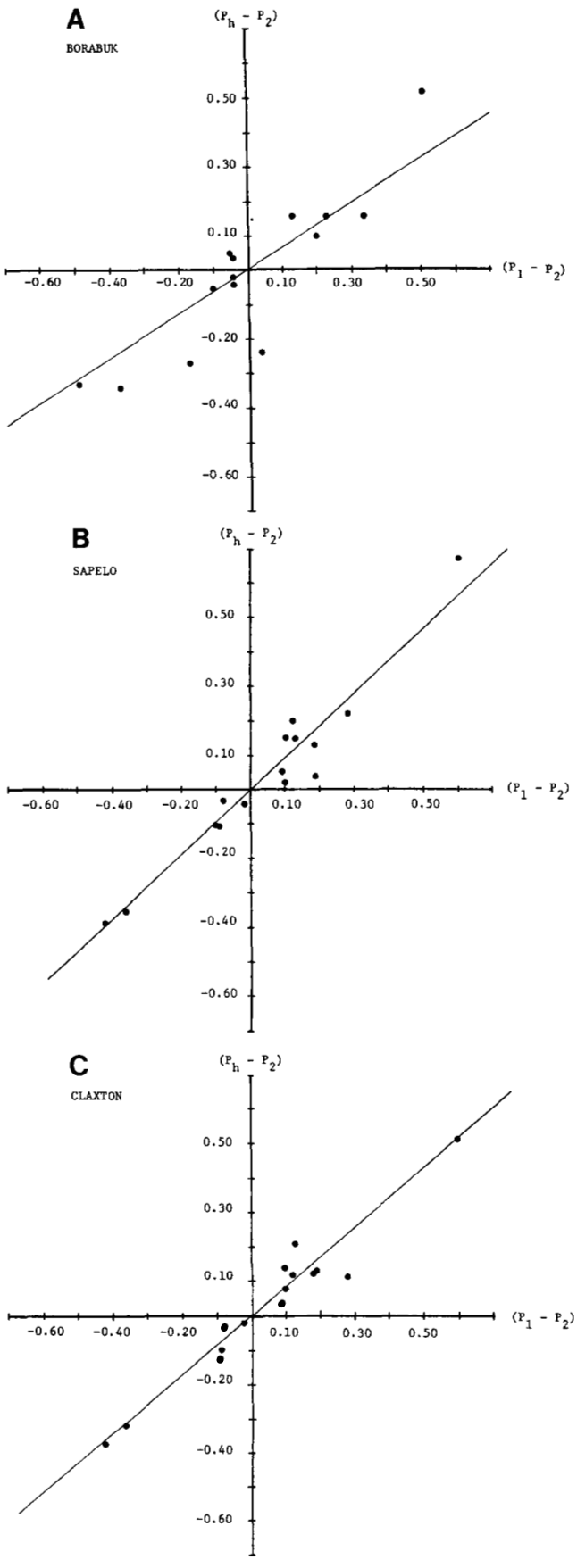

FIGURE 3.—The relationships between $(\tilde{P}_{hi} - P_{2i})$ and $(P_{1i} - P_{2i})$. This relationship should be linear with slope μ (by Equation 2). The slopes of the plotted lines are the estimated proportions of African ancestry in the Afro-American populations and the proportion of Yanomamö ancestry in the Borabuk population.

TABLE 4

Population structure interpretations

Hybrid population	MSE	d.f.	F_{ST}^*	Census size (N_a)	Effective size (N_e)	Ratio N_e/N_a
Claxton	0.017	11	0.013	1287	349	0.27
Sapelo	0.030	11	0.027	211	216	1.02
Borabuk	0.158	10	0.152	≤100	≥18	≥0.18

statistical significance. Despite the broad scatter of admixture estimates from the individual loci, the goodness of fit of the linear admixture model is revealed in Figure 3 by plots of $(\tilde{P}_{hi} - P_{2i})$ versus $(P_{1i} - P_{2i})$, as suggested by Eq. 2. While the graphical displays are compelling, their interpretations must be tempered with the knowledge that the points are heteroscedastic and that interdependence of error terms exists among the points representing alleles segregating at the same locus.

Further analyses of these populations are presented in Table 4. Since a significant χ^2 is not obtained in any of these populations, evidence for natural selection is lacking, and F_{ST} and the effective population size was estimated for each population. The relative ranking of F_{ST}^* for these populations, Claxton (0.013) < Sapelo (0.030) < Borabuk (0.152), conforms to what would be expected from the social demography of these groups. The estimates of N_e are biologically reasonable for Claxton (349) and Borabuk (≥18). This is evidenced by the fact that the ratios of their effective sizes to census sizes are, respectively, 0.27 and ≥0.18. The estimated N_e for Sapelo Island (216) is unusually high since this would suggest that the effective size is greater than the census size.

The percentiles of the bootstrapped distributions for these estimates are presented in Table 5. These distributions are widely dispersed; however, the 95% confidence interval for Sapelo Island [119, 483] still suggests that Sapelo Island had an uncharacteristically large effective population size. A likely explanation for this result is that the Sapelo Island population has received immigrants from other Afro-American communities. In other words, genetic drift on Sapelo Island has been affected by immigrants from both the Caucasian population and other Afro-American populations. Since there were only four recorded cases of miscegenation in the history of Sapelo Island, and very few permanent Caucasian inhabitants of the island (BLUMBERG and HESSER 1971), it is not unlikely that a portion of the Caucasian genes in the Sapelo Island population were actually introduced by admixed Afro-Americans.

TABLE 5

Percentiles of bootstrapped distributions of N_e^*

Percentile	Population		
	Claxton	Sapelo	Borabuk
Min	117	82	7
1	162	109	13
2.5	178	119	14
5	194	128	16
10	216	140	17
25	265	166	21
50	341	205	26
75	457	265	34
90	618	347	46
95	745	413	58
97.5	881	483	72
99	1069	596	98
Max	2305	2150	614

Each distribution is based on $n = 10{,}000$ bootstrapped trials.

DISCUSSION

The objective of this study has been to provide a framework for estimating the effects of genetic drift and admixture simultaneously. Thereby, the value of admixture estimates for analysis of hybrid population structure is enriched, and an appropriate null hypothesis for the detection of natural selection can be formulated. It was shown that Wright's fixation index F_{ST} is closely related to the error variance of allele frequencies predicted by admixture, and that F_{ST} can be estimated from a multiple locus sample of hybrid population allele frequencies, providing that there is information on the allele frequencies in the contributing parental populations.

There are two existing methods for estimating admixture proportions that allow for both evolutionary and sampling error (THOMPSON 1973; WIJSMAN 1984); each of these methods requires independent knowledge of the variance effective population size of the hybrid population over its entire evolutionary history, and that the admixture process was that of the Intermixture Model. However, variance effective sizes of populations are seldom known, and the Intermixture Model does not apply to all admixed populations. Superimposition of the intermixture model onto populations for which gene flow is more appropriate would clearly be misleading. Perhaps the greatest advantage of the estimation procedures developed here is that F_{ST} can be estimated regardless of the underlying population structure, and the drift corrected chi-square test for heterogeneous admixture proportions can always be applied.

Evolutionary variance in admixture estimates overwhelmed sampling variance in analyses of three admixed human populations. This finding is especially important since it was originally argued that drift should not have been a significant force in two of these populations (see WORKMAN, BLUMBERG and COOPER 1963; BLUMBERG and HESSER 1971). The chi-square statistics for heterogeneity of admixture estimates did not attain statistical significance in any of these populations. Earlier, significant chi-squares had been obtained for Claxton and Sapelo Island Blacks (ADAMS and WARD 1973), although these investigators were duly cautious in their interpretation.

In the absence of evidence for natural selection, it is possible to estimate effective population sizes. N_e is a fundamentally important parameter in population genetics that links simplified theoretical models to natural and artificial populations; yet N_e is notoriously difficult to estimate, as noted by LAURIE-AHLBERG and WEIR (1979), HILL (1981), WOOD (1987), and WAPLES (1989) in other contexts. The estimation methods presented here should prove to valuable to the analysis of hybrid populations.

Fundamental to the interpretation of estimates of N_e is knowledge of the distributions the estimates will follow. The bootstrap method suggested here is only a rough approximation. Ideally, if the original number of alleles sampled is large, then the empirical distribution will closely approximate the true distribution. For smaller samples, such as those analyzed for the three example human populations, the interpretations should be cautious. It should also be noted that the bootstrap method is not without underlying assumptions. It clearly assumes that the original data were drawn independently from the same distribution. These assumptions are unlikely to hold considering that alleles may co-segregate within loci and that the distribution of N_e^* may vary from allele to allele. Refinements on this approximation will be a fruitful area for further research.

As a final point, the utility of the procedures developed here should be applicable for purposes beyond analysis of population genetic structure. For example, epidemiological interest in hybrid populations has been increasing. Hybrid populations have recently been shown to present unique opportunities for the estimation of modes inheritance for diseases of complex etiology (CHAKRABORTY and WEISS 1986) and for the estimation of recombination fractions (CHAKRABORTY and WEISS 1988). Precise estimates of ancestral contributions are required in each of these circumstances and the methods provided here will be broadly applicable. The results of this study indicate that drift error must always be considered in admixture estimates and that simultaneous assay of many marker loci will be preferred.

I wish to thank LYNN JORDE, KEN MORGAN, ALAN ROGERS, PETER SMOUSE, TROY TUCKER, BRUCE WEIR and an anonymous reviewer for their comments on earlier drafts of this paper. This work has been greatly improved by the computer simulation studies of JAY SOBUS which led me to the correct derivations of many of the formulas presented here. This research was supported by a National Institutes of Health BSRG grant from the Vice President of Research, University of New Mexico, to JCL.

Note Added in Proof: An anonymous reviewer has shown that the statistics M and MSE, as defined in this paper, can be obtained from a set of equations that do not require eliminating one allele from each locus for computation.

LITERATURE CITED

ADAMS, J., and R. H. WARD, 1973 Admixture studies and detection of selection. Science **180:** 1137–1143.

BLUMBERG, B. S., and J. E. HESSER, 1971 Loci differentially affected by selection in two American Black Populations. Proc. Natl. Acad. Sci. USA **68:** 2554–2558.

BODMER, W. E., and L. L. CAVALLI-SFORZA, 1968 A migration matrix model for the study of random genetic drift. Genetics **59:** 565–592.

CAVALLI-SFORZA, L. L., and W. F. BODMER, 1971 *The Genetics of Human Populations.* Freeman & Co, San Francisco.

CHAGNON, N., 1968 *Yanomamö: The Fierce People.* Holt, Rinehart & Winston, New York.

CHAGNON, N., J. V. NEEL, L. WEITKAMP, H. GERSHOWITZ and M. AYRES, 1970 The influence of cultural factors on the demography and pattern of gene flow from the Makiritare to the Yanomamö Indians. Am. J. Phys. Anthropol. **32:** 339–350.

CHAKRABORTY, R., 1986 Gene admixture in human populations: models and predictions. Yearb. Phys. Anthropol. **29:** 1–43.

CHAKRABORTY, R., and K. M. WEISS, 1986 Frequencies of complex diseases in hybrid populations. Am. J. Phys. Anthropol. **70:** 489–503.

CHAKRABORTY, R., and K. M. WEISS, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc. Natl. Acad. Sci. USA **85:** 9119–9123.

COCKERHAM, C. C., 1969 The variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

COOPER, A. J., B. S. BLUMBERG, P. L. WORKMAN and J. R. MCDONOUGH, 1963 Biochemical polymorphic traits in a U.S. White and Negro population. Amer. J. Hum. Genet. **15:** 420–428.

DIACONIS, P., and B. EFRON, 1983 Computer intensive methods in statistics. Sci. Am. **249:** 116–130.

EFRON, B., 1979 Bootstrap methods: another look at the jackknife. Ann. Statist. **7:** 1–26.

EFRON, B., and G. GONG, 1983 A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Statist. **37:** 36–48.

ELSTON, R. C., 1971 The estimation of admixture in racial hybrids. Ann. Hum. Genet. **35:** 9–17.

GLASS, B., and C. C. LI, 1953 The dynamics of racial intermixture—an analysis based on the American Negro. Am. J. Hum. Genet. **5:** 1–20.

HILL, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. **38:** 209–216.

IMAIZUMI, Y., N. E. MORTON and D. E. HARRIS, 1970 Isolation by distance in artificial populations. Genetics **66:** 569–582.

LAURIE-AHLBERG, C. C., and B. S. WEIR, 1979 Allozyme variation and linkage disequilibrium in some laboratory populations of *Drosophilia melanogaster.* Genetics **74:** 175–195.

LI, C. C., 1976 *First Course in Population Genetics.* Boxwood Press, Pacific Grove, Calif.

LONG, J. C., 1986 The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. Genetics **112:** 629–647.

LONG, J. C., and P. E. SMOUSE, 1983 Intertribal geneflow between the Ye'cuana and Yanomamö: genetic analysis of an admixed village. Am. J. Phys. Anthropol. **61:** 411–422.

MORTON, N. E., 1969 Human population structure. Annu. Rev. Genet. **3:** 53–73.

NEI, M., 1965 Variation and covariation of gene frequencies in subdivided populations. Evolution **119:** 256–258.

NETER, J., and W. WASSERMAN, 1974 *Applied Linear Statistical Models.* R. D. Irwin, Inc., Homewood, Ill.

REED, T. E., 1969 Caucasian genes in American Negroes. Science **165:** 762–768.

REYNOLDS, J., B. S. WEIR and C. C. COCKERHAM, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105:** 767–799.

SMITH, C. A. B., 1969 Local fluctuations in gene frequencies. Ann. Hum. Genet. **32:** 251–260.

THOMPSON, E. A., 1973 The Icelandic admixture problem. Ann. Hum. Genet. **37:** 69–80.

WAPLES, R. S., 1989 A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics **121:** 379–391.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

WIJSMAN, E. M., 1984 Techniques for estimating genetic admixture and applications to the problems of the origin of the Icelanders and the Ashkenazi Jews. Hum. Genet. **67:** 441–448.

WOOD, J. W., 1987 The genetic demography of the Gainj of Papua New Guinea. III. Determinants of effective population size. Am. Nat. **129:** 165–187.

WORKMAN, P. L., 1968 Gene flow and the search for natural selection in man. Hum. Biol. **40:** 260–279.

WORKMAN, P. L., B. S. BLUMBERG and A. J. COOPER, 1963 Selection, gene migration and polymorphic stability in a US White and Negro population. Am. J. Hum. Genet. **15:** 429–435.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugenics **15:** 323–354.

WRIGHT, S., 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution **19:** 395–420.

WRIGHT, S., 1969 *Evolution and The Genetics of Populations. VOL 2, The Theory of Gene Frequencies.* University of Chicago Press, Chicago.

Communicating editor: B. S. WEIR

APPENDIX

Proof That Any Allele Can Be Dropped With the Weighted Least Squares Procedure

Consider a locus with an arbitrary number of alleles $A_1, A_2, \ldots A_Z$, with frequencies $(p_1, p_2, \ldots, p_Z)$ in the hybrid population and $(p_{11}, p_{21}, \ldots, p_{Z1})$ in the first parental population and $(p_{12}, p_{22}, \ldots, p_{Z2})$ in the second parental population. Define $X^T = [(p_{11} - p_{12}), (p_{21} - p_{22}), \ldots, (p_{Z1} - p_{Z2})]$ and $y^T = [(p_1 - p_{12}), (p_2 - p_{22}), \ldots, (p_Z - p_{Z2})]$. The equation for the *Weighted Least Square* estimate of the admixture proportions is given by

$$M = (\mathbf{Xs}^T\mathbf{V}^{-1}\mathbf{Xs})^{-1}\mathbf{Xs}^T\mathbf{V}^{-1}\mathbf{ys} \qquad \text{(A)}$$

where the "shortened" vectors **Xs** and **ys** are obtained from the "full" vectors **X** and **y** by eliminating the Zth element. V has dimension ($Z - 1$ by $Z - 1$) and it is the variance-covariance matrix of the multinomial distribution ($N = 1$). It is easily verified that V^{-1} can

be written out explicitly as

$$v^{-1} = \begin{bmatrix} \frac{1}{p_1} + \frac{1}{p_Z}, & \frac{1}{p_Z}, & \cdots, & \frac{1}{p_Z} \\ \frac{1}{p_Z}, & \frac{1}{p_2} + \frac{1}{p_Z}, & \cdots, & \frac{1}{p_Z} \\ \vdots & \vdots & \vdots\ \vdots & \vdots \\ \frac{1}{p_Z}, & \frac{1}{p_Z}, & \cdots, & \frac{1}{p_{Z-1}} + \frac{1}{p_z} \end{bmatrix}$$

Equation A1 can now be rewritten as follows

$$M = \frac{\sum_{i=1}^{Z-1} y_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_Z} \cdot \sum_{j=1}^{Z-1} x_j \right]}{\sum_{i=1}^{Z-1} x_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_Z} \cdot \sum_{j=1}^{Z-1} x_j \right]}. \qquad \text{(A2)}$$

Let the allele to be dropped from analysis be designated A_d. So far, we have let A_Z be A_d, now we will let A_d be any of the Z alleles. We will show that this leaves Equations 1 and 2 unaffected. Some special notation is required,

$$\sum_{\substack{j=1, \\ j \neq d}}^{Z} x_j$$

will mean that a sum will be taken over all alleles, except A_d. For example, if $d = 3$ and $Z = 5$, then

$$\sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j = (x_1 + x_2 + x_4 + x_5).$$

Now, the estimator for admixture proportions can be rewritten as

$$M = \frac{\sum_{\substack{i=1 \\ i \neq d}}^{Z} y_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]}{\sum_{\substack{i=1 \\ i \neq d}}^{Z} x_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]} \qquad \text{(A3)}$$

To prove that any allele can be dropped, Equation A2 must equal Equation A3. We begin by proving the identity of the numerators

$$N_3 = \sum_{\substack{i=1 \\ i \neq d}}^{Z} y_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]$$

$$= \sum_{\substack{i=1 \\ i \neq d}}^{Z-1} y_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]$$

$$+ y_Z \left[x_Z \cdot \frac{1}{p_Z} + \frac{1}{p_D} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]$$

Now using the identities $y_Z = -\sum_{i=1}^{Z-1} y_i$ and $x_Z = -\sum_{i=1}^{Z-1} x_i$,

$$N_3 = \sum_{\substack{i=1 \\ i \neq d}}^{Z-1} y_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{k=1 \\ j \neq d}}^{Z} x_j \right]$$

$$+ \sum_{i=1}^{Z-1} y_i \left[\sum_{i=1}^{Z-1} x_i \cdot \frac{1}{p_i} - \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right].$$

Since $-x_d = \sum_{j=1\, j \neq d}^{Z} x_j$

$$N_3 = \sum_{\substack{i=1 \\ i \neq d}}^{Z-1} y_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_Z} \cdot \sum_{j=1}^{Z-1} x_j \right]$$

$$+ y_d \left[x_d \cdot \frac{1}{p_d} + \frac{1}{p_Z} \cdot \sum_{j=1}^{Z-1} x_j \right] = N_2.$$

To complete the proof, the identity of the denominators (D_2 and D_3) is now proved

$$D_3 = \sum_{\substack{i=1 \\ i \neq d}}^{Z} x_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]$$

$$= \sum_{\substack{i=1 \\ i \neq d}}^{Z-1} x_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]$$

$$+ x_Z \left[x_Z \cdot \frac{1}{p_Z} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right].$$

Now using the identity and $x_Z = -\sum_{i=1}^{Z-1} x_i$,

$$D_3 = \sum_{\substack{i=1 \\ i \neq d}}^{Z-1} x_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]$$

$$+ \sum_{i=1}^{Z-1} x_i \left[\sum_{i=1}^{Z-1} x_i \cdot \frac{1}{p_i} - \frac{1}{p_d} \cdot \sum_{\substack{j=1 \\ j \neq d}}^{Z} x_j \right]$$

Since $-x_d = \sum_{j=1\, j \neq d}^{Z} x_j$

$$D_3 = \sum_{\substack{i=1 \\ i \neq d}}^{Z-1} x_i \left[x_i \cdot \frac{1}{p_i} + \frac{1}{p_Z} \cdot \sum_{j=1}^{Z-1} x_j \right]$$

$$+ x_d \left[x_d \cdot \frac{1}{p_d} + \frac{1}{p_Z} \cdot \sum_{j=1}^{Z-1} x_j \right] = D_2.$$

Since it has been proved that the denominators of Equations A2 and A3 are equal, and that the numerators of Equations A2 and A3 are equal, it is proved that the same admixture estimate is obtained no matter which allele is eliminated from the system.