DISTRIBUTION OF GENE FREQUENCY AS A TEST OF THE THEORY OF THE SELECTIVE NEUTRALITY OF POLYMORPHISMS[1,2]

R. C. LEWONTIN AND JESSE KRAKAUER

Department of Theoretical Biology and Department of Biology, University of Chicago, Chicago, Illinois 60637

Manuscript received February 14, 1972
Revised copy received January 16, 1973
Transmitted by T. PROUT

ABSTRACT

The variation in gene frequency among populations or between generations within a population is a result of breeding structure and selection. But breeding structure should affect all loci and alleles in the same way. If there is significant heterogeneity between loci in their apparent inbreeding coefficients $F = s_p^2/\bar{p}(1-\bar{p})$, this heterogeneity may be taken as evidence for selection. We have given the statistical properties of F and shown how tests of heterogeneity can be made. Using data from human populations we have shown highly significant heterogeneity in F values for human polymorphic genes over the world, thus demonstrating that a significant fraction of human polymorphisms owe their current gene frequencies to the action of natural selection. We have also applied the method to temporal variation within a population for data on *Dacus oleae* and have found no significant evidence of selection.

THE discovery of vast amounts of polymorphism in sexually reproducing animals and plants since the first reports by HARRIS, by HUBBY and LEWONTIN and by JOHNSON *et al.* in 1966, has given a considerable impetus to the problem of distinguishing between natural selection and essentially non-selective processes, such as restriction of population size, recurrent mutation and migration, on the determination of genetic variation. A synthetic view does not allow an "either-or" approach to this problem, but admits that gene frequency distributions are the result of the interaction of selective and non-selective forces. Yet even in such a view there exists the strong possibility that selection coefficients are of such a magnitude for most genes that variations in time and space and between loci in the array of allele frequencies can be explained in large part without reference to variation in selection coefficients. That is, in the sense of the analysis of variance, the *main effect* of selection in determining gene frequencies is small.

One of the main problems of assessing the role of natural selection lies in the

[1] This paper is dedicated to the memory of Ken-ichi Kojima, who was killed in the full flood of his life on November 14, 1971. He was an imaginative scientist who made many important contributions to evolutionary genetics, and he was a dear friend.

[2] This work was supported by Contract AT(11-1)-1437 of the United States Atomic Energy Commission.

probable size of selection coefficients. Even the most sanguine "selectionist" will not claim that selection coefficients in excess of a few percent are the rule. But the power to detect selection of this magnitude by observing changes in gene frequency, or differences in components of fitness between genotypes, is very low for sample sizes within the range of practicality (see WILSON 1970, for a general theoretical treatment of the estimation problem, and YAMAZAKI 1971, for a specific example in an experimental context). The alternative is then to use information about the spatio-temporal distribution of allelic frequencies on the assumption that these frequencies are in a steady-state distribution resulting from the interaction of the various forces, and so attempt, from the distribution, to estimate the forces. The difficulty with *this* procedure is that various parameters enter in a confounded way into the determination of the observations. For example, the steady-state distribution of allelic frequencies over populations depends critically on terms such as Ns and $N(\mu+m)$ where s is the intensity of selection, μ the mutation rate to one of the alleles, m the migration rate from neighboring populations (in the extremely simplified model of island populations) and N, the effective population size, a number only loosely related to census size. Not only is selection confounded with the other parameters, but those parameters are extremely difficult to measure in practice. Measurements of either migration rates or effective population sizes are, both practical and conceptual reasons, virtually impossible—especially if populations do not conform to simple "island" structure, and intuitive notions about whether N is "very large" or m is "small enough to be ignored" are useless, especially in view of the fact that these two parameters appear as their product, Nm, in the theory! So, any observed distribution of gene frequencies over space or time, if considered to be in a steady state can be explained by a suitable choice of N, m and μ, with s being made arbitrarily small, the more so because the number of populations or time points observed is never really very large and the distribution is poorly known. As an example, PRAKASH, LEWONTIN and HUBBY (1969) showed that the allelic frequencies at 24 loci, 11 of which were polymorphic, were very similar in three widely separated populations of *Drosophila pseudoobscura*, and recent, as yet unpublished, work from the same laboratory extends these similarities to a half dozen more such populations. One explanation of the striking similarity of allelic frequencies among populations 2000 miles apart, even for loci with five to eight alleles segregating, is that these frequencies are in stable equilibrium held by common selective forces in all populations. An equally good fit to the data, however, can be made by the hypothesis that there is no selection and that effective migration over the range of the species is of the order of one individual per generation between neighboring populations (WRIGHT 1951). It is impossible to rule out, by direct evidence, migration rates of that order. (Even allelism of lethals between populations is totally insensitive to small migration rates).

What is required is some method of detecting selection which will cancel out the effects of the breeding structure. In principle, such a method is possible because there is one feature of migration, genetic drift and inbreeding that is quite different from selection. *While natural selection will operate differently for each*

locus and each allele at a locus, the effect of breeding structure is uniform over all loci and all alleles. Inbreeding affects all genes simultaneously and to the same average degree, as does sampling variation and migration. In the absence of selection, the steady-state variation in frequency of an allele at a locus from population to population, or from time to time, is entirely a reflection of the breeding structure of the species and, in fact, the variation of frequency can be used to estimate a parameter F_e, the "effective inbreeding coefficient" for the collection of populations. The effective inbreeding coefficient is a kind of fictitious equivalent to the inbreeding coefficient that would arise after some fixed number of generations in a model of totally isolated populations with no migration between them, perfect panmixia within them, each of a constant size N, that would produce the amount of genetic variation observed among the real populations. The fact that the real populations do exchange genes, do fluctuate in size and are not perfectly panmictic is irrelevant, because these actual conditions are summarized in the "effective" parameter F_e, which has no interest in itself. The estimate is simply

$$\hat{F}_e = \frac{s^2_p}{\bar{p}(1-\bar{p})} \tag{1}$$

where $\hat{F}_e$ = estimate of effective inbreeding,
s^2_p = variance in the frequency of one of two alternate alleles from population to population, and
$\bar{p}$ = mean frequency of the allele over the ensemble of populations.

Now suppose we carry out the estimate in using a number of different loci, each of which will have its own values of s^2_p and $\bar{p}$. If all these loci are selectively neutral, they will all estimate the same F_e since the estimations apply to all genes in the ensemble of populations. Then the collection of $\hat{F}_e$ values actually calculated will be an estimation of a true F_e and there should be no significant heterogeneity among them. On the other hand, if some or all of the loci are under selection, the various $\hat{F}_e$'s will not be estimates of the same F_e because they will be distorted by selection. For the selection loci the s^2_p values and therefore the estimated $\hat{F}_e$ will be too large if selection is different in different populations, while if there is a selection in common, the variance among populations will be too small. Even if all the loci are under some selection, the $\hat{F}_e$ will not be drawn from the same distribution unless selection is acting identically over all alleles. This idea was first used, as far as we know, by CAVALLI-SFORZA (1966), who calculated $\hat{F}_e$ values over a wide range of human groups encompassing the diversity of the species, for fifteen different gene frequencies, representing nine loci. The values of $\hat{F}_e$ from CAVALLI-SFORZA are reproduced as Table 1 of this paper. The considerable variation in $\hat{F}_e$ values from .029 for the Kell blood group locus, to .382 for the R_0 allele of the Rh locus seemed to CAVALLI-SFORZA as too great to explain from sampling error alone, but that impression could not be tested in the absence of any sampling theory. We shall return to the data of Table 1 later.

In principle, then, a test of the homogeneity of $\hat{F}_e$ estimates derived from different loci in steady state will be a test of the homogeneity of selection coefficients across loci—which, in effect, is a test for selection. In practice, the problem is to derive the distribution of $\hat{F}_e$ under the assumption of all being drawn from the

TABLE 1

$\hat{F}$ *values for world wide distribution of human polymorphisms.* n *is the number of groups sampled. Data from* CAVALLI-SFORZA (1966)

System	Allele	n	$\hat{F} = s^2_p/\bar{p}\bar{q}$
ABO	*A*	125	.070
	B	125	.055
	O	125	.081
MN	*MS*	45	.071
	Ns	45	.094
Rhesus	R_0	75	.382
	R_1	75	.297
	R_2	75	.141
	r	75	.172
Duffy	*Fy*	62	.358
Diego	*Di*	64	.093
Kell	*k*	64	.029
Haptoglobin	Hp^1	60	.096
Gm	Gm^a	25	.226
Gc	Gc^1	42	.051
			$\bar{F} = .148$
			$s^2_F = .00741$

same universe and then to use this sampling distribution to test the significance of the difference between two or more $\hat{F}_e$ values from different loci.

TWO MODELS

1. *Variation in space:* Let us postulate a locus with two alleles. *A* and *a* segregating (or fixed) in a large number of populations. The populations are defined simply as sampling units and we know nothing about their actual breeding structure or their degree of isolation from each other. Let p_i be the frequency of allele *A* in the i^{th} population. Let us suppose we choose a random subset of n populations from the entire ensemble and determine the gene frequency p_i in each. We will further suppose that the determination of p_i in each population is based upon sufficient sample size so that the within-population sampling variance is negligible compared with variance among populations. That is, we know each p_i to a first approximation. If we then take the sample variances among the n values of p_i, and the mean, $\bar{p}$, we can calculate one estimate

$$\hat{F}_{e_j} = \frac{(s^2_p)_j}{(\bar{p}\bar{q})_j}.$$

From another subset of n randomly-chosen populations we can calculate another value of $\hat{F}_e$ and so on. Then the values of $\hat{F}_e$ will have a distribution that depends upon the true variance of the p_i, σ^2_p and the true mean, μ_p. Unfortunately it depends upon not just these two first moments, but also on the entire distribution of p_i. This distribution may be of any of a number of shapes: unimodal falling

off on both sides, J-shaped, U-shaped or rectangular, depending upon the underlying parameters N, m, s, etc. which are unknown. Nor can we estimate the form of the distribution directly since we usually cannot sample enough populations to get a picture of it. In practice, of course, the sampling distribution of $\hat{F}$ might be rather insensitive to different shapes and, in the best of all possible worlds might depend only on the ratio $\sigma^2_p/\mu_p(1-\mu_p)$, i.e., on the true value of F. As we will show, for a given general functional form of the parent distribution of p_i, the sampling distribution of $\hat{F}$ does depend only on the ratio $\sigma^2/\mu_p(1-\mu_p)$, that is, on the true value of F; but when we change the form of the parent distribution, the sampling distribution changes non-trivially. This latter, unfortunate fact, leads us to consider an alternate procedure, originally conceived of by C. KRIMBAS and actually utilized by KRIMBAS and TSAKAS (1971).

2. *Variation in time:* If we observe the frequency of an allele in two successive generations in a population, and if that gene is not subject to natural selection, there will still be a change in the gene frequency, Δp, because of the finite size of the population. If there were a very large number of identical populations all starting with the same gene frequency, there would be a variation in gene frequency among the populations after one generation which would be related to an effective inbreeding coefficient by equation 1. We will denote the inbreeding coefficient that arises after a single generation of such drift by f to distinguish it from the result of many generations of the process, F_e. The change, Δp, within any single population can be used to estimate the variance among the populations and in fact

$$s^2 = (\Delta p)^2 \tag{2}$$

is an estimate of the variance with one degree of freedom. Then an estimate of f is, from (1) and (2)

$$\hat{f} = \frac{(\Delta p)^2}{p_0(1-p_0)} \tag{3}$$

where p_0 is the gene frequency in the initial generation. If this is done for many different genes in the same population, then each such estimate, under the hypothesis of no selection, estimates the same true f, and we may apply the same reasoning as for F. The advantage of f over $\hat{F}$, however, is that we know the underlying distribution of the p_i. It must be binomial[3] with mean p_0 and variance $\frac{p_0(1-p_0)}{2N_e}$, because it is a one-stage sample of size $2N_e$ from a population with value p_0. We are not, in this case, plagued with the problem of the unknown underlying gene frequency distribution that affects the sampling distribution of $\hat{F}$. As we shall see, this is a particularly felicitous choice of an underlying distribution.

THE SAMPLING DISTRIBUTION OF $\hat{F}_e$

We wish to find the distribution of the statistic

$$\hat{F}_e = \frac{s^2_p}{\bar{p}(1-\bar{q})}$$

[3] Strictly speaking, only if the variance in offspring number is Poisson, but close enough in any case.

TABLE 2

Statistical of the empirical distributions of $\hat{F}$ *for different underlying distributions of* p. n *is the sample size on which each* $\hat{F}$ *value is calculated,* $\bar{F}$ *the mean,* s^2_F *the variance and* $k = (n-1)\ s^2/F^2$

Distribution	n	$\bar{F}$	s^2_F	$k = (n-1)\ s^2/F^2$
Binomial:	20	.048	2.3×10^{-4}	1.9965
$(.5 + .5)^{21}$	100	.048	4.4×10^{-5}	1.9097
	500	.048	8.9×10^{-6}	1.9306
Binomial:	20	.024	6.2×10^{-5}	2.1528
$(.5 + .5)^{41}$	100	.025	1.2×10^{-5}	1.9200
	500	.024	2.5×10^{-6}	2.1739
Binomial:	20	.025	3.12×10^{-5}	2.083
$(.1 + .9)^{39}$	100	.025	6.25×10^{-6}	1.9200
	500	.026	2.5×10^{-6}	1.8519
				$\bar{k} = 1.9932$
Uniform	20	.34	5.2×10^{-3}	.8990
	20	.34	4.9×10^{-3}	.8378
	100	.34	$.96 \times 10^{-3}$	.8304
	100	.33	1.0×10^{-3}	.9183
	500	.33	2.2×10^{-4}	1.0092
	500	.33	2.1×10^{-4}	.9633
				$\bar{k} = .9113$
U-shaped	20	.536	7.0×10^{-3}	.4875
$p^{-1}\ (1-p)^{-1}$	20	.534	6.7×10^{-3}	.4698
	100	.525	1.2×10^{-3}	.4348
to 2 decimal places	100	.527	1.2×10^{-3}	.4317
	500	.523	2.5×10^{-4}	.4570
	500	.524	2.5×10^{-4}	.4554
				$\bar{k} = .4560$
U-shaped	20	.595	1.77×10^{-2}	.4068
to 5 decimal places	100	.587	3.39×10^{-3}	.3540
	500	.578	6.68×10^{-4}	.4042
				$\bar{k} = .3883$

given that p has some specified distribution among populations and that s^2_p and $\bar{p}$ have been calculated from a sample of n populations. If the p_i were very close to being normally distributed, and if n were large enough so that $\bar{p}$ were very close to the true mean of p, then $(n-1)\hat{F}/\bar{F}$ would be very close to a chi-square distribution with $n-1$ degrees of freedom. In actual fact, however, $\bar{p}$ will vary around the true mean and the distribution of p_i will never really be normal, so we do not know, *a priori*, how useful the chi-square distribution may be. The problems of finding the distribution of $\hat{F}$ analytically seem to us formidable, in general, and even if found would certainly require numerical tabulation; so we have used Monte Carlo simulation to produce the sampling distribution for a number of cases. A hypothetical distribution of p is specified in the form of a table of p values

and the cumulative probabilities of observing a p less than or equal to that value. A random number from a uniform distribution on the interval [0, 1], is then generated and used to pick out a p value. This technique will choose p values in proportion to the probabilities given by the postulated distribution. A sample of n such p values is chosen and from this sample a single value of $\hat{F}$ is calculated according to formula 1. The sampling cycle is repeated 5000 times to produce 5000 $\hat{F}$ values which then represent an empirically derived sample distribution of $\hat{F}$, given the hypothetical underlying distribution of p.

Table 2 shows the results of the simulations for underlying distribution of p that are binomial, flat, or U-shaped, corresponding to common distributions expected for steady state gene frequency distributions without selection. Both a symmetrical and an asymmetrical binomial distribution and two different variances ($N = 21$ and $N = 41$) were tested. Two replicates for each of the uniform and U-shaped distributions are given, to show the reliability of the empirical statistics.

For the binomial distributions we see that the mean value of $\hat{F}$ is unaffected by the asymmetry so that the normalizing effect of dividing the variance in gene frequency by $\bar{p}(1-\bar{p})$ does indeed work. Moreover, the mean $\hat{F}$ turns out to be very close to $1/21 = .0476$ and $1/41 = .0244$, that is to $1/N$, which is exactly what the expected value would be if the denominator of F,$\bar{p}(1-\bar{p})$, had no sampling variance. In fact $\bar{F}$ turns out to be slightly greater than $1/N$ as a result of the small sampling variance of the denominator. The mean value for the uniform distribution also turns out to be very close to the expected value of 4/12, which would be the case if the denominator had no variance. For the U-shaped distribution it is more difficult to compare the observed mean with the ideal since the expectation depends critically on what convention is made concerning the terminal classes. Obviously p=0 and p=1 must be excluded or the mean will be infinite. The more finely-divided the underlying discrete distribution is, the smaller the value of p in the subterminal classes and so the larger the value of F since $\bar{p}$ appears in the denominator. Table 2 shows that an increase of the fineness of subdivision from 2 to 5 decimal places of the U-shaped distribution classes increases the value of $\hat{F}$ by about 15%.

If the distribution of F is in some sense invariant under changes in the parameters of the underlying distribution, there should be a relation between $\bar{F}$ and s^2_F. In any case, s^2_F shoud be inversely proportional to $(n-1)$, the degrees of freedom of $\hat{F}$, and in addition if the distribution of $\hat{F}$ is invariant, we might guess that the variance of $\hat{F}$ will be proportional to $\bar{F}^2$. That is

$$s^2_F = \frac{k\ \bar{F}^2}{(n-1)}\ .$$

In the last column of Table 2 we have calculated $k = (n-1)\ s^2/\bar{F}^2$ for each run. We see that for each form of distribution there is a characteristic k, and that for the binomial distributions, k=2, irrespective of the parameters of the binomial. This value for k is not coincidental. We remarked before that we expected that F might have a distribution related to x^2 when the underlying distribution of p is

roughly normal. If p is binomial, then as noted by WORKMAN and NISWANDER (1970), the *weighted* sum of squares of the p_i divided by $\bar{p}\ (1-\bar{p})$ is algebraically identical to the usual homogeneity statistic calculated from $2 \times n$ tables. In particular χ^2 is distributed with mean $n-1$ and variance $2(n-1)$ where n−1 is the number of degrees of freedom. Now $\hat{F}$ has a mean $\bar{F}$, so $\frac{(n-1)\hat{F}}{\bar{F}}$ will also have a mean of $(n-1)$. What will its variance be? It will be

$$\frac{(n-1)^2 s^2_{\hat{F}}}{\bar{F}^2}.$$

But we have shown in Table 2 that

$$\frac{(n-1)s^2_{\hat{F}}}{\bar{F}^2} = 2$$

so the variance of $\frac{(n-1)\hat{F}}{\bar{F}}$ will be $2\ (n-1)$, the same as the chi-square distribution. It would appear likely then that $\frac{(n-1)\hat{F}}{F}$ is distributed as χ^2 with $n-1$ degrees of freedom. To make a closer check, the empirical cumulative distribution of $\hat{F}$ is plotted in Figure 1 for several of the cases given in Table 2. The coordinates of the abscissa are so arranged that a normally distributed variable will appear as a straight line with a slope equal to its standard deviation. Plotted on this graph are the $\hat{F}$ distributions based on underlying binomial, uniform and U-shaped distributions. In each case the ordinate is in units of $\frac{(n-1)\hat{F}}{\bar{F}}$. For the U-shaped and flat underlying distributions, the distribution of $\frac{(n-1)\hat{F}}{F}$ are very close to normal with mean $n-1$ and standard deviation $\sqrt{k(n-1)}$ were k is given in Table 2. On the other hand, the $\hat{F}$ distribution based on the binomial distribution deviates significantly from the normal, being concave upwards. This concavity means that the distribution is skewed to the left, in excellent agreement with the χ^2_{19} distribution shown by the curved line. It should be noted that this curve is in no way "fitted" to the data. It is simply the chi-square distribution with 19 degrees of freedom.

We may summarize the results of these empirical distribution studies as follows. If the underlying distribution of gene frequencies across populations is binomial or normal, then $\frac{(n-1)\hat{F}}{\bar{F}}$ is distributed as χ^2 with $n-1$ degrees of freedom were n is the number of populations sampled. If the underlying distribution of gene frequencies is much more dispersed than the binomial, then $\frac{(n-1)\hat{F}}{F}$ is normally distributed with mean $(n-1)$ and a variance between a quarter and a half of that for the binomial case. This latter point—that greater dispersion of the underlying gene frequencies *reduces* the variance of $\hat{F}$—will be extremely important for testing hypotheses about the homogeneity of F values.

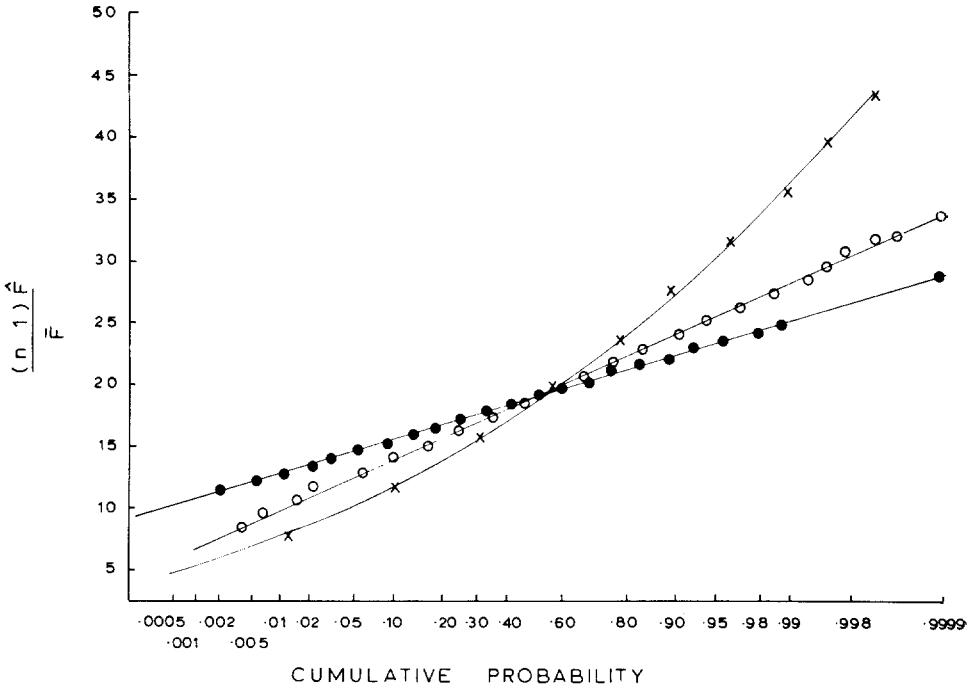

FIGURE 1.—Empirical cumulative distributions of $\hat{F}$ given various underlying distributions of p. Abscissa is in units of probability on a scale arranged to produce a straight line if the distribution is normal. Ordinate in units of $(n-1)\ \hat{F}/\bar{F}$. Crosses: binomial; solid circles: uniform; open circles: U-shaped.

APPLICATION TO SPATIAL VARIATION

In 1966 ARENDS *et al.* published the distribution of 22 allelic frequencies belonging to 15 different loci, distributed over 10 villages of the Yanomama tribe of Indians in the Orinoco Basin. These villages are partially isolated, but do exchange genes, and some villages have been formed from others by a process of budding. Figure 2 shows the distribution of allelic frequencies among the villages for the alleles investigated.

Table 3 gives the $\hat{F}$ value for each allele together with the number of villages over which it has been estimated. The $\hat{F}$ values seem to fall in two groups, nine of them being .036 or less and eleven being greater than .072, with only two being in between these groups. We wish to test the hypothesis that this apparent heterogeneity of $\hat{F}$ values is real. The $\hat{F}$ values observed do not need to be corrected for sampling error within villages since, although the numbers in each village are small, they are nearly complete censuses rather than samples.

Based on our Monte Carlo results we could take two approaches. One would be to test the goodness of fit of the observed distribution of $\hat{F}$ values to one of the sampling distributions. Figure 2 shows that the underlying distribution of p for all the loci are much less disperse than uniform and so fall in the category of the

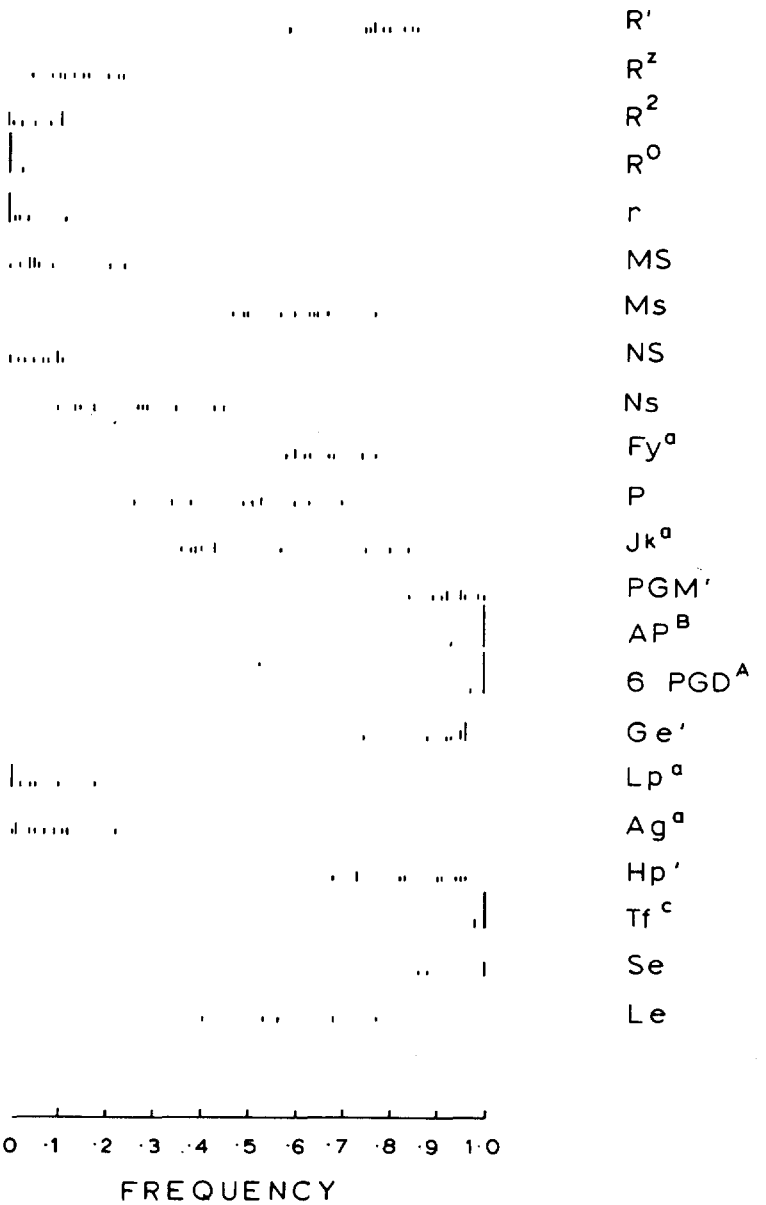

FIGURE 2.—Distribution of allelic frequencies for genetic systems of Table 3. Data from ARENDS *et al.* (1966).

unimodal binomial distribution. We might then test the goodness of fit of the observed distribution of $\hat{F}$ values to a χ^2 distribution with nine degrees of freedom (ten villages). Table 4 shows a comparison between the observed distribution of the 22 $\hat{F}$ values in Table 3 and a χ^2 distribution with nine degrees of freedom (corrected for the observed mean F). Classes were constructed to be a half stand-

TABLE 3

$\hat{F}$ values for different allelic distributions among ten villages of Yanomama Indians n *is the number of villages for each $\hat{F}$. Calculated from data of* AREND*S et al. (1966)*

Locus	Allele	n	$\hat{F}$
Rh	R_1	10	.03329
	R	10	.03002
	R^2	10	.04967
	R^0	10	.03069
	r	10	.07704
MNS	MS	10	.08750
	Ms	10	.03620
	NS	10	.03235
	Ns	10	.07593
Duffy	Fy^a	9	.02067
P	P	10	.07398
Kidd	Jk^a	10	.14594
Phosphoglucomutase	PGM^1	10	.03626
Acid phosphatase	. .	10	.07496
Group component	Gc^1	10	.07326
Lp	Lp^a	10	.09414
Ag	Ag^a	10	.06719
Lewis	Le	5	.08371
Transferrin	Tf^c	10	.01821
ABH-secretor	Se	5	.10417
Haptoglobins	Hp^1	10	.08226
6-phosphogluconic dehydrogenase	A	10	.03058
			$\bar{F} = .06175$
			$s^2_{\hat{F}} = .001052$

TABLE 4

Comparison of the distribution of $\hat{F}$ values of Table 3 and the theoretical χ^2 distribution

F	Observed	Expected	χ^2
.01372–.02744	2	1.98	.00
.02745–.04116	7	3.73	2.84
.04117–.05488	1	4.62	3.62
.05489–.06860	1	3.95	2.20
.06861–.08232	6	3.01	2.97
.08233–.09574	3	1.97	1.58
.09575–.10946	1	1.21	.20
> .10947	1	1.53	
Total	22	22.00	13.41

P = .04

ard deviation of the χ^2 distribution wide, centering on the observed mean F of .06175. The test for goodness of a fit between the observed distribution and the theoretical gave a $\chi^2 = 13.41$ with five degrees of freedom corresponding to $P = .02$, so we judge the observed distribution of $\hat{F}$ to be significantly different from the theoretical sampling distribution. This difference results from the greater heterogeneity of the observed $\hat{F}$, as evidenced in the two modes and is taken as evidence for selection on some of the genes. Of course we cannot tell whether the lower mode is the non-selected group, with diversifying selection accounting for the upper mode, or whether the lower mode is evidence for a common heterotic selection acting over all villages.

A second test of heterogeneity would be the comparison of the observed variance of $\hat{F}$ with the theoretical variance. As we have seen, the theoretical variance of $\hat{F}$ is given by

$$\sigma^2 = \frac{k\bar{F}^2}{(n-1)}$$

with k=2 when the underlying distribution of p is binomial. Applying this rule to the data of Table 3 we have

$$\sigma^2 = .0007624 .$$

We can test whether the observed variance $s^2_F = .001052$ is significantly larger by the ratio $s^2_F/\sigma^2 = 1.380$ which will be distributed as χ^2/d.f. In the analysis, there is a problem of what to do with the two multiple allelic systems, Rh and MNS. The variations in the various alleles at one locus are obviously correlated positively. That is, if one allele is very uniform over villages, other alleles must also tend to be uniform. To choose only one of the alleles arbitrarily, say the one closest to a frequency of .5, or to average F values over alleles at one locus would be to sacrifice information. The effect of using this correlated information will be to overestimate the power of any test because of spurious degrees of freedom. We will compensate for this by removing one degree of freedom for each multiple allelic locus. This would be exactly correct in the case where both alleles at a di-allelic locus were used, and must be nearly so here, since each degree of freedom corresponds to an added linear restriction on the ensemble of $\hat{F}$ values. In this particular case, the number of degrees of freedom is then $22 - 3 = 19$ and the probability of the ratio is $P = .12$, so the difference is not significant. The ratio of variances has then failed to detect the excess heterogeneity of $\hat{F}$ values shown from the goodness of fit test, which itself was significant only at the .02 level. Clearly the conclusions about heterogeneity of the $\hat{F}$ values is doubtful.

A fortunate circumstance makes it possible to perform a second test of this case. After these calculations were made, the Michigan group published more complete data on the Yanomama for 16 of the original 22 allelic frequencies plus a new system, Diego. The 10 original villages plus 27 new ones are included in this new study, and the authors have also calculated $\hat{F}$ values for the different systems (GERSHOWITZ *et al.* 1972; WEITKAMP *et al.* 1972; NEEL, WARD and MACCLUER 1972). This new study shows that the $\hat{F}$ values for the Yanomama are indeed homogeneous since:

1) The F values based on all 37 villages show no trace of bimodality.

2) There is no correlation between $\hat{F}$ values on the ten villages and $\hat{F}$ values from the larger sample.
3) The ratio of observed variance of $\hat{F}$ to expected variance in the larger study is .00413/.000216 = 1.55, which is referred to the χ^2/d.f. distribution with $17 - 3 = 14$ degrees of freedom and has a $P = .10$.

Apparently, then, if selection is operating on the genes in the study, the villages are insufficiently isolated or too recent historically, to make detection possible.

What of the data in Table 1 for the worldwide distribution of gene frequencies? Here the underlying distribution of various gene frequencies is more problematic, since some are close enough to binomial, but others with very high F values are much closer to uniform or even U-shaped distributions on a world-wide basis. However, since the expected variance of $\hat{F}$ is smaller (k is smaller) for these underlying distributions than for the binomial, we can perform a conservative test by using $k = 2.0$. The number of racial groups varies in Table 1 from 25 to 125, so we have used the harmonic mean, 60, for n. The theoretical variance of the $\hat{F}$ for this case is then

$$\sigma^2 = \frac{2.0\ (.148)^2}{59} = .000742$$

while the observed variance is .007416, so we have $s^2/\sigma^2 = 10$.

Again subtracting a degree of freedom for each multiple allelic loci, we compare this ratio with the distribution of χ^2/d.f. for 11 degrees of freedom and obtain a $P \ll .001$. Thus CAVALLI-SFORZA's suggestion that the $\hat{F}$ values are much too heterogeneous to be explained without selection is amply justified. In this case there is even some suggestion of which sort of selection has operated. On a world-wide basis, the most deviant allelic frequencies are generally found in groups that have small populations and are isolated culturally and genetically from other human groups. These include Eskimos, American Indians, Basques and Australian Aborigines, among others. There is no reason to suppose that natural selection will vary more between, say, American Indians and Australian Aborigines, both Stone Age peoples up until recently, than between, say, Europeans and Africans. Thus it is probably their isolation and small population size that has caused the divergence of these isolated groups. Then it is among the gene frequencies that have *not* diverged, those associated with *small F* values, that we should look for selection. In particular we should look for heterotic selection tending to retard divergence among the isolated groups with respect to these loci.

Both sets of human data differ in a significant way from the Monte Carlo sampling scheme on which the distribution of $\hat{F}$ is based. For the Yanomama and the worldwide distribution of gene frequencies, the same populations were characterized for all the loci, while in the Monte Carlo scheme, a new random sample of populations was chosen at each sampling. I am indebted to PROF. C. A. B. SMITH for pointing out this discrepancy to me. The repeated sampling of the same populations would be of no significance if there were no genetic correlations between populations, since each locus would be an independent random sample even though the populations were repeated. However, human populations are hierarchically-related with local populations within races being more closely

related than between races. Even the Yanomama villages are hierarchically related because some have budded off from others in recent times. To see the effect of repeated sampling in a hierarchically-arranged set of populations, we consider the most extreme possible case of two genetically differentiated races, but with all local populations within races identical. If we sample, say ten populations in each race and calculate $\hat{F}$ for many loci, we have really sampled only two different populations, since there is no component of variation within races. Then the appropriate value of $n-1$ in the denominator of the theoretical variance of F is only one instead of nineteen. However, compensating for this inflation of σ^2 by the reduction of n is a reduction in the theoretical variance because of the repeated samples from the same of set of populations. The underlying distribution of allele frequencies in this extreme case is bimodal, with half the populations having one allele frequency and the other half having a different frequency. But we are always choosing one "population" from each mode, whereas in a random sampling scheme we would by chance choose both populations from the same mode half the time. Thus, there are two opposite tendencies acting on the theoretical variance of $\hat{F}$ when repeated samples are taken from the same set of hierarchically-related populations, and we do not know the exact effect of these tendencies since we have not simulated this sampling scheme. For the Yanomama Indians, where the heterogeneity of $\hat{F}$ is on the borderline of significance, we will be made even more cautious. The observed variance for the world $\hat{F}$ values is, however, so much greater than the theoretical variance that it is most unlikely that the altered sampling scheme has much effect.

THE SAMPLING DISTRIBUTION OF $\hat{f}$

Because the use of the temporal inbreeding, $\hat{f}$, will often involve multiple allelic loci, we have simulated temporal sampling at a three-allele locus to determine the effect of the correlation between allelic frequencies. Three initial frequencies at the locus were specified and a random sample of N genes were taken to form a new population. The changes in allele frequency, Δp_1, Δp_2, and Δp_3, were then used to calculate $\hat{f}$ from the relation

$$\hat{f} = \frac{1}{3} \sum_{i=1}^{3} \frac{(\Delta p_i)^2}{p_i(1-p_i)}$$

This calculation was replicated 5000 times for each original distribution of p_i and for each population size N. For each replicate, the simulation was pushed further. Because in nature it is often impossible to get gene frequency data every generation, the sampling scheme was carried out for four generations to check that $\hat{f}$ after four generations is essentially four times $\hat{f}$ after a single generation, as it should be for low levels of inbreeding. Table 5 gives the statistics of these runs in the same form as for Table 2. It should be remembered that for all these runs, $n = 3$ since there are only three Δp's that enter into each $\hat{f}$ calculation.

We see from Table 5 that each of the $\hat{f}_1$ values is almost precisely $1/N$ and that the average $\hat{f}_4/\hat{f}_1 = 3.94$, which is exactly that predicted from the theoretical relation

TABLE 5

Statistics for the empirical distribution of $\hat{f}$ *for different initial gene frequency distributions.* N *is the number of genes in the breeding population* ($=2N_e$), $\hat{f}_1$ *is the single generation* $\hat{f}$ *value,* f_4 *is the four generation value*

Allele distribution	N		$\bar{f}$	s^2_f	$k=\frac{2s^2_f}{\bar{f}^2}$
.333, .333, .333	50	$\hat{f}_1$	$.19787 \times 10^{-1}$	$.38487 \times 10^{-3}$	1.966
		$\hat{f}_4$	$.75214 \times 10^{-1}$	$.48783 \times 10^{-2}$	1.724
		$\hat{f}_4/\hat{f}_1$	3.80		
	100	$\hat{f}_1$	$.97641 \times 10^{-2}$	$.89969 \times 10^{-4}$	1.887
		$\hat{f}_4$	$.38872 \times 10^{-1}$	$.14765 \times 10^{-2}$	1.954
		$\hat{f}_4/\hat{f}_1$	3.98		
	500	$\hat{f}_1$	$.20107 \times 10^{-2}$	$.38835 \times 10^{-5}$	1.921
		$\hat{f}_4$	$.81169 \times 10^{-2}$	$.66818 \times 10^{-4}$	2.028
		$\hat{f}_4/\hat{f}_1$	4.04		
.0625, .3125, .6250	100	$\hat{f}_1$	$.98860 \times 10^{-2}$	$.10154 \times 10^{-3}$	2.078
		$\hat{f}_4$	$.39282 \times 10^{-1}$	$.19256 \times 10^{-2}$	2.431
		$\hat{f}_4/\hat{f}_1$	3.97		
.0625, .4375, .4375	100	$\hat{f}_1$	$.10087 \times 10^{-1}$	$.10145 \times 10^{-3}$	1.994
		$\hat{f}_4$	$.38961 \times 10^{-1}$	$.15854 \times 10^{-2}$	2.089
		$\hat{f}_4/\hat{f}_1$	3.86		
.0625, .0625, .8750	100	$\hat{f}_1$	$.99167 \times 10^{-2}$	$.11082 \times 10^{-3}$	2.254
		$\hat{f}_4$	$.39162 \times 10^{-1}$	$.24201 \times 10^{-2}$	3.156
		$\hat{f}_4/\hat{f}_1$	3.99		
	Average	$\hat{f}_4/f_1$	$= 3.94$		$\bar{k} = 2.123$

$$f_n = 1 - (1 - f_1)^n,$$

which for $f_1 = .01$ gives $f_4 = .0394$. Thus, the average $\hat{f}$ over the three alleles at a tri-allelic locus is behaving exactly according to the theory for a single allele, as it should. The average k value is 2.123 and given the large variation from run to run and the lack of pattern of these variations, the agreement with a value of 2.0 for the χ^2 distribution is good. Indeed, nearly the whole excess over 2.0 is a result of the very extreme value in the last run in the table. Figure 3 shows the cumulative frequency distribution for several cases as compared with a χ^2 distribution (corrected for the mean f) with two degrees of freedom. There is a definite bias in the left half of the distribution with the empirical distribution rising somewhat faster than the χ^2. The right half, however, is in excellent agreement and thus is the part of the distribution that will be used for testing heterogeneity.

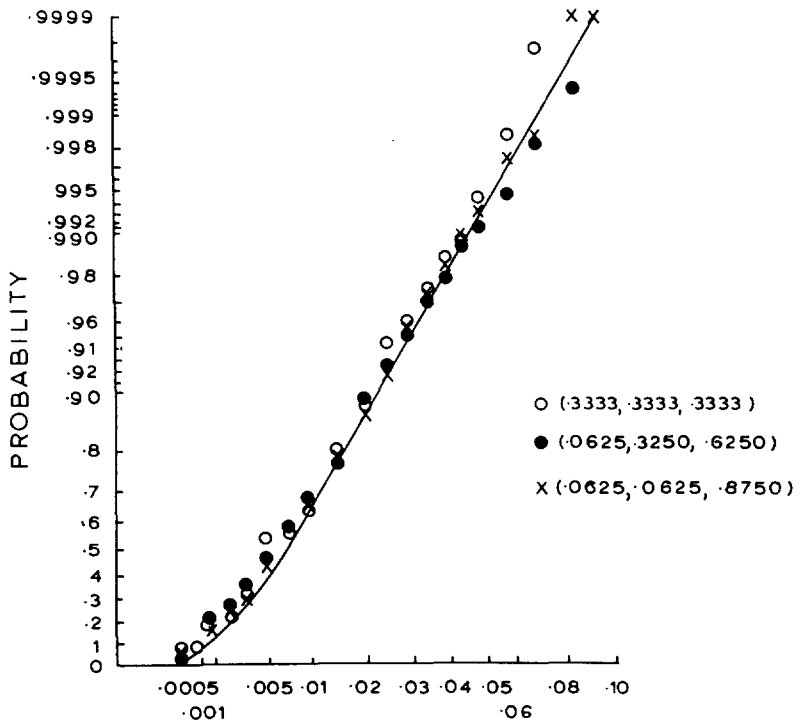

FIGURE 3.—Comparison between empirical cumulative f distributions and the χ^2 distribution with two degrees of freedom.

APPLICATION TO TEMPORAL VARIATION

KRIMBAS and TSAKAS (1971) studied two polymorphic loci controlling the synthesis of esterase enzymes in a natural population of the olive fruit fly, *Dacus oleae*. They took samples in three successive years, 1966, 1967 and 1968, and found eighteen alleles at the A locus and thirteen alleles at the B locus in this population. The alleles changed frequency during the course of the successive samples and they wished to test the hypothesis that the changes were a result of genetic drift. One way to do this would be to estimate f, the single-generation drift inbreeding, for each locus and then ask whether there was a significant difference between the two loci. This would have to be done separately for each pair of successive years since there is no reason to suppose that the effective population size will be the same over two full seasonal cycles. If the f value calculated from gene A and from gene B do not differ significantly, and if this is true both for the 1966–67 and 1967–68 comparisons, there is then strong evidence that selection is not involved.

KRIMBAS and TSAKAS estimated the average f for each locus by a three-step process. Using the relation given as our equation 3, they estimated the gross f over the n alleles at a locus as

$$\hat{f}_g = \frac{1}{n} \sum_{i=1}^{n} \frac{(\Delta p_i)^2}{p_i(1-p_i)}.$$

TABLE 6

Calculation of f *for genes A and B in two successive year comparisons, and the equivalent effective population sizes.* f_g *is the "gross"* $\hat{f}$, n *is the number of alleles in each comparison,* $\hat{f}$ *the final estimate of* f *and* $\hat{N}_e$ *the estimated effective population size. Adapted from* KRIMBAS and TSAKAS (1971)

Source	n	f_g	Sampling correction	$\hat{f}$	N_e
Gene A 1967–68	18	.0116938	.0028526	.0022103 ± .000758	226 ± 77.5
Gene B 1967–68	12	.0117140	.0030019	.0021924 ± .000935	228 ± 97.2
Gene A 1966–67	17	.0062634	.0026574	.0009015 ± .000319	1009 ± 356.8
Gene B 1966–67	13	.0056023	.0028455	.0006892 ± .000281	1451 ± 592.3

This gross f was then corrected to take account of the fact that some of the apparent Δp_i was their own sampling error. Letting M_1 and M_2 be their sample sizes in the two successive years, they calculated a net f

$$f_n = f_g - \left(\frac{1}{2M_1} + \frac{1}{2M_2}\right).$$

Finally they noted that *Dacus oleae* has four generations a year rather than one. At the very low rate of drift actually observed, the variance accumulated after four generations will be almost exactly four times that in one generation, so the final estimate of f, $\hat{f} = f_n/4$.

Table 6 shows the appropriate statistics for each gene in each year comparison. In addition to $\hat{f}$, the estimate of effective population size $\hat{N}_e$ is given, calculated from the reciprocal relation $\hat{f} = 1/(2N_e)$. The standard errors of $\hat{f}$ and $\hat{N}_e$ are obtained from our previous results. The underlying distribution of Δp is binomial as we have pointed out. Therefore the variance of $\hat{f}$, which is the average of n individual f values, each with one degree of freedom would be $\sigma^2_{\hat{f}} = 2\hat{f}^2/n$. However, since the n values are not independent of each other because they are the n alleles at one locus, we reduce the degrees of freedom by one, giving the standard error of $\hat{f}$ as with n alleles at a multiple allelic locus, from our Monte Carlo studies of f,

$$\text{S.E. } \hat{f} = \hat{f}\sqrt{\frac{2}{n-1}}. \tag{4}$$

Obviously genes A and B give the same value in 1967–68 and the difference between A and B in 1966–67, although larger, is not significantly so. For significance at the 5% level, the larger f value would need to be about four times the smaller. Since $N_e = \frac{1}{2f}$, then in large samples

$$\sigma^2_{N_e} \cong \frac{\sigma^2_f}{4f^4} = \frac{1}{2f^2(n-1)} = \frac{2N^2}{n-1}$$

so the standard error of N is of identical form to the standard error of f.

$$\text{S.E. } N_e = N_e\sqrt{\frac{2}{n-1}} \tag{5}$$

SENSITIVITY OF THE METHOD AS A TEST OF SELECTION

In general the use of variation in gene frequencies as a test of selection seems a reasonably powerful one. It was more than adequate to show that some kind of selective phenomena must have operated in the past for the world distribution of human polymorphism. It was marginally powerful enough to give evidence of selection among the Yanomanä tribes where much less differentiation has occurred.

For temporal variation we can determine about what level of selection could be detected by this method. We have shown that the standard error of f is very close to

$$\text{S.E.}\,\hat{f} = f\sqrt{\frac{2}{n-1}}.$$

Suppose we have two estimates of f, a larger, f_L, and a smaller, f_s, each based upon $m = n-1$ degrees of freedom, say $.n$ alleles at each locus, or m populations for each locus. If the larger is R times the smaller, then the standard error of the difference between them is

$$\text{S.E.}\,(f_L - f_s) = f_s\sqrt{\frac{2(R^2+1)}{m}}. \tag{6}$$

To be significantly different for a reasonable value of m, the difference $f_L - f_s$ must be twice its standard error. Then

$$f_L - f_s = f_s(R-1) = 2f_s\sqrt{\frac{2(R^2+1)}{m}} \tag{7}$$

and solving for R gives

$$R = \frac{1 + \sqrt{1 - \left(1 - \frac{8}{m}\right)^2}}{1 - \frac{8}{m}} \qquad (R \geqslant 1). \tag{8}$$

Note that the number of independent observations for each f must be greater than 8, or no difference will be significant. For a case like the KRIMBAS and TSAKAS work, we may let $m = 16$ and we find that

$$R \cong 3.7$$

so that the larger f (or the larger effective population size N_e) must be between three and four times greater than the smaller one for a significant difference. This may seem a great deal, but what does this amount to in terms of selection? The larger f will be the sum of a contribution from selection and from drift. That is

$$\frac{f_L}{f_s} = R = \frac{\hat{f}_{drift} + \frac{(\Delta p)^2_{sel}}{pq}}{\hat{f}_{drift}}. \tag{9}$$

But $(\Delta p)^2_{sel}$ is, for a semidominant gene, approximately $s(pq)^2$, and $f_{drift} = \frac{1}{2N_e}$. Substituting in (9) and solving, we get

$$s = \sqrt{\frac{R-1}{2pq\,N_e}}. \tag{10}$$

For the numerical case we are considering $R \cong 3.7$ and p for the most favorable case would be .5. This gives

$$s = 2.3/\sqrt{N}$$

as the level of selection that could be detected for a gene at intermediate frequency. For a population of about 500 this would mean a selection coefficient of .10. Note that the detectable s goes down only as the square root of $R-1$ so that $R-1$ would need to be 100 times smaller to detect s of .01. Clearly such one-generation tests of f are not adequate when selection coefficients are small. In such cases the tests need to be on $\hat{F}$, which is an equilibrium value, representing the accumulation of a large number of generations of random drift and selection.

What factors bias the test? Linkage disequilibrium (or any other form of historical correlation) between genes will do so, *reducing* the variance among the $\hat{F}$ values by roughly the amount of the squared correlation between the loci. Thus if two loci were completely correlated, irrespective of whether in coupling or repulsion, they would give identical values of $\hat{F}$. If a distribution of $\hat{F}$ values were found to be significantly too *homogeneous*, this would presumably be the explanation. Biases in the direction of increasing the heterogeneity of $\hat{F}$ without selection are difficult to conceive of. One obvious source, preferential migration according to genotype, must be considered as selection in the general sense of the neutrality hypothesis. That is, if different genotypes have different dispositions to migrate, then the genotypes are certainly not physiologically equivalent and, in addition, it is difficult to see how differential migration would not *ipso facto* result in differential fertility and viability patterns.

HOW IMPORTANT IS HISTORY?

A heterogeneous assemblage of F values can, in general, have another source besides selection. Suppose that a species has a number of ancient polymorphisms where F values reflect the accumulated history of the species' breeding structure. Suppose now that a new polymorphism arises as a new allele in some population fairly well isolated from the rest of the species. This allele may rise in frequency and even become fixed in its population by random drift before it has spread effectively to other parts of the species if the local population is small enough and isolated enough. So long as the allele is more or less confined to a single population among many, it will not give a high F value, but should the population proliferate into many new subpopulations and thus become the progenitor of a significant fraction of the species populations, an entire section of the species distribution will have a high frequency of an allele that is absent or virtually so, everywhere else until migration swamps the difference. Thus there will be a high F value for this newly-polymorphic locus, as compared with the more ancient polymorphisms. A high value from such a cause will be distinguishable by the fact that a number of related populations of the species have a high frequency of an allele that is absent or virtually absent elsewhere.

There will also be second mark of such an historical event. On the average it requires $4N$ generations for a new neutral mutant to go to fixation (or a high frequency) in a population of size N (KIMURA and OHTA 1969). But it requires

TABLE 7

Allele frequencies at the Fy *and* Rh *loci for a cross-section of human populations. Averages over many populations and studies are given simply as indications of frequency*

Populations	Fy^a	Genes R_0	R_1	R_2	r
Africans	.04	.60	.14	.07	.15
Europeans	.41	.07	.42	.16	.38
Basques	—	.08	.39	.06	.45
Lapps	.82	—	—	—	—
Hindi speakers	.73	.05	.64	.01	.28
South Asian aborigines	.71	—	—	—	—
Chinese	.90	0	.72	.19	.06
Japanese	.86	—	—	—	—
Malaysians	—	0	.92	.07	0
Amerinds	.75	0	.52	.48	0
Esquimo	1.00	—	—	—	—
Australian aborigines	1.00	.08	.56	.20	0

only between .06N and 2.8N generations on the average, for a polymorphism whose more common allele has a frequency between .99 and .5, respectively, to be fixed by drift (EWENS 1963). So, while the newly-arisen allele goes to high frequency in its original population, all the old polymorphisms in that population will be lost! Thus we should be able to detect high F values that are indicative of "new" polymorphisms rather than selection by first asking whether the F value results from some *related* populations' having a high frequency of an allele that is rare elsewhere in the species. If that is so, we could then ask whether those populations with the unusual allele are monomorphic for the polymorphisms that are common to the rest of the species.

Let us apply these criteria to the world distribution of F values in man. Table 1 shows six high F values corresponding to the Duffy blood group (.358), the four Rh alleles (.382, .297, .172, and .141) and the a specificity of Gm (.226). Table 7 shows sample gene frequencies for the Duffy and the Rh alleles. For Duffy, we see immediately that the frequencies of the Fy^a allele form a spectrum from .04 for black Africans to 1.00 for Australian aborigines, with Caucasians falling in the middle. There is certainly no pattern of a unique allele in related populations. The situation is more complex for the R alleles, but here again no case can be made for a group of related populations sharing an unusual allele as the cause of the high F values. On the other hand the Gm (a) locus does fit such a pattern, since all populations are fixed at 100% Gm^a except Caucasians who have more than 50% of the alternate allele, absent everywhere else. But this one does not fit our second criterion since Caucasians are highly polymorphic, and tend in fact to have intermediate frequencies of alleles at nearly every human polymorphic locus. Thus, there is no evidence that the heterogeneity in F values is the result of an "ancient" low F set of polymorphisms, and a "recent" high F group. It would seem that only selection can explain the heterogeneity.

LITERATURE CITED

ARENDS, T., C. BREWER, N. CHAGNON, M. L. GALLANGO, H. GERSHOWITZ, M. LAYRISSE, J. NEEL, D. SHREFFLER, R. TASHIAN and L WEITKAMP, 1967 Intra-tribal genetic differentiation among the Yanomama Indians of Southern Venezuela. Proc. Nat. Acad. Sci. **57**: 1252–1259.

CAVALLI-SFORZA, L., 1966 Population structure and human evolution. Proc. Roy. Soc. London Ser. B **164**: 362–379.

EWENS, W. J., 1963 The diffusion equation and a pseudo distribution in genetics. J. Roy. Statist. Soc. Ser. B **25**: 405–412.

GERSHOWITZ, H., M. LAYRISSE, Z. LAYRISSE, J. NEEL, N. CHAGNON and M. AYRES, 1972 Genetic structure of a tribal population, the Yanomama Indians. II. Eleven blood group systems and the ABH-Le secretion trait. Ann. Human Genet. **35**: 261–269.

KIMURA, M. and T. OHTA, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61**: 763–771.

KRIMBAS, C. B. and S. TSAKAS, 1971 The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control. Selection or drift? Evolution **25**: 454–462.

NEEL, J. V., R. H. WARD and J. A. MACCLUER, 1972 The genetic structure of a tribal population, the Yanomama Indians. VI. F-statistics (including a comparison with the Makiritare and Xavante). Ann. Human Gnet. **35** (in press).

PRAKASH, S., R. C. LEWONTIN and J. L. HUBBY, 1969 A molecular study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal, and isolated populations of *Drosophila pseudoobscura*. Genetics **61**: 841–858.

PROUT, T., 1969 The estimation of fitness from population data. Genetics **63**: 949–967.

WEITKAMP, L. R., T. ARENDS, M. L. GALLANGO, J. V. NEEL, J. SCHULTZ and D. C. SHREFFLER, 1972 The genetic structure of a tribal population, the Yanomama Indians. III. Seven serum protein systems. Ann. Human Genet. **35**: 271–279.

WILSON, J., 1970 Experimental design in fitness estimation. Genetics **66**: 555–567.

WORKMAN, P. L. and J. D. NISWANDER, 1970 Population studies on Southwestern Indian tribes. II. Local genetic differentiation in the Papago. Amer. J. Human Genet. **22**: T–23.

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

YAMAZAKI, T., 1971 Measurement of fitness at the esterase-5 locus of *Drosophila pseudoobscura*. Genetics **67**: 597–603.