The Population Variance of the Proportion of Genetic Admixture in Human Intergroup Hybrids

(individual admixture proportion/group admixture proportion/California Negroes)

T. EDWARD REED

Departments of Zoology and Anthropology, University of Toronto, Toronto 5, Ontario, Canada

Communicated by James V. Neel, October 4, 1971

ABSTRACT For each individual in a human hybrid population there is a proportion μ_i , whose value is usually unknown, that expresses the fraction of his genes deriving from a specified parental population. The distribution of these individual proportions about the mean proportion μ is not known for any large hybrid population in man. It is of interest to know whether the population variance of individual proportions (μ_i) can be estimated from the variation between different, independent estimates of the mean proportion (μ).

This possibility was tested with data on Negroes of the Oakland, California area, by the use of some of the principles of analysis of variance. Even with a large sample and the useful Duffy blood-group system to indicate admixture, almost no information about the population variance of individual proportions is provided by between-sample variation in estimates of μ . It is concluded that group data on admixture proportions usually do not give useful information about the population variance. It is further concluded that a recent estimate of this variance by Shockley is in error.

Given two human populations, P_1 and P_2 , and a hybrid population H produced by matings between P_1 and P_2 , it may be possible, with suitable data and conditions, to estimate the relative genetic contributions of P_1 and P_2 to H . The minimal requirements are that there is some gene G , whose frequencies in P_1 and P_2 (p_1 and p_2 , respectively) are measurably different, and that, for the time period in question, interbreeding between P_1 and P_2 is the only important process determining the frequency of G in H (selection, drift, and mutation must all be negligible). In practice, $|p_1 - p_2|$ cannot be very small and the time period must be relatively short, say under 500-1000 years. If μ^* is defined as the proportion of genes in H that derive from P_1 , then, by definition, the expected frequency of G in a sample from H , p_h , is $\mu p_1 + (1 - \mu)p_2$. An estimate of μ , M , is given by $(p_h - p_2)/(p_1 - p_2)$.

The genetically most studied human hybrid population is probably the American Negro. Estimates of μ (defining P_1 as "the" Caucasian parental population and P_2 as "the" African parental population) for this hybrid population have recently been reviewed and extended (1). Using the *Fy^a* allele of the Duffy blood-group system, M was 0.2195 ± 0.0093 for a sample of 3146 Negroes from the Oakland, California area. By various criteria, including small standard error, this estimate appears to be the best presently available for U.S. Negroes living outside the Deep South (1).

INDIVIDUAL PROPORTION AND ITS VARIANCE

Each individual in H has his own individual proportion of genes deriving from P_1 . This proportion, for the i th individual, may be termed μ_i . μ is the mean of the μ_i values in H . μ_i is not estimable from single-locus genetic data; all studies to date have presented M values estimating μ . In any hybrid population, μ_i can vary between 0 and 1; M from a single sample gives no information about the distribution of μ_i . It is of interest to enquire whether the population variance of μ_i , σ_i^2 , measuring variation about μ , can be estimated when the μ_i are not estimated. An attempt to estimate σ_i^2 in the Oakland population mentioned above has been recently reported (2). As will be shown, the estimate appears to be incorrect. This estimate used three published estimates (1) of μ , each of them based on either the total sample of 3146 Negroes (estimating with Duffy phenotype data and with ABO phenotype data) or on a sample of 260 Negroes (estimating with Gm phenotype data). Since each of the three loci can supply an independent M value, variability among these estimates would seem to provide information for estimating σ_i^2 . It is important to note, however, that the relative amounts of information about μ provided by these three estimates are very different. Estimating information as the reciprocal of the variance of the estimate, the Duffy estimate is found to provide 16 times as much information as the Gm estimate and 22 times as much as the ABO estimate (1). Clearly the best estimate of σ_i^2 would come from Duffy data alone, by the use of two or more independent samples. This approach will now be tested.

The only useful data available consist of estimates, $M_1, M_2, \dots, M_j, \dots, M_k$ of μ for k different subsamples, e.g., males and females, of the total Negro sample. We ask whether the variation between subsamples in M value can provide any useful information about σ_i^2 . The only relation between these subsample estimates and σ_i^2 appears to be the following: The M_j can be used to calculate (in the analysis of variance sense) a between-subsample mean-square, B , whose expected value is an upper limit for σ_i^2 . If there are real differences among the M_j , there will be a corresponding (but unknown) between-subsample variance, σ_b^2 . Since σ_i^2 , from the analysis of variance point of view, can be considered here as a within-sample (or subsample) variance, we can express the expected value of B , $E(B)$, in terms of σ_i^2 and σ_b^2 :

$$E(B) = \sigma_i^2 + c\sigma_b^2, \quad [1]$$

where c is approximately, or exactly, equal to the mean subsample size. B is an upper limit to σ_i^2 and, if we can show that σ_b^2 is small, B might be a good approximation to σ_i^2 .

* Here and elsewhere, Greek letters represent the true, but unknown, population parameters. Roman letters represent calculated or observed values.

TABLE 1. *Duffy blood-group phenotypes [Fy(a+) or Fy(a-)] and estimates of μ (M) for various subsamples of 3146 Oakland Negroes. Estimates of between-subsample mean-square (B) and its standard error for three subdivisions.**

Subdivided by	Subsample	Number		$M \pm SE^\dagger$	$B \pm SE$
		Fy (a+)	Fy (a-)		
Sex	Males	290	1301	0.2233 ± 0.0128	0.0530 ± 0.2225
	Females	274	1281	0.2155 ± 0.0128	
MN blood-group	M	139	601	0.2305 ± 0.0188	0.0795 ± 0.1929
	MN	276	1303	0.2137 ± 0.0126	
	N	149	678	0.2206 ± 0.0175	
ABO blood-group	A	140	726	0.1969 ± 0.0162	0.5681 ± 0.4342
	B	128	477	0.2615 ± 0.0221	
	O	271	1269	0.2152 ± 0.0128	
	AB	25	110	0.2271 ± 0.0433	

Unpublished data from a study reported in ref. 1; West European Caucasians were used.

* Calculations done with 16-digit accuracy, but rounded to four decimal places.

† The estimate for the total sample is 0.2195 ± 0.0093 .

No other relation between the M_j and σ_1^2 is apparent. Denoting the size and M estimate of the total sample by N_T and M_T , and the corresponding quantities for the j th subsample by N_j and M_j , we can calculate the value of B for k subsamples from

$$B = \left\{ \sum_{j=1}^k N_j M_j^2 - N_T M_T^2 \right\} / (k - 1). \quad [2]$$

We must also estimate the variance of B , $V(B)$, in order to judge how much information about B , and therefore about σ_1^2 , this procedure yields. We may use the method of differentials, commonly used for finding variances of functions (3), to derive the approximate variance of B . If the calculated variances of the M_j are denoted by $V(M_j)$ (estimated by the formula given in ref. 1),

$$V(B) \doteq [4/(k - 1)^2] \sum_{j=1}^k N_j^2 (M_j - M_T)^2 V(M_j). \quad [3]$$

If $V(B)$ is large, the estimate of B may not be meaningful.

The sample of 3146 Oakland Negroes can be subdivided in various ways (Reed, T. E., unpublished data), and the Duffy distribution in these subsamples can be used with Eqs. [2] and [3]. Table 1 gives the Duffy blood-group phenotypes and corresponding M_j for each subsample produced by subdivision of this one sample according to three independent criteria: sex, MN blood-group phenotype, and ABO blood-group phenotype. The proportions of Fy(a+) individuals and the values of M_j do not vary significantly among the subsamples of any of three sets. An upper limit of σ_1^2 should consequently be well estimated by B . The estimates of B and its standard error, however, show that this approach is not useful. The standard errors of B are all very large, indicating that there is little information about B or about σ_1^2 . All three estimates are compatible with a true B value anywhere in the range from zero to 0.4 or greater, and therefore the estimates of σ_1^2 may also have this range. There is almost no information here about σ_1^2 .

DISCUSSION

There are no known associations between the Duffy system and sex, MN phenotype, or ABO phenotype in Caucasians (4, 5) or in Negroes (4), and the data of Table 1 give no

evidence of significant heterogeneity among the M values. The between-subsample mean-square estimates should therefore be reasonable estimates of an upper limit of σ_1^2 , the within-subsample variance. Since none of the three estimates is usable (they are compatible with a σ_1^2 of up to 0.4 and a standard deviation of μ_1 up to 0.6 or more), it appears that even with Duffy data from a large sample, σ_1^2 cannot be usefully estimated from group data. The one previous estimate (2) of σ_1^2 in the Oakland Negro population, estimated on the basis of separate M estimates from Duffy, ABO, and Gm data by an obscure method, is 0.047 ± 0.03 . This estimate appears to be incorrect because the standard error presented is far too small (especially considering that it is derived in part from the relatively uninformative ABO and Gm M estimates) [see Note Added in Proof]. Shockley does not clearly distinguish between μ_1 and μ or between σ_1^2 or $V(M)$, even though he is trying to estimate σ_1^2 from group data (2). A further problem is that natural selection may produce real (but unrecognized) differences among the M values from different loci, and consequently inflate the between-subsample variance.

It seems that there may be no substitute for the individual estimation of μ_1 if one wishes to determine the distribution of μ_1 about its mean, μ . However, for the American Negro at least, accurate individual estimation of μ_1 may not be possible soon. There are very few genes whose frequencies (i) throughout the former slaving areas of Africa are at least moderately well known and (ii) are usefully different from Caucasian frequencies (1).

NOTE ADDED IN PROOF

It should be noted that, contrary to Shockley (2), the three M estimates for Oakland (1) do not demonstrate "discordance" since they do not differ significantly among themselves.

I thank Profs. William J. Schull and Charles F. Sing for discussions and comments.

1. Reed, T. E., *Science*, **165**, 762 (1969).
2. Shockley, W. E., *Proc. Nat. Acad. Sci. USA*, **68**, 1390A (1971).
3. Kendall, M. G., *The Advanced Theory of Statistics* (Charles Griffin & Co., London, 1947), Vol. 1, 3rd ed., p. 208.
4. Reed, T. E., *Amer. J. Hum. Genet.*, **20**, 142 (1968).
5. Shreffler, D. C., C. F. Sing, J. V. Neel, H. Gershowitz, and J. A. Napier, *Amer. J. Hum. Genet.*, **23**, 150 (1971):