A Sampling Theory of Selectively Neutral Alleles in a Subdivided Population

Elisabeth R. Tillier and G. Brian Golding

Department of Biology, York University, North York, Ontario M3J 1P3, Canada

Manuscript received August 21, 1987
Revised copy accepted March 16, 1988

ABSTRACT

Ewens' sampling distribution is investigated for a structured population. Samples are assumed to be taken from a single subpopulation that exchanges migrants with other subpopulations. A complete description of the probability distribution for such samples is not a practical possibility but an equilibrium approximation can be found. This approximation extracts the information necessary for constructing a continuous approximation to the complete distribution using known values of the distribution and its derivatives in randomly mating populations. It is shown that this approximation is as complete a description of a single biologically realistic subpopulation as is possible given standard uncertainties about the actual size of the migration rates, relative sizes of each of the subpopulations and other factors that might affect the genetic structure of a subpopulation. Any further information must be gained at the expense of generality. This approximation is used to investigate the effect of population subdivision on Watterson's test of neutrality. It is known that the infinite allele, sample distribution is independent of mutation rate when made conditional on the number of alleles in the sample. It is shown that the conditional, infinite allele, sample distribution from this approximation is also independent of population structure and hence Watterson's test is still approximately valid for subdivided populations.

THE sampling distribution of alleles from a finite population was determined by EWENS (1972). This theory assumes that a finite sample of selectively neutral alleles is obtained from a single finite and randomly mating population. This population undergoes random genetic drift at a rate determined by its effective size (N_e), and the alleles in this population mutate to new allelic forms at a rate μ. These are the sole processes determining the distribution of alleles. The formulas for this distribution permit a complete description of the sampling properties of alleles from a population and have answered many questions concerning hypothesis testing. For example, the distribution demonstrates that, to estimate $N_e\mu$, the frequency array of alleles in the sample does not contain any more information than does the number of distinct alleles in the sample.

Since its original derivation, the distribution has been extensively studied and its applicability extended. This has included tests of neutrality (*e.g.*, EWENS 1974; EWENS and GILLESPIE 1974; WATTERSON 1978), generalizations of the distribution (*e.g.*, KINGMAN 1977), infinite sized populations with random selection (GILLESPIE 1977), and the addition of small selective effects on the alleles (*e.g.*, WATTERSON 1977). In addition, the theory has had a generally stimulating effect on various areas within population genetics.

Almost every population that exists in nature does not mate at random and one of the common factors which prevents random mating is isolation by distance (WRIGHT 1943). The idea that populations may be partially isolated and that they may be related by migration was realized early by WRIGHT (1940). The extensive work within this area has been reviewed by FELSENSTEIN (1977).

Methods of estimating gene flow and population structure are available (WRIGHT 1951; WEIR and COCKERHAM 1984; SLATKIN 1981, 1985) but there can be many sources of bias for these estimates, rendering their accuracy uncertain [see SLATKIN (1985) for discussion]. Knowledge of population properties that are independent of population structure is therefore desirable.

It is the purpose of this note to examine some of the effects of population structure on the sampling distribution of alleles. A general solution of the sampling distribution in subdivided populations is not possible. It is possible to find a numerical solution for simple cases (*i.e.*, when the numbers of genes sampled is small and when the number of subpopulations considered is small) but this is impracticable for larger samples. We show here that an approximation can be used to extend EWENS sampling distribution to a subpopulation that exchanges migrants with other populations. This approximation incorporates all information which is generally applicable to any subpopulation and is accurate for many situations.

It is further shown that WATTERSON's (1978) test of neutrality is still approximately valid for subdivided populations. This is because any simple infinite allele approximation of the sampling distribution will be independent of both mutation and migration rates when conditional on the number of alleles sampled. Thus, any conditional test will incorporate all of the properties of subdivided populations that hold for any pattern of migration between populations.

METHOD

The model: Let the population be subdivided into s subpopulations. No restrictions will be placed on s in the following. Only samples which originate from a single subpopulation will be considered at length but we must also examine the effects of migration between the various subpopulations. Let m_{lj} designate the probability that a randomly chosen allele from subpopulation l originated in the previous generation from subpopulation j. Designate by

$$m = \sum_{\substack{j=1 \\ j \neq l}}^{s} m_{lj}$$

the total probability that a randomly chosen allele from subpopulation l is a migrant in the previous generation from any other subpopulation. Let the lth subpopulation consist of N_l diploid individuals which undergo random mating internally within each subpopulation. The total population size is $N_T = \sum_{l=1}^{s} N_l$. Generations are assumed to be discrete and non-overlapping. Mutation occurs at a rate μ per gamete per generation and there are only K distinct alleles possible at a locus.

Coefficients: To describe the subpopulations a probability approach will be taken similar to that used in GOLDING (1984). This involves setting up a system of variables to describe the various samples that are possible and then deriving a system of recursion equations that can be solved for the probabilities of these samples.

Many different samples are possible and samples might not only contain different allelic types but also may originate from more than one subpopulation. To distinguish between possible samples a vector notation is used. Designate possible samples by the vector $S_l = \{n_1, n_2, \ldots, n_K\}$, where the subscript l indicates samples from the lth subpopulation. The quantity n_i designates the number of alleles sampled that fall into the ith identity class. Those n_i that are zero can be ignored and the number of nonzero n_i define the number of alleles in the sample (designated by k). For example, when $n_1 = 2$ and $n_i = 0$ for all $i > 1$, this represents a sample of size $n = 2$ from population l with both genes identical in state ($k = 1$). When $n_1 = 1$, $n_2 = 1$ and $n_i = 0$ for all $i > 2$, this represents a sample of $n = 2$ genes that are not identical in state ($k = 2$). Suppressing zero elements, these samples can be written as $\{2\}$ and $\{1, 1\}$. The numbering of alleles is arbitrary and is intended only to keep track of the identity relationships between alleles in the samples and not to designate a particular allele. For example, the sample $\{1, 2, 3\}$ is equivalent to the sample $\{3, 0, 2, 1\}$. All samples are assumed to be drawn at random, without replacement. Any sample from a single population will consist of k ($k \leq K$) distinct alleles, with a total of $n = \sum_{i=1}^{K} n_i$ genes sampled.

RESULTS

Recursion Equation: A general recursion equation can be found for the sampling probabilities using standard probability arguments. The general system of simultaneous equations for each subpopulation for arbitrary samples is not difficult to find but it is quite lengthy. Considering samples from only a single population permits some simplification.

Let $E(S_l)$ designate the probability of some sample S with k alleles from subpopulation l. The expectation is taken over conceptually replicate populations. Let $E(S_l[n_i - 1])$ designate the probability of a sample S but with the number of genes of the ith allelic type decreased by 1. Similarly, let $E(S_l[n_i - 1, n_j + 1])$ designate the probability of the sample S with the number of genes of the ith allelic type reduced by one, and the number of genes of the jth allelic type increased by one. Let $E(S_l^j[n_{i-1}^1])$, be the probability of the same sample from subpopulation l with one fewer gene of the ith allelic type and an allele of the ith type sampled from another subpopulation j. Ignoring terms smaller than $1/2N_l$, μ and m_{lj} the recursion equation for $E(S_l)$ is

$$\begin{aligned} E'(S_l) = &\left(1 - \frac{n(n-1)}{4N_l} - n\mu - nm\right)E(S_l) \\ &+ \sum_{i=1}^{k} \frac{n_i(n_i-1)}{4N_l} E(S_l[n_i - 1]) \\ &+ \sum_{i=1}^{k} \mu\delta(n_i - 1) \sum_{\substack{j=1 \\ j \neq i}}^{k+1} \frac{K - (k + \delta(n_j) - 1)}{K - 1} \\ &\times E(S_l[n_i - 1, n_j + 1]) \qquad (1) \\ &+ \sum_{i=1}^{k} n_i\mu\delta(\delta(n_i - 1)) \sum_{\substack{j=1 \\ j \neq i}}^{k+1} \frac{1}{K-1} \\ &\times E(S_l[n_i - 1, n_j + 1]) \\ &+ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq l}}^{s} n_i m_{lj} E(S_l^j[n_{i-1}^1]) \end{aligned}$$

where $E(S_l[n_{K+1} + 1]) = 0$. The prime indicates the value in the next generation and $\delta(x)$ is a Dirac delta function such that $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$. Equation 1 is derived using probability arguments as in GOLDING (1984).

The first term in this recursion relationship is the probability that there has been no change from one generation to the next. That is, none of the n genes sampled are duplicate copies of one gene from the previous generation, none of the genes are mutants and none are migrants. The second term calculates the probability that any pair of genes are identical in state due to a common ancestor in the previous generation. In this case, the identity between these two alleles is assured and the probability that the remaining alleles are identical is given by the probability that the original sample of alleles, less one, are identical. The third and fourth terms correspond to the probability that one of the alleles is a mutant since the last generation but that the sample specified by S_l is still obtained. The last term is the probability of obtaining the sample when any one of the genes (of the ith allelic type) is a migrant from the jth subpopulation.

For example, the homozygosity can be calculated by considering the probability of picking two identical genes from population l. In this case $n = 2$, $k = 1$, $n_1 = 2$, $n_i = 0$ for all $i > 1$ and the recursion relationship is

$$E'(\{2\}) = \left(1 - \frac{1}{2N_l} - 2\mu - 2m\right) E(\{2\}) + \frac{1}{2N_l} + 2\mu \frac{1}{K-1} E(\{1, 1\}) + \sum_{\substack{j=1 \\ j \neq l}}^{s} 2m_{lj} E\left(\begin{Bmatrix} 1 \\ 1 \end{Bmatrix}\right).$$

The first term on the right measures the probability of no change. The second term is the probability that these two genes were made identical by descent. The third is the probability that the two identical genes were made identical in state by mutation. The last term is the probability that one of the genes is a recent migrant from subpopulation j; where $E(\{{}^{1}_{1}\})$ indicates a sample of one gene from subpopulation N and one gene of the same allelic type from subpopulation j. This equation was first determined by MALÉCOT (1948). That this method could be extended up to samples of four genes from a single population was first shown by EWENS and KIRBY (1975).

The known solution: It is not possible to solve, in closed form, the general system of equations for arbitrary migration rates between an arbitrary number of subpopulations. Attempts to solve some simplified systems using computer-based algebraic manipulation languages suggests that even for small samples, the corresponding solution consists of the ratio of two very high degree polynomials. However, when $m = 0$ we know from WATTERSON (1976) that the resulting distribution is given by

$$\hat{E}(S_l) = \frac{K!}{(K-k)!} \frac{\prod_{i=1}^{k} \prod_{j=0}^{n_i - 1} \left(j + \frac{\theta}{K-1}\right)}{\prod_{i=0}^{n-1} \left(i + \frac{K\theta}{K-1}\right)} \tag{2}$$

where $\theta = 4N_l\mu$ and where the caret designates the expectation at equilibrium. As demonstrated previously in GOLDING (1984), Equation 2 does not include the number of ways of obtaining a particular sample of alleles. Thus to make this result identical to EWENS (1972), multiply Equation 2 by the number of distinct combinations of k alleles falling into n_i classes.

The derivatives: Although a general solution is not practicable, it is possible to ask how this solution changes when $m \neq 0$. But in order to do this, we do not want to place any conditions on m_{lj}, the pattern of migration. Interestingly, this is indeed possible by considering the derivatives of $\hat{E}(S_l)$ around $N_l m = 0$. Multiplying Equation 1 by $4N_l$ and solving at equilibrium gives

$$\begin{aligned} n(n - 1 + \theta + 4N_l\mathrm{m})\hat{E}(S_l) &= \sum_{i=1}^{k} n_i(n_i - 1)\hat{E}(S_l[n_i - 1]) \\ &+ \sum_{i=1}^{k} \theta\delta(n_i - 1) \sum_{\substack{j=1 \\ j \neq i}}^{k+1} \frac{K - (k + \delta(n_j) - 1)}{K - 1} \\ &\times \hat{E}(S_l[n_i - 1, n_j + 1]) \\ &+ \sum_{i=1}^{k} n_i\theta\delta(\delta(n_i - 1)) \sum_{\substack{j=1 \\ j \neq i}}^{k+1} \frac{1}{K-1} \\ &\times \hat{E}(S_l[n_i - 1, n_j + 1]) \\ &+ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq l}}^{s} n_i 4N_l m_{lj} \hat{E}(S_l^i[n_{i-1}^1]). \end{aligned} \tag{3}$$

Taking the derivative of this equation with respect to $4N_l m$, gives

$$\begin{aligned} n(n - 1 + 4N_l m + \theta)\frac{d\hat{E}(S_l)}{d(4N_l m)} + n\hat{E}(S_l) &= \sum_{i=1}^{k} n_i(n_i - 1)\frac{d\hat{E}(S_l[n_i - 1])}{d(4N_l m)} \\ &+ \sum_{i=1}^{k} \theta\delta(n_i - 1) \sum_{\substack{j=1 \\ j \neq i}}^{k+1} \frac{K - (k + \delta(n_j) - 1)}{K - 1} \\ &\times \frac{d\hat{E}(S_l[n_i - 1, n_j + 1])}{d(4N_l m)} \\ &+ \sum_{i=1}^{k} n_i\theta\delta(\delta(n_i - 1)) \sum_{\substack{j=1 \\ j \neq i}}^{k+1} \frac{1}{K-1} \end{aligned} \tag{4}$$

$$\times \frac{d\hat{E}(S_l[n_i - 1, n_j + 1])}{d(4N_lm)}$$

$$+ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq l}}^{s} n_i 4N_l m_{lj} \frac{d\hat{E}(S_l^i[n_{i-1}^1])}{d(4N_lm)}$$

$$+ \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq l}}^{s} n_i \frac{d4N_l m_{lj}}{d(4N_lm)} \hat{E}(S_l^i[n_{i-1}^1]).$$

The last term is necessary since

$$m = \sum_{\substack{j=1 \\ j \neq l}}^{s} m_{lj}.$$

Evaluating this equation at $m = 0$ ($m_{lj} = 0$, for all j and thus the subpopulation l is completely isolated) yields a great simplification. Since samples from different subpopulations are independent when there is no migration, the value of $\hat{E}(S_l^i[n_{i-1}^1])$ is known. It is either $1/K$ (if $n_i > 1$) or $1 - (k - 1)/K$ (if $n_i = 1$), times the probability of the remainder of the sample from subpopulation l. Including these values makes the recursion relationship (4) independent of the population structure.

A solution can be found by a multiple series of proofs by induction; first for $\hat{E}(\{n\})$, then for $\hat{E}(\{n, 1\})$ and so on. Continuing this, the derivatives of the probabilities (evaluated at $m = 0$, $m_{lj} = 0$) can be found as

$$\left.\frac{d\hat{E}(S_l)}{d(4N_lm)}\right|_{m=0} = \frac{K!}{(K-k)!} \frac{\prod_{i=1}^{k} \prod_{j=0}^{n_i-1} \left(j + \frac{\theta}{K-1}\right)}{\prod_{i=0}^{n-1} \left(i + \frac{K\theta}{K-1}\right)} \left[\frac{1}{K} \sum_{i=1}^{k} \sum_{j=0}^{n_i-1} \left(j + \frac{\theta}{K-1}\right)^{-1} - \sum_{i=1}^{n-1} \left(i + \frac{K\theta}{K-1}\right)^{-1}\right]. \quad (5)$$

The solution can be confirmed by substitution into Equation 4. Note that $d\hat{E}(S_l)/d(4N_lm)\,|_{m=0} = (1 - 1/K)d\hat{E}(S_l)/d(\theta)\,|_{m=0}$. The rate of change of the probabilities with respect to migration and mutation differ by a factor which is dependent only on the number of alleles.

It is apparent when $4N_lm \gg 1$, such that $4N_lm_{lj} \gg 1$ for all l, j, that the solution to the recursion Equation 1 is again well known. In this case the probabilities are given by Equation 2 with N_l replaced with N_T. We have still more knowledge of the probabilities since their derivatives with respect to N_lm asymptotically approach zero when $N_lm_{lj} \gg 1$ for all l, j. This is because the rate of change of the probabilities slows as the amount of migration increases.

It is not possible to obtain any more knowledge of the general pattern of the sampling distribution. To show this, consider the requirements for determining the second derivatives of the probabilities around $N_lm = 0$. Evaluation of the second derivatives requires knowledge of the derivatives and probabilities of samples with alleles chosen from more than one subpopulation. To find $d^2\hat{E}(S_l)/d(4N_lm)^2$ requires values for

$$d\hat{E}(S_l^i[n_{i-1}^1])/d(4N_lm)$$

which depends on the functional form of

$$\hat{E}(S_l^i[n_{i-1}^1]).$$

This would in turn depend on the particular migration rates between subpopulations and could easily be different among the subpopulations. For example, if a large circular system of subpopulations is considered then, with very small migration rates, the probability of identity between alleles chosen from neighboring subpopulations will be much larger than that probability when alleles are chosen from subpopulations on opposite sides of the circle. This would not be true for subpopulations that exchange migrants at random regardless of physical distance between the subpopulations. Hence, the derivatives would necessarily be different for these two models.

At the other extreme of migration, the derivatives are known to approach zero when $N_lm \gg 1$ but the rate of approach to this limit is dependent on the migration structure of the population. To show this an example has been constructed. The system of equations necessary to describe a sample of up to five alleles from one subpopulation was derived under the infinite allele model of KIMURA and CROW (1964). This subpopulation is one of either four circularly arranged populations or four populations with random migration between each. These systems were then solved numerically. For circular migration, a system of 32 linear equations is required. The migration structure is defined with $m_{12} = m_{23} = m_{34} = m_{41} = m$ and $m_{lj} = 0$ otherwise. The solutions of these equations are compared with the numercial solutions for four subpopulations that have equal migration rates between each of the subpopulations ($m_{lj} = m/3$ for all l, j; leading to a system of 16 linear equations). In both cases, all subpopulations are assumed to be of equal sizes.

The first derivative of the probabilities $\hat{E}(\{5\})$ and $\hat{E}(\{4, 1\})$ around $N_lm = 0$ with $\theta = 0.1$ have the values given by Equation 4 (-1.59 and 0.16, respectively) and are indeed zero when $N_lm \gg 1$. The behavior of the probabilities was found for both the circular population structure and the random migration scheme when the rate of migration is large. Both models of migration show different behaviors when $N_lm \gg 1$ for both $\hat{E}(\{5\})$ and $\hat{E}(\{4, 1\})$. The derivatives

with respect to $1/4N_lm$ when $4N_lm$ is large are 0.09 and -0.02 with the random migration scheme and are 0.1 and -0.03 with the circular migration scheme respectively for $\hat{E}(\{5\})$ and $\hat{E}(\{4, 1\})$. Similarly, in the circular migration scheme, the second derivatives of $\hat{E}(\{5\})$ and $\hat{E}(\{4, 1\})$, evaluated at $N_lm = 0$ with $\theta = 0.1$, are 16.0 and -1.76, respectively, while for random migration they are 12.12 and -1.43. Therefore, the behavior of the probabilities when $N_lm \gg 1$ and the second derivative evaluated at $N_lm = 0$ are both dependent on the actual migration structure.

A PADÉ APPROXIMATION

In total there are four pieces of information available about samples from a subpopulation which apply generally. These are (1) the sample distribution when $m = 0$, (2) the derivatives with respect to N_lm when $m = 0$ (as derived above), (3) the sample distribution when $N_lm_{lj} \gg 1$, and (4) the derivatives with respect to N_lm_{lj} when $N_lm_{lj} \gg 1$. Each of these is independent of the population structure. These pieces of information can be combined into a single continuous approximation with one restriction. Here, we impose the limitation that m_{lj} and m are generally of the same order of magnitude. In this way, when the m_{lj}'s are very large, they can be approximated, with little loss of accuracy, by setting them equal to an equivalent large m value. An approximation can be found which interpolates all of these properties. Of several approximations tried, a Padé approximation, the ratio of two polynomials (ATKINSON 1978; p. 206), appeared to work best. The Padé that combines the four properties listed above is given by

$$\hat{E}(S_l) = \frac{A(C - A) + BC(4N_lm)}{C - A + B(4N_lm)} \qquad (6)$$

where A is given by Equation 2, B is given by Equation 5 and C is given by Equation 2 with N_l replaced with N_T. This approximation has value A when $4N_lm = 0$, a value of C when $4N_lm \gg 1$ (equivalently, $m_{lj} \gg 1$ for all l, j), a derivative (with respect to $4N_lm$) equal to zero when $4N_lm \gg 1$ and equal to B when $4N_lm = 0$.

CONDITIONAL SAMPLE PROBABILITY

The test of neutrality developed by WATTERSON (1978) makes use of the sampling probabilities conditional on k, the number of distinct alleles in the sample. In light of this, it was interesting to look at the behavior of the conditional probabilities in the approximation when migration is present.

It was found that the conditional probabilities of samples given k, are independent of the migration rate as $K \to \infty$. This result is easily shown. When K is infinite, the Padé approximation can be written as the product of two terms: one that is dependent on the actual distribution of the alleles in the sample (the n_i's) but independent of $4N_lm$, and another that is independent of the n_i's and contains $4N_lm$. When the probability is made conditional on k, the term dependent on $4N_lm$ is a constant and thus can be cancelled. When K is finite, the n_i's can not be factored and this result is no longer true.

DISCUSSION

The approximation given here extracts all of the information that is generally applicable to any structured population. Of course, more information could be extracted which would be specific for a particular model (or set of models) with known migration rates, but it seems clear that relatively small changes in these parameters could easily alter the resulting solution. Given the uncertainties with which such quantities are measured, the approximation given here may be just as accurate.

It is necessary to compare the Padé approximation to the exact solution under different circumstances. The exact solution was obtained by deriving the required systems of equations and then solving numerically. We have restricted comparisons to the infinite allele model and samples of up to ten genes of one allelic type (sampled from one of two populations), samples of eight genes of two allelic types (sampled from one of two populations) or samples of two genes (from one of ten populations). These systems require up to 134 equations and the number of equations increases exponentially with the number of subpopulations and genes sampled.

The probabilities of the samples {3}, {5} and {10} are given in Figure 1 with $\theta = 0.1$ when sampling is from one of two populations of equal size. For comparison the Padé approximations using Equation 6 are also given. It is evident from Figure 1 that the approximation is very accurate over a wide range of values for the migration rates (the results do not change qualitatively if different mutation rates are used). Indeed, for $\hat{E}(\{3\})$ and $\hat{E}(\{5\})$, the approximation and the true probability are visually indistinguishable. The error between the actual probability and the approximation is greatest with intermediate values of migration. This is sensible, since this is the region where the least amount of information is available to construct the approximation.

It also appears that the error increases as the number of genes sampled (n) increases. This may, in part, be due to the greater difference in the size of the probabilities when $4N_lm = 0$ and $4N_lm \gg 1$. If the probability is relatively large when $4N_lm = 0$ and relatively small when $4N_lm \gg 1$, the approximation must span a greater range and the error may depend on the magnitude of the largest versus the smallest value in the approximation.

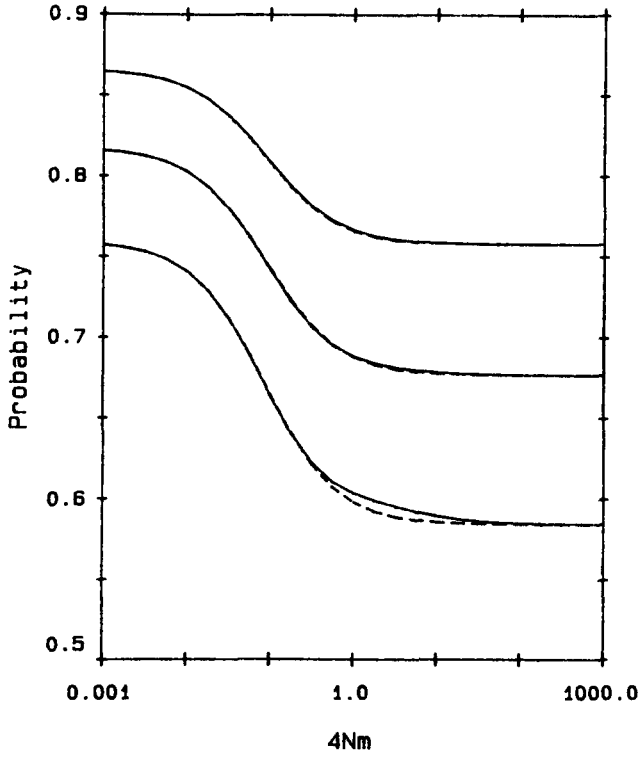

FIGURE 1.—The probabilities of sampling three genes ({3}, upper curve), five genes ({5}, middle curve) or ten genes ({10}, lower curves) all of which are identical ($K \rightarrow \infty$ and $\theta = 0.1$). All genes are chosen from one of two subpopulations. The Padé approximation to these probabilities is given as a dashed line.

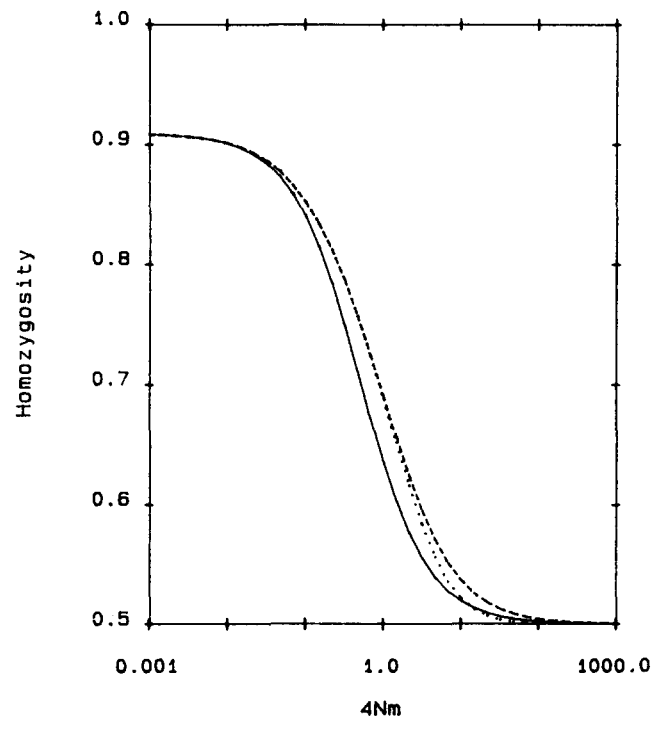

FIGURE 2.—The homozygosity (probability of sample {2}) in one of ten subpopulations of equal size as given by the Padé approximation (solid line), with a random pattern of migration (solid line), with a circular pattern (dashed line) and with a linear array of subpopulations (dotted line; genes are sampled from one of the middle subpopulations in the array of ten). $K \rightarrow \infty$ and $\theta = 0.1$.

The accuracy of the Padé approximation was also checked for an increased number of subpopulations with several migration schemes. Figure 2 shows the homozygosity within one of ten subpopulations. All subpopulations are of equal size and the migration pattern is either random, circular or linear. The Padé approximation is quite good and agrees perfectly with the random migration pattern; less well for other migration patterns. The differences in the sample probabilities that can be caused by different patterns of migration are obvious from the figure. Nevertheless, each pattern of migration has all of the properties used to construct the approximation. From this figure, it can be seen that in a real situation, even a minor amount of migration between populations that are not immediately adjacent in the case of circular or linear migration could greatly change the sample probabilities. This migration might be so small as to be impossible to detect. It should also be noted, that such a close fit between the approximation and the true homozygosity with random migration is not generally found for other probabilities.

Each of the above probabilities are relatively well behaved functions of mutation and migration. For these probabilities, the Padé approximation has been found to be reasonably accurate. However, there are also classes of sample probabilities that are less well behaved and less accurately approximated.

One such class of samples occurs when $A > C$ and $B > 0$ (or $C > A$ and $B < 0$), where A, B and C are defined after Equation 6. In this case, there is a positive value of $4N_fm$ for which the denominator of the Padé equation is zero. This causes a local singularity in the approximation (Figure 3). This singularity will never occur when there is only one allelic type in a sample (such as for the homozygosity). For these samples B is always negative and A greater than C.

These singular points can be predicted *a priori* from the values of A, B and C and the problem can be avoided by considering a different approximation. One way to do this is to force the second derivative at $4N_fm = 0$ to be zero, as is commonly done with spline approximations. This restricted Padé approximation is given by

$$\hat{E}(S_i) = \frac{A + CD(4N_fm) + CD^2(4N_fm)^2}{1 + D(4N_fm) + D^2(4N_fm)^2} \qquad (7)$$

$$D = \frac{B}{C - A}.$$

This approximation is usually a more well behaved function of $4N_fm$ than the Padé approximation, as

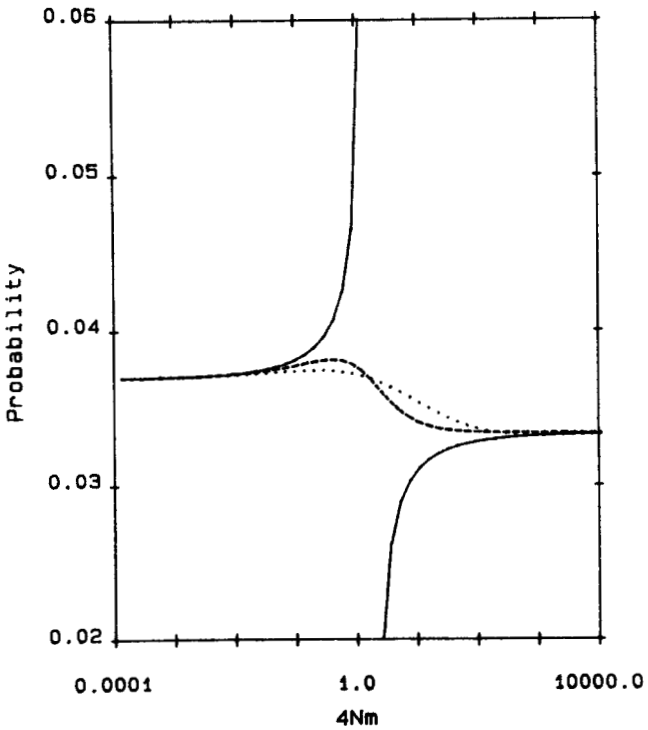

FIGURE 3.—The probability of sample {5, 1} (dotted line), the Padé approximation (solid line) and the restricted Padé approximation (dashed line), for a sample chosen from one of two populations ($K \rightarrow \infty$ and $\theta = 0.5$).

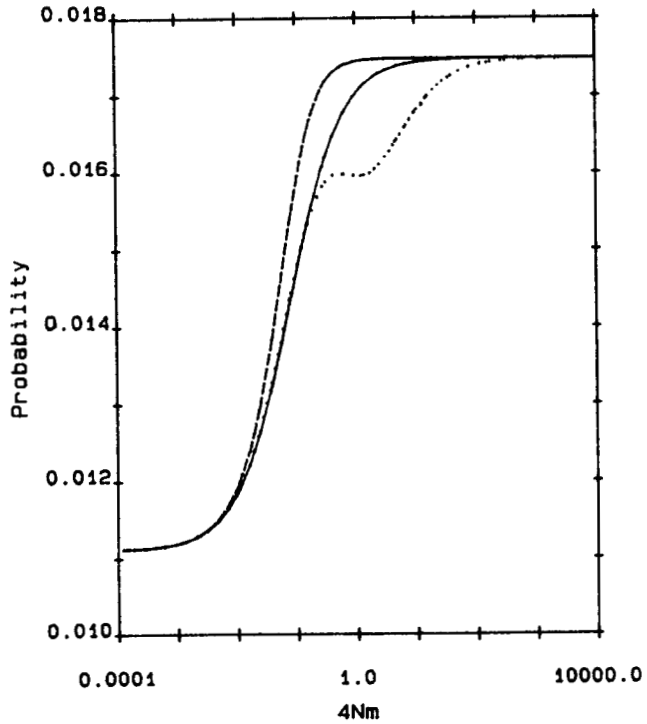

FIGURE 4.—The probability of sample {7, 1} (dotted line), the Padé approximation (solid line) and the restricted Padé approximation (dashed line), for a sample chosen from one of two populations ($K \rightarrow \infty$ and $\theta = 0.1$).

shown in Figure 3. Here the probability of the sample {5, 1} is given when K is infinite and $\theta = 0.5$ for the Padé approximation, for this restricted Padé approximation and for the actual probability. As can be seen each approximation is accurate only near the values of $4N_1m$ from which information was extracted.

The error of the Padé approximation may also increase when the actual sampling probability is not monotonic with increasing migration rate. An example of this case is given in Figure 4, for the sample {7, 1} with $\theta = 0.1$. The nonmonotonic shape of this curve is dependent on the actual migration pattern between subpopulations. This type of behavior can not be predicted *a priori* and hence can not be corrected. Because of this dependence on migration, the nature of this nonmonotonic behavior (even its presence) may change with small changes in the migration rates.

The Padé approximations given here are only two of many possible functions that could be written which include the four known pieces of information on the probability distribution with migration. Any function that includes these four properties could potentially be used to approximate the distribution. Most reasonable approximations will yield sampling probabilities, conditional on k (the number of alleles in the sample), that are constant with migration under the infinite allele model. This is due to the fact that the term containing migration will factor when the probability is made conditional on k.

The accuracy of using a constant to approximate the conditional probabilities was investigated in a population split into two subpopulations. Since the WATTERSON test uses cumulative conditional probabilities, it was necessary to observe the behavior of the error as the conditional probabilities are summed. The probabilities of the samples {7, 1}, {6, 2}, {5, 3} and {4, 4} were made conditional on $k = 2$ (by dividing by the sum of these probabilities) and then summed in order of increasing homozygosity. The result is shown in Figure 5. The relative error is less than 10% and it decreases relatively quickly as different probabilities are added. Again, the error is largest for intermediate values of migration, as expected.

Since WATTERSON's test uses conditional probabilities and since the approximations given here incorporate all available information, we conjecture that Watterson's test (or any other test of conditional probabilities) is the best that can be constructed with complete generality.

MARUYAMA (1977) has analyzed the distribution of gene frequencies within an arbitrary subpopulation which forms part of either a circular or a lattice stepping stone model. He constructed an approximation to these distributions by first determining the

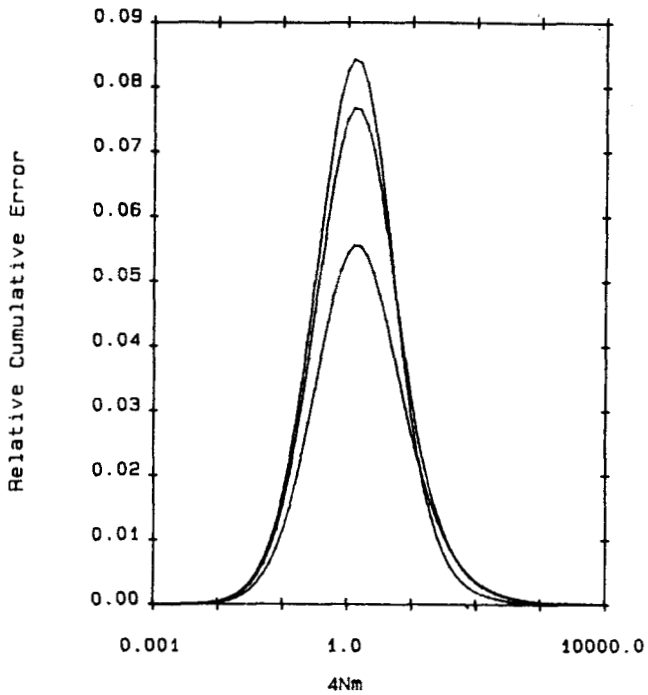

FIGURE 5.—The relative error in approximating the cumulative probabilities conditional on $k = 2$ for the samples {4, 4}, {5, 3} and {6, 2} ($K \rightarrow \infty$ and $\theta = 0.1$). The samples originate from one of two subpopulations with equal population size. The top curve corresponds to the relative error of the conditional probability of sample {4, 4}, the second is that of the cumulated conditional probabilities of the {4, 4} and {5, 3}, and the bottom curve is that of the cumulated conditional probabilities of the {4, 4}, {5, 3} and {6, 2} samples.

local homozygosity within one subpopulation. A value of θ is calculated from this homozygosity using the relationship between θ and homozygosity in the absence of migration. The probabilities of samples are then approximated using standard formulas without migration but with the altered value of θ. This approximation and that given here are numerically very similar but distinct. In addition, the approximation given here is independent of a theoretical knowledge of the local homozygosity.

MARUYAMA (1974) was also able to demonstrate the remarkable property that some population parameters are independent of the geographical structure of the population. We have used here those properties of a single subpopulation that are also independent of structure. Because each property is itself independent of the patterns of migration, the approximation is also independent of population structure (with a K allele model of mutation).

The probabilities derived here are known to be related to allele frequency moments. To an approximation on the order of $1/N$, there is a one to one relationship between these probabilities and the expected frequency of allelic combinations. This was shown by MALÉCOT (1948), who noted that $\hat{E}(\{2\})$ is approximately $E\left(\sum_{i=1}^{K} p_i^2\right)$, where p_i is the frequency of the ith allele in the population. In general

$$E(\{n_1, n_2, n_3, \ldots\}) = E\left(\sum_{i}^{K}\sum_{j\neq i}^{K}\sum_{k\neq i,j}^{K} \ldots p_i^{n_1} p_j^{n_2} p_k^{n_3} \ldots\right) + 0\left(\frac{1}{N}\right)$$

where each of the sums extends over all alleles different from those preceding (for more applications of similar relationships see COCKERHAM and WEIR 1973). The above results therefore apply to all equilibrium gene frequency moments and to any quantity that can be expressed as a function of gene frequencies.

We would like to thank B. S. WEIR, J. H. GILLESPIE and several reviewers for their very helpful comments in the preparation of this manuscript. This work was supported by Natural Sciences and Engineering Research Council of Canada grant U0336 to G.B.G. Additional support was provided by the Canadian Institute for Advanced Research.

LITERATURE CITED

ATKINSON, K. E., 1978 *An Introduction to Numerical Analysis.* John Wiley and Sons, New York.

COCKERHAM, C. C., and B. S. WEIR, 1973 Descent measures for two loci with some applications. Theor. Popul. Biol. **4:** 300–330.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

EWENS, W. J., 1974 A note on the sampling theory for infinite alleles and infinite sites models. Theor. Popul. Biol. **6:** 143–148.

EWENS, W. J., and J. H. GILLESPIE, 1974 Some simulation results for the neutral allele model with interpretations. Theor. Popul. Biol. **6:** 35–57.

EWENS, W. J., and K. KIRBY, 1975 The eigenvalues of the neutral alleles process. Theor. Popul. Biol. **7:** 212–220.

FELSENSTEIN, J., 1976 The theoretical population genetics of variable selection and migration. Annu. Rev. Genet. **10:** 253–280.

GILLESPIE, J. H., 1977 Sampling theory for alleles in a random environment. Nature **266:** 443–445.

GOLDING, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KINGMAN, J. F. C., 1977 The population structure associated with the Ewens sampling formula. Theor. Popul. Biol. **11:** 274–283.

MALÉCOT, G., 1948 *Les Mathématiques de l'Hérédité.* Masson et Cie. Paris.

MARUYAMA, T., 1974 A simple proof that certain quantities are independent of the geographical structure of population. Theor. Popul. Biol. **5:** 148–154.

MARUYAMA, T., 1977 Stochastic problems in population genetics. In: *Lecture Notes in Biomathematics,* Vol. 17. Springer-Verlag, New York.

SLATKIN, M., 1981 Estimating gene flow in natural populations. Genetics **16:** 97–159.

SLATKIN, M., 1985 Rare alleles as indicators of gene flow. Evolution **39:** 53–65.

WATTERSON, G. A., 1976 The stationary distribution of the infinitely-many neutral alleles diffusion model. J. Appl. Prob. **13:** 639–651.

WATTERSON, G. A., 1977 Heterosis or neutrality? Genetics **85:** 789–814.

WATTERSON, G. A., 1978 The homozygosity test of neutrality. Genetics **88:** 405–417.

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

WRIGHT, S., 1940 Breeding structure of populations in relation to speciation. Am. Nat. **74:** 232–248.

WRIGHT, S., 1943 Isolation by distance. Genetics **28:** 114–138.

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

Communicating editor: B. S. WEIR