Evolution, 43(6), 1989, pp. 1236–1251

TEMPORAL VARIATION IN ALLELE FREQUENCIES: TESTING THE RIGHT HYPOTHESIS

ROBIN S. WAPLES
National Marine Fisheries Service, Northwest Fisheries Center,
2725 Montlake Blvd. East, Seattle, WA 98112

Abstract.—Although standard statistical tests (such as contingency chi-square or G tests) are not well suited to the analysis of temporal changes in allele frequencies, they continue to be used routinely in this context. Because the null hypothesis stipulated by the test is violated if samples are temporally spaced, the true probability of a significant test statistic will not equal the nominal α level, and conclusions drawn on the basis of such tests can be misleading. A generalized method, applicable to a wide variety of organisms and sampling schemes, is developed here to estimate the probability of a significant test statistic if the only forces acting on allele frequencies are stochastic ones (i.e., sampling error and genetic drift). Results from analyses and simulations indicate that the rate at which this probability increases with time is determined primarily by the ratio of sample size to effective population size. Because this ratio differs considerably among species, the seriousness of the error in using the standard test will also differ. Bias is particularly strong in cases in which a high percentage of the total population can be sampled (for example, endangered species). The model used here is also applicable to the analysis of parent-offspring data and to comparisons of replicate samples from the same generation. A generalized test of the hypothesis that observed changes in allele frequency can be satisfactorily explained by drift follows directly from the model, and simulation results indicate that the true α level of this adjusted test is close to the nominal one under most conditions.

Received December 14, 1987. Accepted April 17, 1989

Analysis of temporal variation of allele frequencies is one of the most promising means of studying microevolutionary processes. It is not surprising, therefore, that in addition to empirical studies on a wide range of organisms (e.g., Dobzhansky, 1943; Gaines and Krebs, 1971; Williams et al., 1973; Wall et al., 1980; Franklin, 1981; Bowen, 1982; Barker et al., 1986; Jacobson et al., 1986), there have been numerous theoretical treatments of this topic (Charlesworth and Giesel, 1972; Lewontin and Krakauer, 1973; Pamilo and Varvio-Aho, 1980; Nei and Tajima, 1981; Pollak, 1983). Theoretical analyses necessarily concern various aspects of the sampling processes involved, both in choosing gametes to form the next generation (genetic drift) and in choosing the sample for genetic analysis. Several studies (Fisher and Ford, 1947; Templeton, 1974; Schaffer et al., 1977; Gibson et al., 1979; Wilson, 1980; Watterson, 1982; Mueller et al., 1985) have developed means of testing alternative hypotheses regarding the causes of temporal changes in allele frequency. These tests, however, have not been widely used by others. For routine analysis of temporal variability, biologists still commonly use a standard statistical test (such as the homogeneity chi-square test) to determine whether the observed differences are "significant" (Krimbas and Tsakas, 1971; Koehn and Williams, 1978; Cavener and Clegg, 1981; Cornejo de Caminos et al., 1981; Mihok et al., 1983; Smith et al., 1983; Johnson and Black, 1984; Gyllensten, 1985; Apfelbaum and Blanco, 1985; Korpelainen, 1986; Black and Krafsur, 1986).

Gibson et al. (1979) pointed out that a standard test is not appropriate for the analysis of temporal changes, because the null hypothesis typically of interest (that observed differences in allele frequencies are due entirely to stochastic processes in a finite population) does not coincide with the null hypothesis stipulated by the test. Clearly, the use of significant test results as evidence for nonrandom forces acting on allele frequencies is unwarranted if drift alone will also lead to a high percentage of "significant" results. Here, a method is described to estimate the probability of a significant test statistic under pure drift conditions for a variety of realistic sampling schemes. A straightforward extension of the method leads to a generalized chi-square test that takes genetic drift into consideration.

Materials and Methods

The Model

The following model is for the contingency chi-square test of homogeneity of allele frequencies, but the conclusions are essentially the same for other commonly used tests (e.g., G test, t test). If A_1 and A_2 are the numbers of A alleles at a locus in two samples (of S_1 and S_2 diploid individuals), then the corresponding numbers of "non-A" alleles are $B_1 = 2S_1 - A_1$, and $B_2 = 2S_2 - A_2$. This suggests the 2 × 2 table of the numbers of alleles of each kind in the two samples shown in Table 1. With the nominal α level set to 0.05, the null hypothesis is rejected if $X^2 > 3.84$, where

$$X^2 = \frac{2(S_1 + S_2)(A_1B_2 - B_1A_2)^2}{4S_1S_2(A_1 + A_2)(B_1 + B_2)} .$$

But what is the proper null hypothesis? The model for this test stipulates that S_1 and S_2 are binomial samples with identical probability P of choosing an A allele at each trial. If this assumption holds, then the proportion of A alleles in the two samples ($x = A_1/[2S_1]$ and $y = A_2/[2S_2]$) are independent estimates of the same parameter, P. The variance of x [$V(x)$] and the variance of y [$V(y)$] around the parametric value P are just the binomial variances for samples of size $2S_1$ and $2S_2$:

$$V(x) = E(x - P)^2 = \frac{P(1 - P)}{2S_1} \quad (1a)$$

$$V(y) = E(y - P)^2 = \frac{P(1 - P)}{2S_2} . \quad (1b)$$

Furthermore, because the samples are independent,

$$\begin{aligned} V(x - y) &= V(x) + V(y) \\ &= P(1 - P)\left(\frac{1}{2S_1} + \frac{1}{2S_2}\right) \\ &= \mathrm{BV} \quad (2) \end{aligned}$$

where BV is used to denote the variance of $(x - y)$ if sampling is binomial.

The assumptions of the model can be made explicit by noting that the quantity $(x - y)^2/V(x - y)$ is distributed as chi-square. The test statistic for the 2 × 2 contingency test can also be computed as

TABLE 1. The 2 × 2 table of the numbers of A and non-A alleles in two samples of S_1 and S_2 individuals (see text).

	Number of alleles		
Sample	A	non-A	Sum
1	A_1	B_1	$2S_1$
2	A_2	B_2	$2S_2$
Sum	$A_1 + A_2$	$B_1 + B_2$	$2(S_1 + S_2)$

$$\frac{(x - y)^2}{P(1 - P)\left(\frac{1}{2S_1} + \frac{1}{2S_2}\right)} = \frac{(x - y)^2}{\mathrm{BV}} .$$

As the parameter P is generally not known, it is estimated by $\hat{P}$, the weighted (by sample size) mean of x and y.

Drawing multiple samples from a natural population (as in the analysis of temporal variation) generally results in deviations from the above model. To illustrate, assume now that samples of S_0 and S_t individuals are taken from the same population in generations 0 and t, and let P be the frequency of an allele in the gamete pool preceding generation 0. The initial sample can be considered to be a binomial draw from the initial gamete pool, so $V(x)$ is identical to (1a). However, $V(y)$ is larger than (1b), because allele frequencies in the second sample also reflect t generations of genetic drift. As a result, $V(x - y)$ will generally not equal BV, the variance upon which the test statistic is based, and the probability of a significant test statistic will deviate from the presumed α level. In particular, if $V(x - y)$ exceeds BV, the probability of a significant test statistic is greater than α, even if no directional forces are involved.

It is thus clear that the null hypothesis stipulated by the model is generally not satisfied under a sampling scheme involving temporal variation. However, if the true value of $V(x - y)$ can be estimated, two alternative approaches are still available. First, note that the expression $(x - y)^2/V(x - y)$ is still a chi-square variate and can form the basis for a statistical test. That is, if the true value of $V(x - y)$ is used instead of BV in the denominator of the chi-square test, one can test the hypothesis that the observed differences in allele frequency can be explained entirely by genetic drift and sampling error. More will be said about this approach later. Second, if the standard test is used (BV in the denominator), it is possible to predict the probability of obtaining a significant test statistic. That is, the power of the test to detect departures from the null hypothesis (two independent binomial samples from the same probability distribution) can be evaluated. If we write $X^2 = (x - y)^2/V(x - y)$, $X^{2'} = (x - y)^2/\mathrm{BV}$, it is clear that $X^{2'} = CX^2$, where $C = V(x - y)/\mathrm{BV}$ is a scaling factor. Consider now the probability of a significant test statistic if the standard test is used (i.e., the probability that $X^{2'} \geq 3.84$):

$$\begin{aligned}\Pr[X^{2'} \geq 3.84] &= \Pr[CX^2 \geq 3.84] \\ &= \Pr\left[X^2 \geq \frac{3.84}{C}\right] \\ &= \Pr\left[X^2 \geq \frac{3.84}{\dfrac{V(x - y)}{\mathrm{BV}}}\right]. \end{aligned} \tag{3}$$

The probability of a significant test statistic is thus a function of $V(x - y)$ and can be obtained from tables of the χ^2 distribution (or, equivalently, from tables of the normal frequency distribution by computing the probability that a standard normal variate $[|Z|]$ is at least $1.96/\sqrt{C}$). $V(x - y)$, in turn, is a function of sample size, effective and total population sizes, elapsed number of generations, and the method of sampling.

To examine the effects of these parameters on the variance in allele frequencies, I will consider a model similar to that described above in more detail (see Waples [1989] for additional details). A diploid population with effective population size N_e is sampled for genetic analysis in generations 0 and t. P, S, x, and y are as defined above. Discrete generations are assumed, but results will also be applicable in many cases to organisms with overlapping generations (see Pollak, 1983). Because only the expectations under pure drift conditions are of interest, selection, migration, and mutation are ignored.

Nei and Tajima (1981) pointed out that there are two possible ways of sampling the individuals to be analyzed genetically (Fig. 1). The same two sampling plans will be considered here, although the model of Nei and Tajima differs from the present one in other respects. In sampling plan I, individuals are sampled after reproduction or are replaced before reproduction occurs, whereas in plan II the sample for genetic analysis in generation 0 is taken before reproduction and is not replaced. Waples (1989) showed that for both plans,

$$V(x) = E(x - P)^2 = \frac{P(1 - P)}{2S_0} \tag{4}$$

$$\begin{aligned} V(y) &= E(y - P)^2 \\ &= P(1 - P)\left[1 - \left(1 - \frac{1}{2S_t}\right) \cdot \left(1 - \frac{1}{2N_e}\right)^t\right] \end{aligned} \tag{5}$$

$$\begin{aligned} V(x - y) &= E(x - y)^2 \\ &= P(1 - P) \\ &\quad \cdot\left[\frac{1}{2S_0} + 1 - \left(1 - \frac{1}{2S_t}\right) \cdot\left(1 - \frac{1}{2N_e}\right)^t\right] \\ &\quad - 2\mathrm{Cov}(x, y). \end{aligned} \tag{6}$$

As mentioned above, $V(x)$ in (4) is identical to the binomial variance in (1a), but $V(y)$ in (5) is larger than that in (1b) because of genetic drift (represented by the term $(1 - 1/[2N_e])^t$).

Although $V(x)$ and $V(y)$ are the same under the two sampling plans, $V(x - y)$ will not in general be the same, because the covariance term differs for the two plans. In plan II (sampling before reproduction), the samples of $2S_0$ genes for genetic analysis and $2N_e$ genes representing the effective population size in generation 0 are mutually exclusive and can be considered to be independent binomial draws from the initial gamete pool. Therefore, frequencies in samples S_0 and S_t are uncorrelated and $\mathrm{Cov}(x, y) = 0$. If sampling is after reproduction (plan I), the samples of $2S_0$ genes and $2N_e$ genes in generation 0 may overlap and are therefore positively correlated with respect to P. This positive correlation, arising from the fact that both samples are derived from a finite population of size N, means that $\mathrm{Cov}(x, y)$ will also be positive. Waples (1989) showed that $\mathrm{Cov}(x, y) = P(1 - P)/(2N)$ if sampling is according to plan I, where N is the total population size in generation 0. Substituting the appropriate covariance terms into (6) yields

$$V(x - y) = P(1 - P) \cdot \left[\frac{1}{2S_0} + 1 - \left(1 - \frac{1}{2S_t}\right) \cdot \left(1 - \frac{1}{2N_e}\right)^t - \frac{1}{N}\right] \quad (7a)$$

$$V(x - y) = P(1 - P) \cdot \left[\frac{1}{2S_0} + 1 - \left(1 - \frac{1}{2S_t}\right) \cdot \left(1 - \frac{1}{2N_e}\right)^t\right] \quad (7b)$$

for plan I and plan II, respectively.

Equations (7a) and (7b) allow us to evaluate $V(x - y)$ and, using (3), the probability of a significant test statistic for any values of N, N_e, S_0, S_t, and t. Results of these analyses are given below; a nominal α level of 0.05 is assumed. For simplicity, a constant sample size, $S = S_0 = S_t$ is considered for most of the following. If sample sizes differ in generations 0 and t, then S in the following can be interpreted as the harmonic mean sample size: $\tilde{S} = 2/[1/(2S_0) + 1/(2S_t)]$. To see this, note that the total contribution of sampling error to $V(x - y)$ is given by $1/(2S_0) + [1 - 1/(2N_e)]^t/(2S_t)$. Unless t is very large or N_e very small, $[1 - 1/(2N_e)]^t$ will be close to 1 and $\tilde{S}$ will be an adequate description of sample size. Similarly, if N_e is not constant over time, then the appropriate value to use in (7a) or (7b) is the harmonic mean of the single-generation N_e's (Nei and Tajima, 1981).

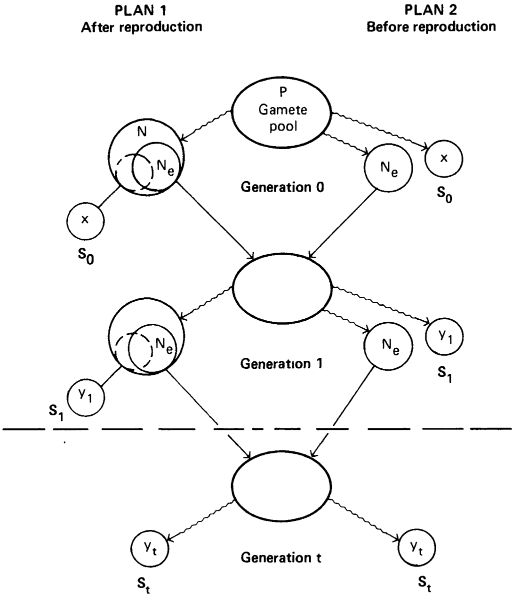

FIG. 1. Two sampling plans considered in the analysis. In both plans, P is the frequency of an allele in the gamete pool preceding generation 0; x and y_t are allele frequencies in samples (of S_0 and S_t individuals) for genetic analysis taken at generations 0 and t, respectively; N is the total population size at the time of the initial sample; and N_e is the variance effective population size. Plan I: sample S_0 is taken after reproduction, so it may contain some of the $2N_e$ genes representing effective population size. Allele frequencies in samples S_0 and S_t are positively correlated with respect to P, because they are derived from the same population (size N) at generation 0. Plan II: the sample is taken before reproduction and not replaced, so the samples S_0 and N_e are mutually exclusive and can be considered to be independent binomial draws from the initial gamete pool. Total population size is not a factor, and x and y_t are uncorrelated.

Modifying the Standard Chi-Square Test.—The need for an appropriate test for the study of temporal variation was recognized long ago. In their study of allele-frequency change in the moth *Panaxia dominula*, Fisher and Ford (1947) designed a test to take the effects of genetic drift into consideration. Variations of this approach have been used subsequently by other authors, but none has been adopted generally, perhaps because each was derived for specific organisms (Mueller et al., 1985) or sampling schemes (Fisher and Ford, 1947; Schaffer et al., 1977; Gibson et al., 1979; Wilson, 1980; Watterson, 1982). However, Waples (1989) showed that different sampling plans can be treated in a uniform way, and this permits a more generalized adjusted test applicable to a wide variety of organisms.

It was noted above that the quantity $(x - y)^2/V(x - y)$ is distributed as χ^2 and can be used as the basis for a test, provided $V(x - y)$ can be estimated. This can be done using Equations (7a) and (7b) for the two sampling plans. A significant test statistic would imply that something besides genetic drift and sampling error must be invoked to explain the changes observed.

The method of estimating P to use in (7a) and (7b) requires some consideration. A simple weighting by sample size is not appropriate for temporally spaced samples. Both x and y are unbiased estimates of P, but for the same sample size, x is a more precise estimate, because it has a smaller variance. The problem is to choose weighting terms (β and $1 - \beta$) for x and y to minimize the variance of $\hat{P} = \beta x + (1 - \beta)y$. This can be done by choosing β such that

$$\beta = \frac{V(y) - \mathrm{Cov}(x, y)}{V(x - y)}. \qquad (8)$$

If $\mathrm{Cov}(x, y) = 0$ (plan II), this is equivalent to weighting by the reciprocals of the variances; if sampling actually is binomial, this weighting procedure reduces to that of the standard (unadjusted) chi-square test, as it should.

Multiple Alleles.—If the standard chi-square test is used for more than $K = 2$ alleles, the probability of a significant test result is higher than would be the case for a diallelic locus. The probability is given by

$$\Pr\left[X^2 \geq \frac{Q_{(K-1)}}{C}\right]$$

where $Q_{(K-1)}$ is the critical χ^2 value for $K - 1$ degrees of freedom and $C = V(x - y)/BV$. This expression reflects the fact that all the parameters that affect C (S_0, S_t, N_e, t, and N) are identical for every allele at a locus. See the Appendix for an adjusted chi-square test using multiple alleles.

Multiple Loci.—When data are available for more than a single gene locus, a common practice is to perform an overall test by summing the chi-square values and degrees of freedom over all loci ($X^2{}_T = X^2{}_1 + X^2{}_2 + \ldots + X^2{}_L$, where L = the number of loci). If an adjusted test is used for each locus, this combined test should still have probability α of indicating statistical significance by chance alone, provided that nothing besides drift and sampling error is involved. If the single-locus tests are not adjusted to account for drift, the probability of a significant combined test depends on the total number of degrees of freedom ($d = \Sigma[K_i - 1]$, where the summation is over all loci). If sample sizes are the same at each locus, the quantity C is the same for each locus as well as for each allele, and the probability of a significant test is

$$\Pr\left[X^2{}_T \geq \frac{Q_{[d]}}{C}\right]. \qquad (9)$$

For example, if $C = 2$ and data for ten diallelic loci are used, the probability of a significant test result for each individual locus is $\Pr[X^2 \geq 3.84/2] = \Pr[X^2 \geq 1.92] \approx 0.17$, whereas the corresponding probability for the combined test is $\Pr[X^2 \geq Q_{[10]}/2] = \Pr[X^2 \geq 9.155] \approx 0.52$. Thus, use of the combined chi-square test without correction can lead to seriously erroneous conclusions.

Multiple Samples.—If samples are available from more than two time periods, the situation is more complicated, because allele frequencies in all samples taken after generation 0 are positively correlated with respect to P, regardless of the sampling plan. For samples S_a and S_b taken in generations a and b (frequencies y_a, y_b), the magnitude of the covariance is a function of the number of generations beyond generation 0 that the two samples shared a common evolutionary pathway. In plan-II sampling, both samples are derived from the same gamete pool (frequency P_a) preceding generation a, so their covariance involves a generations of genetic drift. In fact, $\mathrm{Cov}(y_a, y_b)$ is just the variance of P_a with respect to P [$V(P_a \mid P) = E(P_a - P)^2$], so in plan-II sampling

$$\mathrm{Cov}(y_a, y_b) = P(1 - P)\left[1 - \left(1 - \frac{1}{2N_e}\right)^a\right] \qquad (10a)$$

If sampling is after reproduction, the common history of samples S_a and S_b further extends to the N_a individuals in the population in generation a, so in plan-I sampling

$$\mathrm{Cov}(y_a, y_b) = P(1 - P)\left[1 - \left(1 - \frac{1}{2N_e}\right)^a \cdot \left(1 - \frac{1}{2N_a}\right)\right]. \qquad (10b)$$

An adjusted chi-square test (using $t_j + 1$ samples and therefore having t_j degrees of

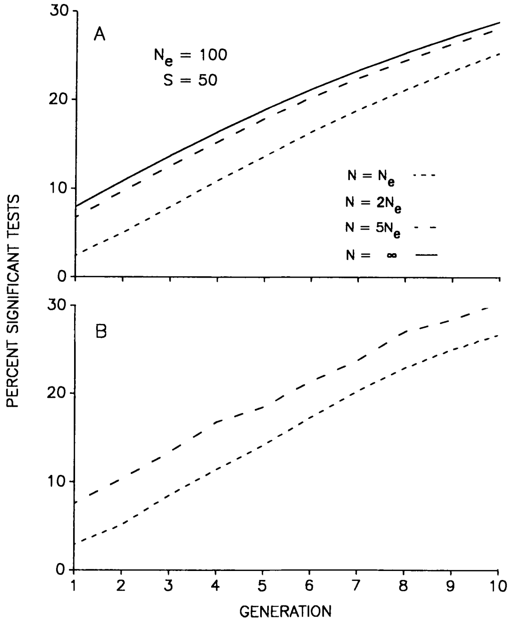

FIG. 2. Effects of total population size (N) on the probability of a significant test statistic (standard chi-square test of equality of allele frequencies in samples taken a number of generations apart) for plan-I sampling. A) Expected results obtained analytically using method described in the text; B) results observed in simulations. If $N > 2N_e$, results for plan-I sampling are similar to those for plan-II sampling (equivalent to setting $N = \infty$).

freedom) that takes drift and the above covariances into consideration is

$$X^2 = \begin{bmatrix} x - \hat{P} \\ y_{t1} - \hat{P} \\ \vdots \\ y_{tj} - \hat{P} \end{bmatrix}^{\mathrm{T}} \Sigma^{-1} \begin{bmatrix} x - \hat{P} \\ y_{t1} - \hat{P} \\ \vdots \\ y_{tj} - \hat{P} \end{bmatrix}. \qquad (11)$$

In the above, x is the sample allele frequency in generation 0, y_{t_i} is the sample frequency in generation t_i, $\hat{P}$ is the weighted mean estimate of P (see below), and Σ^{-1} is the inverse of the variance-covariance matrix.

The diagonal elements of Σ are variances of sample allele frequencies, which are given by (4) for the first sample and by (5) for the rest; the off-diagonal elements are covariances given by (10a) or (10b), depending on the sampling plan. Equations (4), (5), (10a), and (10b) are functions of P, which is estimated by $\hat{P}$:

$$\hat{P} = \frac{\mathbf{1}^{\mathrm{T}}\Sigma_*^{-1}\mathbf{P}}{M}. \qquad (12)$$

In (12), $\mathbf{P}$ is a column vector of the sample allele frequencies, $\mathbf{1}^{\mathrm{T}}$ is a row vector of 1's of the same length as $\mathbf{P}$, Σ_* is Σ omitting the common term $P(1 - P)$ for each element, and M is the sum of all the elements in Σ_*^{-1}.

If effective population size varies over time and multiple samples are taken, then the above terms of the form $[1 - 1/(2N_e)]^t$ can be replaced by $\Pi\,[1 - 1/(2N_{e_j})]$, where N_{e_j} is the effective population size in the jth generation.

Computer Simulations.—To evaluate the accuracy of results obtained analytically and to consider properties of the adjusted chi-square test, computer simulations were performed as described by Waples (1989). In each simulation, samples were drawn in the initial generation and for ten consecutive generations thereafter according to plan I or plan II. For each set of initial parameters P, N, N_e, and S ($S = S_0 = S_t$), 5,000 replicates of this process were performed, and the percentage of trials yielding a statistically significant test statistic was recorded. Simulations using three alleles at a locus were also performed. Chi-square tests were not performed for replicates in which an allele was absent in all samples being considered.

RESULTS

Plan I versus Plan II.—The only difference between $V(x - y)$ for plans I and II [compare (7a) and (7b)] is the covariance term, $-P(1 - P)/N$, for plan I. As N increases, the covariance diminishes, it makes less difference whether sampling is before or after reproduction, and $V(x - y)$ for plan I converges to the value for plan II. In the limit ($N = \infty$), the two sampling plans are equivalent. In practice, the difference between the two plans is minor unless the ratio $r = N/N_e$ is small. Figure 2A shows that, if N is at least twice as large as N_e, ignoring N and estimating $V(x - y)$ as if sampling were according to plan II gives a close approximation to the true probability of a significant test statistic. This result was also observed in the simulations (Fig. 2B). However, if effective and total population sizes are nearly equal, a reasonably accurate estimate of N is necessary, particularly in the early generations.

It is important to note that, in this context, N is the size of the population subject

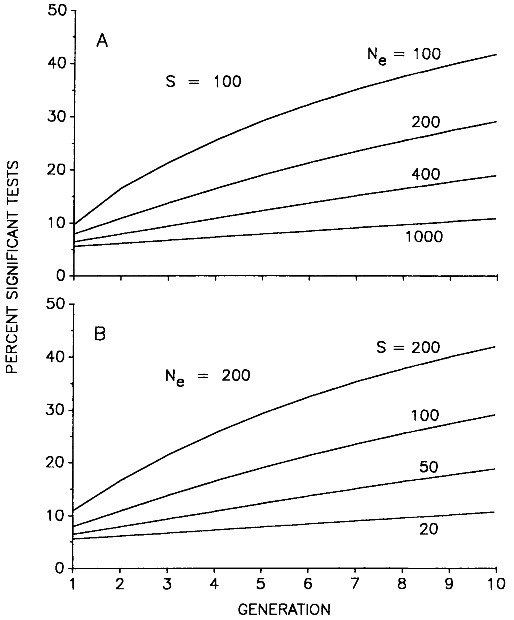

FIG. 3. The probability of a significant test statistic as a function of A) effective population size and B) sample size. Results were obtained analytically assuming plan-II sampling.

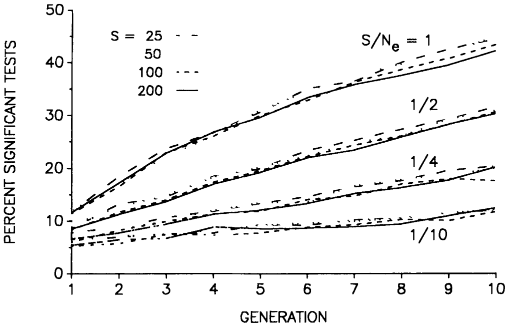

FIG. 4. The probability of a significant test statistic as a function of the ratio of sample size to effective population size (S/N_e). Results were obtained from simulations using $S = S_0 = S_t = 25$, 50, 100, and 200 and values of N_e in the proportions $S/N_e = 1$, ½, ¼, $^1/_{10}$. Sampling was according to plan II. The probability of a significant test is sensitive to the ratio S/N_e but not to absolute values of S or N_e.

to sampling in generation 0 and that for a given species the ratio N/N_e may vary depending on the life-history stage sampled. If only adults are taken, N may represent the number of breeding individuals, which may be approximately the same as N_e. On the other hand, if juveniles of a very fecund species are sampled, N is likely to be very large, and it matters little whether the sample is taken before or after reproduction. To reduce complexity of most of the remaining analyses, only plan II sampling will be considered.

Agreement with Analytical Results.—As illustrated in Figures 2 and 5, the percentage of significant tests observed in the simulations agreed well with expectations based on analytical results. Nevertheless, in simulations using moderate allele frequencies, there was a consistent, slight elevation in the ratio of observed to expected numbers of significant tests. This ratio is a function of sample size, and it declined from about 1.10 for $S = 25$ to 1.02 for $S = 100$, averaged over all simulations. Analysis of normal probability plots suggests that this result is due to a slight departure from normality of the quantity $(x - y)$. In any event, the effect is small unless samples of very limited size are taken.

Effective Population Size.—The probability of a significant test statistic is, as expected, strongly dependent on effective population size (Fig. 3A). Smaller N_e leads to larger variance in allele frequencies and a higher percentage of statistically significant tests.

Sample Size.—The probability of a significant test also increases dramatically with sample size (Fig. 3B). This is because a larger sample will more accurately reflect the real temporal differences in allele frequencies caused by genetic drift.

Ratio of Sample Size to Effective Population Size.—Comparison of a number of graphs such as those in Figure 3 for various values of S and N_e yields an interesting result: it is not the actual values of sample size or effective population size that are important in determining the probability of a significant test statistic, but rather their ratio. Thus, the curve in Figure 3A for $S = 100$ and $N_e = 200$ is virtually identical to the curve for any combination of S and N_e in the proportion 1:2.

This result can be demonstrated analytically by noting that, unless t is very large or N_e very small, a good approximation for (7b) is

$$V(x - y) \approx P(1 - P)\left(\frac{1}{2S_0} + \frac{1}{2S_t} + \frac{t}{2N_e}\right).$$

The probability of a significant test result is determined by the ratio $V(x - y)/\mathrm{BV}$:

$$\frac{V(x-y)}{\mathrm{BV}} \approx \frac{P(1-P)\left(\frac{1}{2S_0}+\frac{1}{2S_t}+\frac{t}{2N_e}\right)}{P(1-P)\left(\frac{1}{2S_0}+\frac{1}{2S_t}\right)}$$

$$= 1 + \left[\frac{2S_0S_t}{(S_0+S_t)}\right]\frac{t}{2N_e}$$

$$= 1 + \frac{\tilde{S}t}{2N_e}.$$

The same dependence on the ratio S/N_e holds for sampling after reproduction, because the covariance term (of order $1/N$) will generally be small compared to the term $\tilde{S}t/(2N_e)$. Again, however, if $r = N/N_e$ is small, the effects of total population size may be important in the early generations.

The strength of the dependence on the ratio S/N_e is illustrated in Figure 4, which shows results for simulations using sample sizes of $S = 25$, 50, 100, and 200 and $N_e = S$, $2S$; $4S$, and $10S$. Almost all of the variation between simulations is due to differences in the ratio S/N_e, with the actual values of S or N_e having little effect.

Initial Allele Frequency.—Allele frequency appears in (2) and (6) only as the term $P(1 - P)$, which is a factor of both equations. Thus, the ratio $V(x - y)/\mathrm{BV}$, as well as the probability of a significant test result, is theoretically independent of initial allele frequency. In practice, some effects of allele frequency are expected if P is close to 0 or 1, particularly for small samples. In simulations using $N_e = 100$ and $S= 50$, initial allele frequency had little effect on the probability of a significant test statistic for either two- or three-allele systems, unless alleles with initial frequency less than 0.05 were included (Fig. 5). For more extreme allele frequencies, the observed frequency of significant tests was less than expected based on (3), and the deviations increased with the number of generations between samples.

Multiple Alleles and Multiple Loci.—So far I have only considered data for a single locus with two alleles. Figure 5 shows the observed percentage of significant chi-square tests based on simulated data for $L = 1$, 2, 5, or 10 loci, with $N_e = 100$ and $S = 50$ and setting $t = 5$. For each value of L in each simulation, 2,000 multilocus chi-square values were computed by sampling with replacement from the single-locus values generated in the simulations. For diallelic loci (Fig. 5A), the observed percentage of significant tests was close to that expected (solid lines), even for $P = 0.95$, but was much lower than expected for $P = 0.99$. Note that the slight elevation of the ratio of observed to expected numbers of tests noted above for single loci was somewhat more pronounced for $L = 5$ or $L = 10$. For simulations with three alleles at a locus, alleles with frequency as small as 0.01 had even less effect on the probability of a significant test (Fig. 5B).

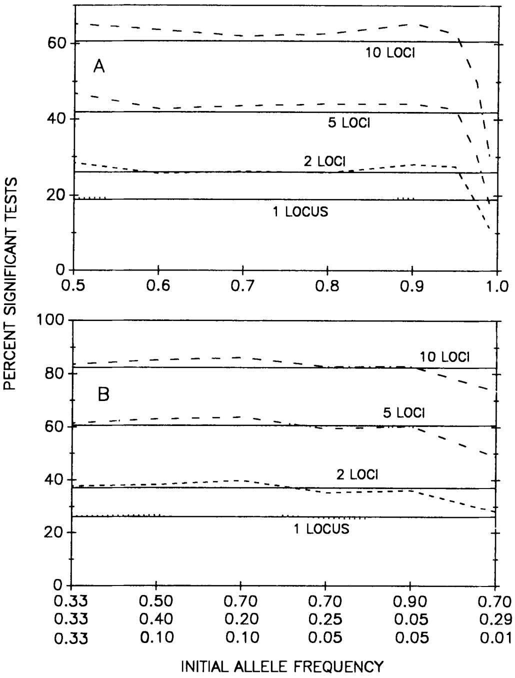

FIG. 5. The probability of a significant test statistic for a combined chi-square test using data for multiple gene loci. Broken lines are results obtained from simulations using A) two or B) three alleles at each locus and a range of initial allele frequencies; solid lines are expected probabilities obtained analytically. In each simulation, effective population size was 100, sample size was 50, samples were taken five generations apart, and sampling was according to plan II.

The Adjusted Test.—Adjusted chi-square tests that take drift into consideration were also performed for each replicate in the simulations. Figure 6 shows results of adjusted

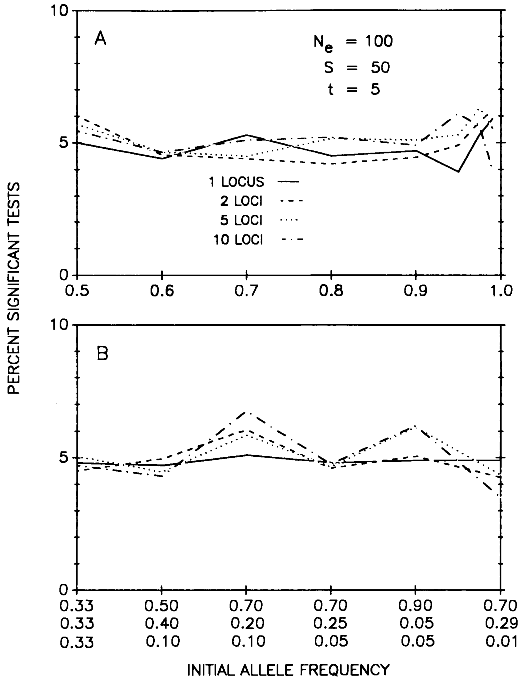

FIG. 6. The percentage of statistically significant adjusted chi-square tests, using data for 1–10 gene loci. Tests were adjusted to account for genetic drift as described in the text. Results were obtained from simulations using A) two or B) three alleles at each locus. Initial conditions were as in Figure 5: effective population size was 100, sample size was 50, samples were taken five generations apart, and sampling was according to plan II.

tests using two samples, 1–10 loci (each with two or three alleles), and the same initial conditions as in Figure 5. Even when the frequency of one allele was as low as 0.01, the observed proportion of significant tests was very close to the nominal $\alpha = 0.05$ level.

The adjusted tests depicted in Figure 7 were computed using multiple samples, with the initial sample taken in generation 0 and subsequent samples taken each generation for a total of three, six, or 11 samples. Again, under most circumstances the observed proportion of significant tests was close to the expected 5%. However, the true α level of the test was higher than expected if six or 11 samples were considered and initial allele frequencies were extreme. That this effect was more pronounced when more samples were considered probably reflects the fact that, after a number of generations, allele frequencies initially close to 0 or 1 are likely to be affected by boundary conditions. True α levels closer to 0.05 were observed in simulations using larger sample sizes or larger effective population size.

Special Cases

Parent-Offspring Data.—If the data consist of allele frequencies for parents and their offspring, then clearly, sampling was according to plan I (sample taken after reproduction or replaced before reproduction). Probability of a significant test depends on the ratio $V(x - y)/\mathrm{BV}$, which, using (7a) and (2), cancelling the common term $P(1 - P)$, and setting $t = 1$, can be written in the form given below. Ignoring the last term in the numerator, which will be small, $V(x - y) \approx \mathrm{BV}$ if $N = 2N_e$, in which case use of the standard test will not be misleading. If $N > 2N_e$, the standard test can be expected to produce "significant" results more often than the nominal α level, while the test will be conservative if N_e is more than half the total number of adults subject to sampling. In interpreting parent-off-

Parent-offspring data:

$$\frac{V(x-y)}{\mathrm{BV}} = \frac{\dfrac{1}{2S_0} + 1 - \left(1 - \dfrac{1}{2N_e}\right)\left(1 - \dfrac{1}{2S_t}\right) - \dfrac{1}{N}}{\dfrac{1}{2S_0} + \dfrac{1}{2S_t}}$$

$$= \frac{\dfrac{1}{2S_0} + \dfrac{1}{2S_t} + \dfrac{1}{2N_e} - \dfrac{1}{N} - \dfrac{1}{4N_eS_t}}{\dfrac{1}{2S_0} + \dfrac{1}{2S_t}}$$

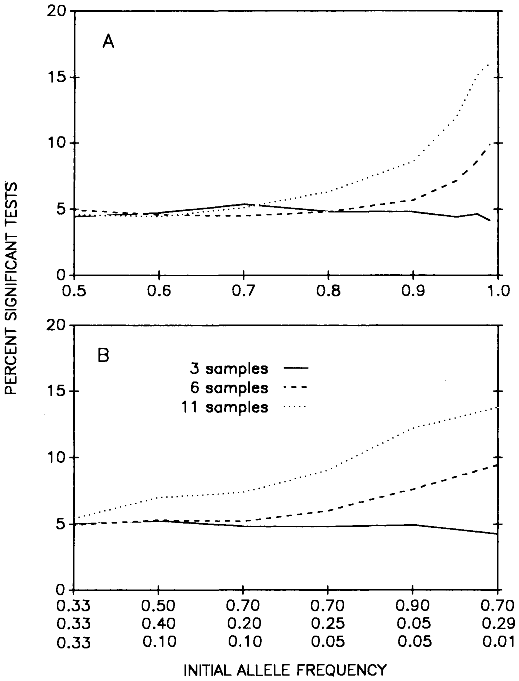

FIG. 7. The percentage of statistically significant adjusted chi-square tests, using data for 3–11 temporally spaced samples taken in successive generations. Tests were adjusted to account for genetic drift as described in the text. Results were obtained from simulations using A) two or B) three alleles at each locus. Initial conditions: effective population size was 100, sample size was 50, and sampling was according to plan II.

spring gene frequency data, therefore, it is important that both total and effective population sizes be considered.

Replicate Samples in the Same Generation.—If two or more replicate samples are taken in the same generation, drift is not a factor and the variance of allele frequencies for both samples is given by (4). However, the covariance of allele frequencies in the two samples depends on how the samples are drawn, and the principles behind sampling plans I and II are applicable here. The relevant question is whether the samples, once they are drawn, are replaced before subsequent ones are taken. (It is assumed that within each sample, individuals are taken without replacement.) If entire samples are not replaced, they are independent with respect to P, so a standard statistical test is appropriate. If, however, the initial samples are replaced and those individuals are subject to being chosen again, frequencies in the samples will be positively correlated with respect to P, $V(x - y)$ will be less than BV, and the test will be conservative.

TABLE 2. Probability of a significant chi-square test for comparison of allele frequencies between 1966 and 1967 in the *Dacus oleae* population studied by Krimbas and Tsakas (1971). Probability is shown for various possible values of effective population size (N_e) and number of generations (t) between samples.

		Generation		
Gene	N_e	$t = 3$	$t = 4$	$t = 5$
A	100	0.45	0.50	0.54
	300	0.25	0.30	0.33
	500	0.18	0.22	0.25
	1,000	0.12	0.14	0.16
	10,000	0.06	0.06	0.06
B	100	0.43	0.49	0.53
	300	0.24	0.28	0.32
	500	0.17	0.21	0.24
	1,000	0.11	0.13	0.15
	10,000	0.06	0.06	0.06

Numerical Examples

The following examples illustrate the practical application of the methods described here. Krimbas and Tsakas' (1971) data for *Dacus oleae* sampled in 1966 and 1967 are of interest, because of their statement that "the population . . . seems to be at a stable equilibrium for genes *A* and *B*" (Krimbas and Tsakas, 1971 p. 456). The significance of the change (or lack thereof) in allele frequencies can be evaluated as follows. Sample sizes were large: 474 flies in 1966 and 312 flies in 1967 for locus *A*; 469 flies in 1966 and 281 flies in 1967 for locus *B*. The authors estimated that four generations elapsed between the two samples. As pointed out elsewhere (Nei and Tajima, 1981; Pollak, 1983), sampling in this case was according to plan II, so total population size was not a factor. For both loci, many alleles were observed in at least one of the samples (17 at locus *A* and 13 at locus *B*). For each allele at a given locus, the parameters S_0, S_t, N_e, N, and t are the same, so the quantity $C = V(x - y)/\text{BV}$ is constant. Between loci, only S_0 and S_t change. Thus, for locus *A*, the probability of a significant chi-square test for a given allele can be computed by substituting appropriate values in (7b) and referring to (3):

$$V(x-y) = P(1-P)\left[\frac{1}{948} + 1 - \left(1 - \frac{1}{2N_e}\right)^4\left(1 - \frac{1}{624}\right)\right]$$

$$= P(1-P)\left[1.001055 - 0.9984\left(1 - \frac{1}{2N_e}\right)^4\right]$$

$$\frac{V(x-y)}{\mathrm{BV}} = \frac{P(1-P)\left[1.001055 - 0.9984\left(1 - \frac{1}{2N_e}\right)^4\right]}{P(1-P)\left(\frac{1}{948} + \frac{1}{624}\right)}$$

$$= \frac{1.001055 - 0.9984\left(1 - \frac{1}{2N_e}\right)^4}{0.00266}$$

$$= 376.336 - 375.338\left(1 - \frac{1}{2N_e}\right)^4.$$

The probability of a significant test is

$$\Pr\left[X^2 \geq \frac{3.84}{\frac{V(x-y)}{\mathrm{BV}}}\right] = \Pr\left[X^2 \geq \frac{3.84}{376.336 - 375.338\left(1 - \frac{1}{2N_e}\right)^4}\right].$$

For locus B, a similar approach indicates that this probability is

$$\Pr\left[X^2 \geq \frac{3.84}{351.87 - 350.87\left(1 - \frac{1}{2N_e}\right)^4}\right].$$

Table 2 shows these probabilities for N_e in the range of 10^2–10^4. For $t = 4$ and $500 < N_e < 1{,}000$ (as estimated for 1966–1967 by Krimbas and Tsakas [1971], Nei and Tajima [1981], and Pollak [1983]), it is clear that one can expect about ¼–⅛ of such chi-square tests to be significant. Calculation of X^2 values for each individual allele (using only those alleles with expected numbers >1) resulted in three out of 12 significant tests for locus A (alleles A_6, A_8, and A_s) and two of 12 for locus B (alleles B_l and B_{45}). Overall, the percentage of significant tests (5/24 ≈ 1/5) was about as large as could be expected from drift and sampling error and does not provide evidence of genetic equilibrium.

This conclusion, however, may be misleading if some of the parameters were incorrectly estimated. For example, Krimbas and Tsakas (1971) estimated four generations per year but acknowledged a possible range of 3–5. If five generations rather than four elapsed between samples, more change would be expected: the probability of a significant test is about ⅓–½ for N_e in the range of 100–300 (Table 2). Thus, the observed changes in allele frequency were more modest than expected if effective population size was this small and five generations separated the samples.

To test the hypothesis that observed changes for a particular allele are compatible with drift, one computes $X^2 = (x - y)^2/V(x - y)$, using $V(x - y)$ as derived above. Again, a range of N_e values were considered because of the uncertainty regarding this parameter. Because taking drift into consideration lowers the test statistic, it is not necessary to consider alleles with nonsignificant results using the standard test. Two of the alleles showing "significant" differences between years (alleles A_s and B_{45}) also changed by more than can be expected from drift alone, provided N_e was as large as 1,000 (Table 3). Changes for the other three alleles

TABLE 3. Adjusted X^2 values for test of the hypothesis that the changes in allele frequency for *Dacus oleae* during 1966–1967 can be attributed to genetic drift. Asterisks indicate significant chi-square test ($P < 0.05$; $d.f. = 1$). The elapsed number of generations is assumed to be four. Values for $N_e = \infty$ are those for a standard (unadjusted) chi-square test. Alleles not shown exhibited nonsignificant differences in standard tests.

	Gene A			Gene B	
N_e	A_6	A_8	A_s	B_1	B_{45}
100	0.75	0.59	0.85	1.20	0.86
500	2.54	2.00	2.85	2.67	2.84
1,000	3.63	2.85	4.07*	3.16	4.00*
10,000	5.91*	4.64*	6.63*	3.79	6.36*
∞	6.35*	4.99*	7.13*	3.87*	6.81*

TABLE 4. Chi-square values for a test of the hypothesis that allele-frequency changes over eight generations in the moth *Panaxia dominula* can be attributed to genetic drift (data from Fisher and Ford [1947]). Sampling was according to plan I; results are dependent both on effective population size (N_e) and on total population size (N). X^2 values below the stepped line are larger than the critical vlaue ($X^2 = 14.07$; $d.f. = 7$; $P = 0.05$). A) Frequencies were arcsine(square root)-transformed before using (11) to compute the test statistic; B) test statistics computed using untransformed allele frequencies.

	N			
N_e	1,000	3,000	5,000	∞
A.				
400	13.01	11.14	10.86	10.48
500	14.48	12.28	11.96	11.52
600	15.76	13.26	12.90	12.40
700	16.90	14.12	13.72	13.18
800	17.93	14.89	14.46	13.88
1,000	19.77	16.24	15.74	15.08
B.				
700	13.69	11.78	11.50	11.12
800	14.66	12.55	12.25	11.83
1,000	16.42	13.93	13.58	13.10
1,100	—[a]	14.56	14.18	13.67
1,200	—[a]	15.15	14.74	14.20

[a] Test not performed for $N_e > N$.

can be adequately explained by drift unless effective population size was very large ($N_e > 10{,}000$). Thus, there is evidence from one allele at each locus of greater-than-expected change in allele frequencies, although the evidence is not compelling given the relatively large number of tests performed for each locus, all but one of which are independent.

For the second example, we return to Fisher and Ford's (1947) study, which was the first attempt to determine whether observed changes in allele frequency were too large to be reasonably attributed to drift. Fisher and Ford monitored the frequency of the *medionigra* gene in *Panaxia dominula* over eight successive generations; in that period, frequency varied from 4.3% to 11.1% in samples of $S = 117$–986 moths, and estimates of minimum total population size (based on mark-recapture data) ranged from 1,000 to at least 6,000. Fisher and Ford used a test similar to (11) to show that the observed changes were too large to be attributed to drift even in a population of constant size 1,000. However, they did not consider the possibility that effective population size might be less than the total number of individuals, and they also used some approximations in their formulas for $V(x)$ and $V(y)$.

I reanalyzed these allele frequency data using (11) and considering a range of values for N_e (Table 4). The moths were released after examination, so sampling was according to plan I and N was considered as well. I followed Fisher and Ford's (1947) use of the angular transformation of allele frequencies [$\theta = \sin^{-1}(\sqrt{x})$, where x is allele frequency], because nearly all were in the range 0.05–0.1. Under the assumption that $N = N_e = 1{,}000$ each year, the test statistic was $X^2 = 19.77$, $d.f. = 7$, $P < 0.01$ (Table 4A), in good agreement with the value of 19.73 obtained using Fisher and Ford's (1947) method (they reported a value of $X^2 = 20.8064$, but this was based on data in their table 20, in which there is an addition error for column A). However, the test statistic was smaller for $N > 1{,}000$ or $N_e < 1{,}000$, and there are several combinations of N_e and N that yield nonsignificant results ($N_e < 700$, $N > 3{,}000$; $N_e < 500$, $N \geq 1{,}000$; etc.). Given that Fisher and Ford's (1947) estimates of N were 1,000 or more every year and that the ratio N_e/N is fairly small for some natural populations, more information is required before one can rule out drift as a possible explanation for the changes observed.

A further caveat of relevance here is that transformation of the *Panaxia* data results in a larger test statistic than is obtained using the raw data (Table 4B). This result appears to be typical of the angular transformation (Box et al., 1978 pp. 134–135). Because the rejection rate under the null hypothesis for the adjusted test using untransformed data was equal to or higher than the nominal alpha level in the simulations (Fig. 7), use of a more sensitive test would not seem to be desirable. To evaluate the relative merits of using raw versus transformed data in this example, a simulation was performed using plan-I sampling with $N = N_e = 1{,}000$ (Fisher and Ford's [1947] assumed values), $S = 300$ (harmonic mean sample size for 1939–1946), and $P = 0.9$ (approximate population frequency in 1939). Allele frequencies were followed for eight consecutive generations in each of 5,000 replicates. Neither method yielded consistently higher values; overall, 5.00% of the adjusted tests were significant using raw data, and 4.84% were significant using transformed data. In this case, N_e and S were large enough that there was no elevation in type-I error rate by either method, even for $P = 0.9$. Although overall results for the two methods were similar, they lead to different X^2 values in virtually every replicate (as also seen in Table 4), and choice of method to use for a particular data set can affect the conclusions. If the more conservative X^2 values are used (Table 4B; using raw data), one would conclude that allele-frequency changes in *Panaxia* for 1939–1946 were compatible with a drift model, even if effective population size was rather large ($N_e \approx 1{,}000$).

Discussion

The analysis of temporal changes in allele frequency can lead to erroneous conclusions unless care is taken to stipulate the hypothesis being tested. Although it has long been recognized that standard tests (such as the contingency chi-square test) of the equality of allele frequencies are not appropriate for temporally spaced samples, they continue to be used routinely in this context. Given this fact, it is important to have a perspective for interpreting the meaning of "significant" differences. The approach outlined in this paper provides this perspective. Results shown in Figure 5 demonstrate that the probability of a significant test statistic can be reliably estimated for presumptive values of N_e. Knowledge of the true α level of the standard test is necessary if results of such tests are to be used in drawing conclusions about evolutionary forces.

Because the probability of a significant test is largely determined by the ratio of sample size to effective population size, the potential bias in using the standard test will vary considerably with the organism studied and the method of sampling. If S/N_e is small ($S/N_e < 0.1$), then the probability of a significant test is not much greater than the presumed α level even after many generations of genetic drift (Fig. 4), and use of the standard test may not lead to serious problems. However, the probability of an incorrect conclusion increases rapidly as the ratio S/N_e increases, and for many organisms, an appreciable fraction of the effective population size can be sampled (e.g., Gaines et al. [1978] estimated the total size of local *Microtus* populations they sampled to be less than 100). In particular, treatment of data for endangered species will require special care. For such species, it may often be possible to sample essentially the entire population, in which case sample size may equal or exceed effective population size, and use of the standard test can lead to serious errors.

The problems associated with using the standard test are not a consequence of limited sample size and cannot be overcome by taking larger samples; in fact, the reverse is true. With very small samples, ignoring drift may not lead to serious errors, because random changes due to finite N_e are small compared to sampling error. As larger samples are taken, the magnitude of sampling error diminishes, and the test becomes more sensitive to the effects of genetic drift.

If the hypothesis of interest is that observed differences can be satisfactorily explained by stochastic factors alone (genetic drift and sampling error), then the best strategy is to modify the standard test to account for finite population size. If there is a specific alternative hypothesis to be tested, then one of the tests proposed for selection (e.g., Templeton, 1974; Watterson, 1982) may be appropriate. Often, however, there is no specific alternative hypothesis, and a generalized test is most useful. A significant test in this case would imply that some factor(s) other than drift must be responsible for the observed changes (e.g., natural selection, migration, mutation, nonrandom sampling, faulty data). The advantage of such a test is that genetic drift is incorporated into the null hypothesis, rather than being lumped with various other factors as the alternative hypothesis.

The method described here leads to a more generalized adjusted test than has been available to date. Simulation results indicated that the probability of a type-I error using the adjusted test is very close to the nominal $\alpha = 0.05$ level under most conditions. Because the significance level of the adjusted test is only as accurate as the estimates of N_e (plan II) or N and N_e (plan I), results obtained by the methods described in this paper should be regarded as approximations. It is recommended that these methods be used with a range of values for parameters that must be estimated to gain an idea of the conditions under which the conclusions are valid.

It should also be kept in mind that short-term effects of natural selection may be very difficult to detect unless fitness differences among genotypes are large (Nicholas and Robertson, 1976; Watterson, 1982; Pollak, 1983). Failure to reject the null hypothesis that allele-frequency changes can be explained by sampling error and drift is thus not necessarily a powerful indication that other factors are not involved. Such a test, however, can still be very valuable in determining whether the influence of other factors is small enough that they can safely be ignored.

Acknowledgments

I thank J. Felsenstein, R. Kappenman, P. Sampson, P. Smouse, E. Thompson, and J. Wilson for helpful discussions of this topic. J. Wilson and J. Pella critically read a previous version of the manuscript, and two anonymous reviewers provided helpful comments. Part of this research was supported by a National Research Council Research Associateship to the author.

Program Availability

IBM-PC compatible computer programs to perform the analyses described in the text are available; please include a blank, formatted 5.25-inch diskette with the request.

Literature Cited

Appfelbaum, L. I., and A. Blanco. 1985. Temporal variation of allele frequencies in populations of *Akoden dolores* (Rodentia, Cricetidae). Theoret. Appl. Genet. 70:569–572.

Barker, J. S. F., P. D. East, and B. S. Weir. 1986. Temporal and microgeographic variation in allozyme frequencies in a natural population of *Drosophila buzzatii.* Genetics 112:577–611.

Black, W. C., IV, and E. S. Krafsur. 1986. Temporal and spatial trends in allozyme frequencies in house fly populations, *Musca domesticus* L. Theoret. Appl. Genet. 71:673–681.

Bowen, B. S. 1982. Temporal dynamics of microgeographic structure of genetic variation in *Microtus californicus.* J. Mammal. 63:625–638.

Box, G. E. P., W. G. Hunter, and J. S. Hunter. 1978. Statistics for Experimenters. Wiley, N.Y.

Cavener, D. R., and M. T. Clegg. 1981. Temporal stability of allozyme frequencies in a natural population of *Drosophila melanogaster.* Genetics 98: 613–623.

Charlesworth, B., and J. T. Giesel. 1972. Selection in populations with overlapping generations. II. Relations between gene frequency and demographic variables. Amer. Natur. 106:388–401.

Cornejo de Caminos, S., E. H. Bucher, and A. Blanco. 1981. Temporal variations of allele frequencies in the eared dove (*Zenaida auriculata*). Biochem. Genet. 19:1163–1167.

Dobzhansky, Th. 1943. Genetics of natural populations. IX. Temporal changes in the composition of populations of *Drosophila pseudoobscura.* Genetics 28:162–186.

Fisher, R. A., and E. B. Ford. 1947. The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. Heredity 1:143–174.

Franklin, I. R. 1981. An analysis of temporal variation at isozyme loci in *Drosophila melanogaster,* pp. 217–236. *In* J. B. Gibson and J. G. Oakeshott (eds.), Genetic Studies of *Drosophila* Populations. Proceedings of the 1979 Kiola Conference. Australian National University, Canberra, Australia.

Gaines, M. S., and C. J. Krebs. 1971. Genetic changes in fluctuating vole populations. Evolution 25:702–723.

Gaines, M. S., L. R. McClenaghan, Jr., and R. K. Rose. 1978. Temporal patterns of allozymic variation in fluctuating populations of *Microtus ochrogaster.* Evolution 32:723–739.

Gibson, J. B., N. Lewis, M. A. Adena, and S. R. Wilson. 1979. Selection for ethanol tolerance in two populations of *Drosophila melanogaster* segregating alcohol dehydrogenase allozymes. Aust. J. Biol. Sci. 32:387–398.

GYLLENSTEN, U. 1985. Temporal allozyme frequency changes in density fluctuating populations of willow grouse (*Lagopus lagopus* L.). Evolution 39:115–121.

JACOBSON, L. D., R. L. TORBLAA, AND R. H. MORMAN. 1986. Temporal variation in allelic frequencies of sea lamprey ammocoetes. Copeia 1986:199–202.

JOHNSON, M. S., AND R. BLACK. 1984. Pattern beneath the chaos: The effect of recruitment on genetic patchiness in an intertidal limpet. Evolution 38: 1371–1383.

KOEHN, R. K., AND G. C. WILLIAMS. 1978. Genetic differentiation without isolation in the American eel, *Anguilla rostrata.* II. Temporal stability of geographic patterns. Evolution 32:624–637.

KORPELAINEN, H. 1986. Temporal changes in the genetic structure of *Daphnia magna* populations. Heredity 57:5–14.

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 25:454–460.

LEWONTIN, R. C., AND J. KRAKAUER. 1973. Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. Genetics 74: 175–195.

MIHOK, S., W. A. FULLER, R. P. CANHAM, AND E. C. MCPHEE. 1983. Genetic changes at the transferrin locus in the red-backed vole (*Clethrionomys gapperi*). Evolution 37:332–340.

MUELLER, L. D., B. A. WILCOX, P. E. EHRLICH, D. G. HECKEL, AND D. D. MURPHY. 1985. A direct assessment of the role of genetic drift in determining allele frequency variation in populations of *Euphydryas editha.* Genetics 110:495–511.

NEI, M., AND F. TAJIMA. 1981. Genetic drift and estimation of effective population size. Genetics 98: 625–640.

NICHOLAS, F. W., AND A. ROBERTSON. 1976. The effect of selection on the standardized variance of gene frequency. Theoret. Appl. Genet. 48:263–268.

PAMILO, P., AND S. VARVIO-AHO. 1980. On the estimation of population size from allele frequency changes. Genetics 95:1055–1057.

POLLAK, E. 1983. A new method for estimating the effective population size from allele frequency changes. Genetics 104:531–548.

SCHAFFER, H. E., D. YARDLEY, AND W. W. ANDERSON. 1977. Drift or selection: A statistical test of gene frequency change over generations. Genetics 87: 371–379.

SMITH, M. W., M. H. SMITH, AND R. K. CHESSER. 1983. Biochemical genetics of mosquitofish. I. Environmental correlates, and temporal and spatial heterogeneity of allele frequencies within a river drainage. Copeia 1983:182–193.

TEMPLETON, A. R. 1974. Analysis of selection in populations observed over a sequence of consecutive generations. Theoret. Appl. Genet. 45:179–191.

WALL, S., M. A. CARTER, AND B. CLARKE. 1980. Temporal changes of gene frequencies in *Cepaea hortensis.* Biol. J. Linn. Soc. 14:303–317.

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

WATTERSON, G. A. 1982. Testing selection at a single locus. Biometrics 38:323–331.

WILLIAMS, G. C., R. K. KOEHN, AND J. B. MITTON. 1973. Genetic differentiation without isolation in the American eel, *Anguilla rostrata.* Evolution 27: 192–204.

WILSON, S. R. 1980. Analyzing gene-frequency data when effective population size is finite. Genetics 95: 489–502.

Corresponding Editor: P. W. Hedrick

APPENDIX

Generalization to Multiple Alleles.—Procedures are given below for a generalized chi-square test (R samples × K alleles, using either sampling plan) that takes genetic drift into consideration. In such a test, it is necessary to consider covariances of frequencies for different alleles sampled at different times. For illustration, we will consider an example with three alleles (initial frequencies A, B, and C) and three samples (taken in generations $0, j$, and t), but extension to $K > 3$ and $R > 3$ should be obvious. As in the text, subscripted S indicates the number of individuals sampled in a given generation, and we will use a similar notation for total population size (N). Note that only $K - 1$ alleles at a locus are independent, so only A and B are considered below.

The first step is to obtain estimates of the initial frequencies of each allele being considered ($\hat{A}$, $\hat{B}$, etc.). This can be done as described in the text (see "multiple samples") for a single allele. Next, create a vector $\mathbf{P}$ of length $R(K - 1)$ whose elements are differences between (estimated) initial allele frequencies and those observed. For our example, $\mathbf{P} = [A_0 - \hat{A}, A_j - \hat{A}, A_t - \hat{A}, B_0 - \hat{B}, B_j - \hat{B}, B_t - \hat{B}]$. Finally, we construct a covariance matrix of order $R(K - 1) \times R(K - 1)$ and compute the test statistic [with $(R - 1)(K - 1)$ degrees of freedom] as described in the text [Equation (11)].

The diagonal elements of this matrix are variances of allele frequencies in samples taken at different times. These variances are given by (4) or (5) and differ among alleles only in the constant term [$\hat{A}(1 - \hat{A})$, $\hat{B}(1 - \hat{B})$, etc.]. The off-diagonal elements are covariances and can be computed as follows. Consider first sampling plan I. The covariances of A (or B) with itself at different generations were given in the text

$$\mathrm{Cov}(A_0, A_t) = \frac{A(1-A)}{2N_0} \qquad \text{(A1a)}$$

$$\mathrm{Cov}(A_j, A_t) = A(1 - A) \cdot \left[1 - \left(1 - \frac{1}{2N_e}\right)^j \left(1 - \frac{1}{2N_j}\right)\right] \qquad \text{(Alb)}$$

with comparable expressions for allele B. In contrast to (Ala) and (Alb), covariances involving different alleles at the same locus are always negative. If we let A and B represent any two of the $K - 1$ alleles being considered, these terms are as follows:

$$\mathrm{Cov}(A_0, B_0) = -\frac{AB}{2S_0} \tag{Alc}$$

$$\mathrm{Cov}(A_0, B_t) = -\frac{AB}{2N_0} \tag{Ald}$$

$$\mathrm{Cov}(A_j, B_j) = -AB\left[1 - \left(1 - \frac{1}{2N_e}\right)^j\left(1 - \frac{1}{2S_j}\right)\right] \tag{A1e}$$

$$\mathrm{Cov}(A_j, B_t) = -AB\left[1 - \left(1 - \frac{1}{2N_e}\right)^j\left(1 - \frac{1}{2N_j}\right)\right]. \tag{A1f}$$

Results for sampling plan II are simpler, because N is not a factor:

$$\mathrm{Cov}(A_0, A_t) = 0 \tag{A2a}$$

$$\mathrm{Cov}(A_j, A_t) = A(1 - A)\left[1 - \left(\frac{1}{2N_e}\right)^j\right] \tag{A2b}$$

$$\mathrm{Cov}(A_0, B_0) = \frac{AB}{2S_0} \tag{A2c}$$

$$\mathrm{Cov}(A_0, B_t) = 0 \tag{A2d}$$

$$\mathrm{Cov}(A_j, B_j) = -AB\left[1 - \left(1 - \frac{1}{2N_e}\right)^j\left(1 - \frac{1}{2S_j}\right)\right] \tag{A2e}$$

$$\mathrm{Cov}(A_j, B_t) = -AB\left[1 - \left(1 - \frac{1}{2N_e}\right)^j\right]. \tag{A2f}$$

The equivalence of plans I and II for large N can be verified by inserting $N = \infty$ in (A1a), (A1b), (A1d), and (A1f); one obtains (A2a), (A2b), (A2d), and (A2f). As mentioned in the text, if N_e varies over time, the terms $[1 - 1/(2N_e)]^t$ can be replaced by $\Pi[1 - 1/(2N_{e_j})]$, N_{e_j} being the effective population size in the jth generation.