Theoretical Population Biology **53**, 166–174 (1998)
Article No. TP971355

TPB

Segregating Sites in Wright's Island Model

John Wakeley[1]
Department of Biological Sciences, Rutgers University, New Jersey

Received April 10, 1997

Expressions for the expectation and variance of the number of segregating sites in samples from an island model of population subdivision are derived. For small samples, an arbitrary number of demes can be accommodated. Results for larger samples are derived under the assumption of an infinite number of demes. However, simulations indicate that the latter results will hold quite well for the finite-island model in many cases. A new estimator of the population migration rate is proposed and is shown to outperform the widely used pairwise method. © 1998 Academic Press

1. INTRODUCTION

Migration among subpopulations, or demes, shapes the genetic structure of populations and species (Slatkin, 1985; Slatkin, 1987b). Thus, it has figured prominently in evolutionary theory, *e.g*. in Wright's (1977) shifting balance theory. Wright's (1931) island model, which first included an infinite, then later a finite (Maruyama, 1970; Latter, 1973) number of demes, is the most commonly used model to describe gene flow in subdivided populations. This paper presents expressions for the expectation and variance of the number of segregating, or polymorphic, sites in samples from the island model. These expressions can be applied to molecular sequence data, such as DNA sequence data.

The number of segregating sites, S, is a valuable descriptor of genetic variation, which, together with theoretical predictions, can be used to make inferences about population history. For example, in a single Wright–Fisher population, S can be used to estimate the fundamental parameter θ, or four times the effective population size times the neutral mutation rate (Watterson, 1975). Estimates of θ based on S have better statistical properties, *e.g*. smaller standard errors, than estimates made from the average number of differences between pairs of sequences in a sample (Tajima, 1983). It is demonstrated below that the same is true of estimates of the population migration rate in the island model.

Coalescent theory (Kingman, 1982a; Kingman, 1982b; Tajima, 1983; Hudson, 1983a; Tavaré, 1984) has reshaped population genetics because it focuses on the history of a sample, rather than characteristics of the population as a whole, and because it provides a convenient means of examining the statistical properties of that history. The coalescent, together with the assumption of neutral, infinite sites mutation (Kimura, 1969), greatly simplifies the derivation of the expectation and variance of S; see Hudson's (1990) thorough review. Namely, it allows us to consider genealogical and mutational processes separately.

The genealogy of a sample traces the history of common ancestor or coalescent events back to the common ancestor of the entire sample. Let t be the total length of the genealogy in generations and u the neutral mutation rate per gene per generation. It is assumed that the number of mutations occurring over the entire genealogy is Poisson distributed with mean ut. Then,

$$E[S]=uE[t] \tag{1}$$

and

$$\mathrm{Var}[S]=uE[t]+u^2\,\mathrm{Var}[t]. \tag{2}$$

If the diploid effective population size is N, then time is typically measured in units of $2N$ generations. In this

[1] E-mail: jwakeley@rci.rutgers.edu.

0040-5809/98 $25.00

case, t is replaced by $T = t/(2N)$ and u is replaced by $\theta/2 = 2Nu$ in (1) and (2).

The moments of T are known for several models of a population, but not for the island model. The best-known case is a sample of size n from a neutral Wright–Fisher population, *i.e.* without any subdivision:

$$E[T] = \sum_{i=1}^{n-1} \frac{2}{i}, \qquad \mathrm{Var}[T] = \sum_{i=1}^{n-1} \left(\frac{2}{i}\right)^2. \tag{3}$$

Using (1) and (2), these give the familiar results:

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i} \tag{4}$$

and

$$\mathrm{Var}[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \tag{5}$$

(Watterson, 1975). Expressions (3) through (5) represent the expectations and variances over all possible histories of a sample.

2. THEORY

The coalescent is modelled as a Markov chain, with a state space that encompasses all possible sample configurations. Assume that a total of R states are possible and that P_{ij} is the probability of transition from state i to state j in one time unit (usually one generation). These P_{ij} are generally considered to be small, so the time back to a change in state, *i.e.* a coalescent or migration event, is approximately exponentially distributed with parameter

$$P_{i*} = \sum_{j=1,\, j \neq i}^{R} P_{ij}, \tag{6}$$

and the probability that this event is a transition from state i to state j is P_{ij}/P_{i*} (Hudson, 1983b).

After Kaplan *et al.* (1988) and Notohara (1990), the following recursions give the expectation and variance of the total length, T_i, of all the branches in the genealogy of a sample of type i:

$$E[T_i] = \frac{n}{P_{i*}} + \sum_{j=1,\, j \neq i}^{R} \frac{P_{ij}}{P_{i*}} E[T_j] \tag{7}$$

and

$$\mathrm{Var}[T_i] = \left(\frac{n}{P_{i*}}\right)^2 + \sum_{j=1,\, j \neq i}^{R} \frac{P_{ij}}{P_{i*}} (\mathrm{Var}[T_j] + E[T_j]^2) - \left(\sum_{j=1,\, j \neq i}^{R} \frac{P_{ij}}{P_{i*}} E[T_j]\right)^2. \tag{8}$$

A sample from a subdivided population consists of a number of sequences, n, taken from a number of different demes, d, which may be different than the total number of demes, D. Let n_i, $1 \leqslant i \leqslant n$, be the number of demes from which i sequences were sampled. Then, $\sum_{i=1}^{n} n_i = d$ and $\sum_{i=1}^{n} in_i = n$. As is standard, n is assumed to be much smaller than the deme size, N ($n \ll N$) and some results given below also assume that $n \ll D$.

2.1. Arbitrary D, Small Samples

Let $t(n_1, n_2, ..., n_n)$ represent the total length of the genealogy of the sample with configuration $(n_1, n_2, ..., n_n)$. Two genes can assume either of two possible sample configurations, (2, 0) or (0, 1), and the configuration of their common ancestor can only be (1). Under the finite island model, each gene has a probability m of having immigrated from one of the $D-1$ other demes in the previous generation. It is assumed that m is small enough and N large enough that terms of $O(m^2)$, $O(N^{-2})$, and smaller can be ignored. If i genes are sampled from a single deme, the probability of a coalescent event, *i.e.* that two of them are descended from a common ancestor in the previous generation, is equal to $i(i-1)/(4N)$. Thus, the probability of transition from (0, 1) to (2, 0) is equal to $2m$, the probability of transition from (2, 0) to (0, 1) is equal to $2m/(D-1)$, and the probability of transition from (0, 1) to (1) is equal to $1/(2N)$.

Then, (7) gives

$$E[t(0,1)] = \frac{2}{2m + 1/(2N)} + \frac{2m}{2m + 1/(2N)} E[t(2,0)] \tag{9}$$

and

$$E[t(2,0)] = \frac{2}{2m/(D-1)} + E[t(0,1)], \tag{10}$$

since $E[t(1)] = 0$ by definition. These then lead to the well-known results: $E[t(0,1)] = 4ND$ and $E[t(2,0)] = 4ND + (D-1)/m$ (Li, 1976; Slatkin, 1987a, Strobeck, 1987).

If we measure time, now T, in units of $2ND$ generations, and if we let $M = 2NmD/(D-1)$, then (9) and (10) become

$$E[T(0,1)] = \frac{2}{2M(D-1)+D} + \frac{2M(D-1)}{2M(D-1)+D} E[T(2,0)] \qquad (11)$$

and

$$E[T(2,0)] = \frac{2}{2M} + E[T(0,1)], \qquad (12)$$

which give $E[T(0,1)] = 2$ and $E[T(2,0)] = 2 + 1/M$, as expected. The parameter, M, is a natural choice to measure migration rates, as it is estimated easily using F_{ST}-based methods and since D is generally unknown (Hudson *et al.*, 1992). Also, Takahata and Nei (1984) point out that two formulations of the finite island model, with slightly different definitions of the migration rate, m, are made interchangeable by the factor $D/(D-1)$. If $\theta = 4NDu$ and $S(n_1, n_2, ..., n_n)$ is the number of segregating sites in the sample $(n_1, n_2, ..., n_n)$, then (1) gives $E[S(0,1)] = \theta$ and $E[S(2,0)] = \theta + \theta/(2M)$. The variances are obtained similarly, using (8) and (2).

With this recursive method, we can generate expressions for larger and larger samples. For $n = 3$

$$E[T(0,0,1)] = 3, \qquad (13)$$

$$E[T(1,1,0)] = 3 + \frac{1}{M}, \qquad (14)$$

$$E[T(3,0,0)] = 3 + \frac{3}{2M}, \qquad (15)$$

and for $n = 4$

$$E[T(0,0,0,1)] = \frac{11}{3} - \frac{M(D-1)}{3[(2M+3)(M+1)D+2M]}, \qquad (16)$$

$$E[T(1,0,1,0)] = \frac{11}{3} + \frac{1}{M} - \frac{(2M+3)D-2M}{6[(2M+3)(M+1)D+2M]}, \qquad (17)$$

$$E[T(0,2,0,0)] = \frac{11}{3} + \frac{1}{M} + \frac{(2M+3)(M+1)D + M(2M+5)}{3(2M+1)[(2M+3)(M+1)D+2M]}, \qquad (18)$$

$$E[T(2,1,0,0)] = \frac{11}{3} + \frac{3}{2M} + \frac{M(2M+3)}{3(2M+1)[(2M+3)(M+1)D+2M]}, \qquad (19)$$

$$E[T(4,0,0,0)] = \frac{11}{3} + \frac{11}{6M} + \frac{M(2M+3)}{3(2M+1)[(2M+3)(M+1)D+2M]}. \qquad (20)$$

Results for $n = 5$ can also be obtained (not shown) and may be useful, but this method is not going to give general expressions for arbitrary sample sizes.

However, these results for samples of two, three, four, and five genes suggest that the following approximation may be appropriate in some cases:

$$E[T(n_1, n_2, ..., n_n)] \approx 2\left(\sum_{i=1}^{n-1} \frac{1}{i} + \frac{1}{2M} \sum_{i=1}^{d-1} \frac{1}{i}\right). \qquad (21)$$

Takahata (1991) suggested a corresponding expression, his Eq. (13) for the total depth of a genealogy as an interpolation between results for a high migration limit and a low migration limit of the finite island model.

Figure 1 shows the relative error when (16) is approximated using (21), which in this case is equal to 11/3, as a function of M and for four different values of D. With a sample of four genes, the approximation is quite accurate. Plots of (17) through (20) give results of essentially the same form and magnitude, except that (21) sometimes overestimates and other times underestimates the true value. While (21) is within 1% of (16) for all values of M and D and is often much closer, the approximation is at its worst for intermediate values of M, which are the ones of most biological interest. The dependence on D is weak, especially when D is large.

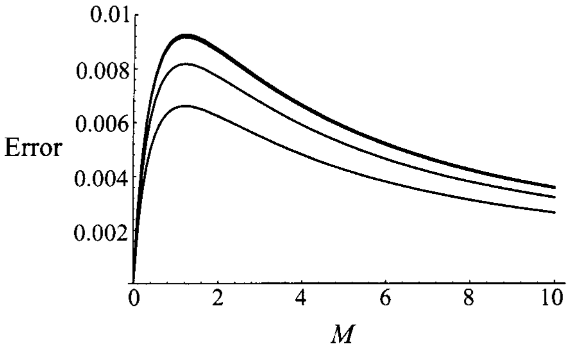

FIG. 1. The relative error of using (21) to approximate (16), the extent to which (21) overestimates (16) as a fraction of the value of (16), plotted as a function of M. The four curves represent different values of D: from bottom to top, 4, 10, 100, 1000. The curves for $D=100$ and $D=1000$ are nearly indistinguishable.

2.2. Large-D Approximation

In the limit as D approaches infinity, (16) through (18) converge on values that differ from those that would be obtained using (21). This is reflected in Fig. 1. Equations (19) and (20), however, converge on (21) as the number of demes increases. In fact, as D goes to infinity, (21) holds exactly for the terminal sample configurations, $(n-2, 1, 0, 0, ...)$ and $(n, 0, 0, ...)$ for any given n. Equation (7) gives

$$\begin{aligned} E[T(n-2, 1, 0, 0, ...)] &= \frac{n}{P_*} + \frac{M(n-2)}{P_*} E[T(n-3, 0, 1, 0, 0, ...)] \\ &+ \frac{M(n-2)(n-3)}{P_*} E[T(n-4, 2, 0, 0, ...)] \\ &+ \frac{2M[D-(n-1)]}{P_*} E[T(n, 0, 0, ...)] \\ &+ \frac{D}{P_*} E[T(n-1, 0, 0, ...)], \end{aligned} \tag{22}$$

where

$$\begin{aligned} P_* = M(n-2) + M(n-2)(n-3) \\ + 2M[D-(n-1)] + D \end{aligned} \tag{23}$$

and

$$E[T(n, 0, 0, ...)] = \frac{1}{M(n-1)} + E[T(n-2, 1, 0, 0, ...)], \tag{24}$$

As D increases relative to n, (22) becomes

$$\begin{aligned} E[T(n-2, 1, 0, 0, ...)] &= \frac{2M}{2M+1} E[T(n, 0, 0, ...)] \\ &+ \frac{1}{2M+1} E[T(n-1, 0, 0, ...)], \end{aligned} \tag{25}$$

so that, substituting into (24), and since $E[T(1)]=0$, we have

$$E[T(n, 0, 0, ...)] = 2\left(1+\frac{1}{2M}\right) \sum_{i=1}^{n-1} \frac{1}{i}, \tag{26}$$

which is equivalent to (21) with $n=d$. Putting (26) in (25) shows that (21) holds also for the sample $(n-2, 1, 0, 0, ...)$.

Under this large-D approximation—just as (22) became (25)—the history of any sample is greatly simplified. There are so many demes relative to the sample size that, for every sample except $(n, 0, 0, ...)$, only two types of transitions occur with any appreciable frequency: (1) coalescent events within demes, and (2) migration events to demes that contain no ancestors. Thus, the ancestors spread out and/or coalesce until they reach the configuration $(n', 0, 0, ...)$, where $d \leqslant n' \leqslant n$, and (26) applies. Because n/P_* in (7) and (8) goes to zero as D increases relative to n, it takes a negligible amount of time for the sample to reach this terminal configuration. Thus, a modified (7) becomes

$$\begin{aligned} E[T(n_1, n_2, ..., n_n)] \\ = 2\left(1+\frac{1}{2M}\right) \sum_{n'=d}^{n} P[n' \mid (n_1, n_2, ..., n_n)] \sum_{i=1}^{n'-1} \frac{1}{i}, \end{aligned} \tag{27}$$

where $P[n' \mid (n_1, n_2, ..., n_n)]$ is the probability that there are $n'-d$ migration events on the way to the terminal configuration $(n', 0, 0, ...)$. Similarly, it can be shown that

$$\text{Var}[T(n, 0, 0, ...)] = 4\left(1+\frac{1}{2M}\right)^2 \sum_{i=1}^{n-1} \frac{1}{i^2} \tag{28}$$

and

$$\begin{aligned} \text{Var}[T(n_1, n_2, ..., n_n)] \\ = 4\left(1+\frac{1}{2M}\right)^2 \left\{ \sum_{n'=d}^{n} P[n'] \left[\sum_{i=1}^{n'-1} \frac{1}{i^2} + \left(\sum_{i=1}^{n'-1} \frac{1}{i} \right)^2 \right] \right. \\ \left. + \left[\sum_{n'=d}^{n} P[n'] \sum_{i=1}^{n'-1} \frac{1}{i} \right]^2 \right\}, \end{aligned} \tag{29}$$

where $P[n']$ is an abbreviation for $P[n' \mid (n_1, n_2, ..., n_n)]$. Then, (1) and (2) can be used to get $E[S(n_1, n_2, ..., n_n)]$ and $\mathrm{Var}[S(n_1, n_2, ..., n_n)]$.

2.2.1. *A sample from a single deme.* In single-deme samples of two (Slatkin, 1987a; Strobeck, 1987) and three —see (13), above—the expectation of T is the same as in a population of size ND without subdivision. However, this is not true for samples of four; see (16), above, and Tajima's (1989) two-deme simulation study. Given an expression for $P[n' \mid (n_1, n_2, ..., n_n)]$, (27) and (29) can be used to compute means and variances for any sample under this large-D approximation. The case of n sequences sampled from a single deme $(n_n = 1)$ provides some background necessary to calculate $P[n' \mid (n_1, n_2, ..., n_n)]$ for a sample of arbitrary configuration.

Coalescent events within the single deme and migration events out of that deme, but not into demes that contain any ancestral genes, characterize the (instantaneous) history of the sample $(n_n = 1)$. That is, they determine the value of n' in the terminal configuration $(1 \leqslant n' \leqslant n, 0, 0, ...)$. The relative probability that the first event is a migration event is Mn, and the relative probability that the first event is a coalescent event is $n(n-1)/2$. Both migration and coalescence decrease the number of ancestral lineages remaining in the deme by one, so after the first event these become $M(n-1)$ and $(n-1)(n-2)/2$. This is true regardless of whether the first event was a migration or a coalescence. After $n-1$ events have occurred, the sample is in configuration $(n', 0, 0, ...)$, where (26) and (28) apply.

Thus, the passage of the sample into the terminal configuration, $(n', 0, 0, ...)$ is described by the expansion of

$$\left[Mn + \frac{n(n-1)}{2}\right] \times \left[M(n-1) + \frac{(n-1)(n-2)}{2}\right] \cdots [2M+1]. \qquad (30)$$

The term involving M^i signifies a history of i migration events and $n-1-i$ coalescent events, *i.e.* that $n' = i+1$. This is given by

$$\frac{n!}{2^{n-1}} |s(n, i+1)| (2M)^i, \qquad (31)$$

where $s(n, i+1)$ are Stirling numbers of the first kind; *e.g.*, see Abramowitz and Stegun (1964). Dividing (31) by (30) leads to

$$P[n' \mid (n_n = 1)] = \frac{|s(n, n')| (2M)^{n'}}{(2M)_{(n)}}, \qquad (32)$$

where $(2M)_{(n)} = (2M)(2M+1) \cdots (2M+n-1)$. Thus, the distribution of n' for the sample $(n_n = 1)$ is identical to the distribution of the number of alleles in Ewens' sampling formula (Ewens, 1972; Karlin and McGregor, 1972), but with infinite alleles mutation is replaced by infinite demes migration.

Using (27), the expected total length of the genealogy of the sample $(n_n = 1)$ is

$$E[T(n_n = 1)] = 2\left(1 + \frac{1}{2M}\right) \sum_{n'=2}^{n} \frac{|s(n, n')| (2M)^{n'}}{(2M)_{(n)}} \sum_{i=1}^{n'-1} \frac{1}{i}. \qquad (33)$$

The expected number of segregating sites is $\theta/2$ times (33). The variance of $T(n_n = 1)$ can be obtained similarly using (29) and (32). Putting in $n = 4$, (33) reduces to the limit of (16) as D approaches infinity. Figure 2 plots the relative error of using (21) to approximate (33) and shows that in contrast to Fig. 1, where the relative error is always less than 1%, (21) overestimates (33) substantially as the sample size increases.

An integral representation of (33) is also possible. Since

$$\sum_{j=1}^{i-1} \frac{1}{j} = \int_0^1 \frac{1 - x^{i-1}}{1 - x}\, dx, \qquad (34)$$

we have

$$E[T(n_n = 1)] = 2\left(1 + \frac{1}{2M}\right) \int_0^1 (1-x)^{-1} \left[1 - \frac{(2Mx)_{(n)}}{x(2M)_{(n)}}\right] dx. \qquad (35)$$

2.2.2. *Arbitrary sample configurations.* The general sample, $(n_1, n_2, ..., n_n)$ is made up of d single-deme

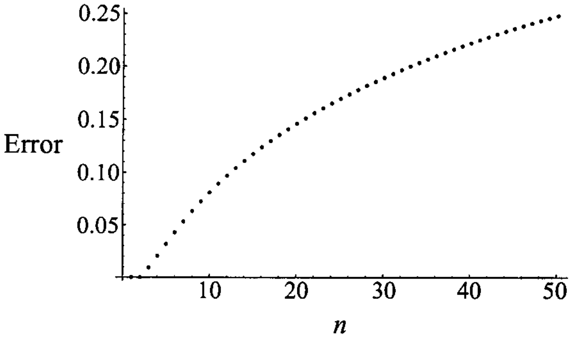

FIG. 2. The relative error of using (21) to approximate (33) as a function of the sample size, n. M is assumed to be equal to one. The third point from the left above $(n = 4)$ is essentially identical to the point on the curve for $D = 1000$ in Fig. 1 where $M = 1.0$.

samples. Under the large-D approximation, the history of this sample involves coalescent events within each deme and migration events out of the sampled demes into unoccupied demes until the configuration $(d \leqslant n' \leqslant n, 0, 0, \ldots)$ is reached. The relative rates of migration and coalescence for a deme containing i lineages are Mi and $i(i-1)/2$, as before, and these do not depend on the numbers of lineages in other demes. Thus, the passage of the sample $(n_1, n_2, \ldots, n_n)$ into the terminal configuration $(d \leqslant n' \leqslant n, 0, 0, \ldots)$ is described by the expansion of a product of term(s) (30):

$$\prod_{i=1}^{n}\left[\frac{i!}{2^{i-1}}(2M+i-1)(2M+i-2)\cdots(2M+1)\right]^{n_i}. \tag{36}$$

The term involving M^i signifies a history of i migration events and $n-d-i$ coalescent events, i.e. that $n' = i + d$. Then, analogous to (32),

$$P[n' \mid (n_1, n_2, \ldots, n_n)] = \frac{A(n')(2M)^{n'}}{\prod_{i=1}^{n}[(2M)_{(i)}]^{n_i}}, \tag{37}$$

where $A(n')$ is the coefficient of $(2M)^{n'}$ in the expansion of the denominator of (37). Inserting (37) into (27) and taking advantage of (34), we have

$$E[T(n_1, n_2, \ldots, n_n)] = 2\left(1+\frac{1}{2M}\right)\int_0^1 (1-x)^{-1} \times \left\{1 - \frac{\prod_{i=1}^{n}[(2Mx)_{(i)}]^{n_i}}{x\prod_{i=1}^{n}[(2M)_{(i)}]^{n_i}}\right\} dx. \tag{38}$$

This can be shown to reproduce all of the expressions derived using the method of Section 2.1 when the limit of those expressions is taken as D goes to infinity. Similarly, putting (37) in (29) gives $\mathrm{Var}[T(n_1, n_2, \ldots, n_n)]$. Last, (1) and (2) give the expectation and variance of the number of segregating sites in the sample.

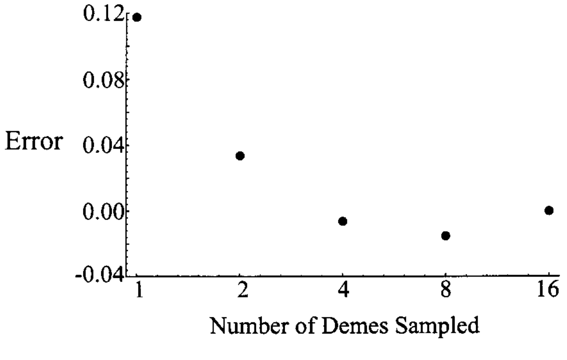

FIG. 3. The relative error of using (22) to approximate (38) when 16 genes are sampled evenly from different numbers of demes; *i.e.*, the first point on the left is when all 16 genes are sampled from a single deme, the second is when eight sequences are taken from two different demes, and the last point on the right is when a single gene is sampled from each of 16 demes.

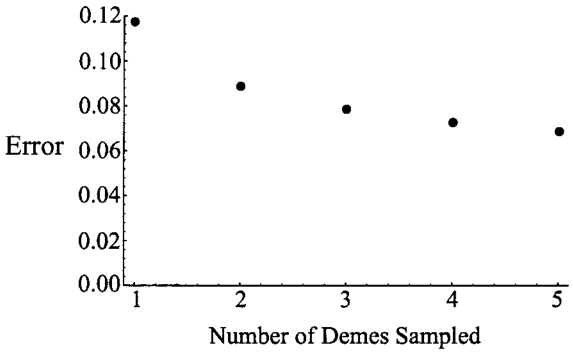

FIG. 4. The relative error of using (21) to approximate (38) as additional demes are sampled, and 16 sequences are taken from each deme. Thus, the leftmost point above is identical to the leftmost point in Fig. 3.

Figure 3 plots the error of using (21) as an approximation to the general expression, (38) when 16 sequences are sampled evenly from different numbers of demes. The error decreases rapidly as the sample is spread out across more demes. This is expected because (21) holds exactly for the sample $(n, 0, 0, \ldots)$. Figure 3 reiterates another point: sometimes (21) gives an overestimate and sometimes it gives an underestimate. In Fig. 3, the error decreases both because the sample is taken from a greater number of demes and because, as this happens, the sample size per deme decreases (see Fig. 2). Thus, Fig. 4 plots the error of using (21) as more demes are sampled, but the sample size per deme remains constant. This reduces the error also, but less rapidly than when the per-deme sample size also decreases.

3. DISCUSSION

The accuracy of the approximation (21) in predicting the expected total length of the genealogy of a sample is impressive. Figure 2 shows clearly, though, that (21) cannot be relied upon in all cases. However, the similarity in form of (21) to Watterson's (1975) formula, given here as (4), provides guidance in designing a sampling strategy for the measurement of DNA polymorphism. We can draw a parallel to the case of a single, randomly mating population (Watterson, 1975; Tajima, 1983) and surmise that samples of five to 10 sequences can provide accurate estimates of levels of polymorphism. Equation (21) also implies that, if our only interest is in overall levels of polymorphism, we need not sample more than one gene per deme. Of course, if our aim is to contrast within and among-dame polymorphisms, for instance in estimating M, then we should take multiple gene copies from several different demes.

The main new results presented here (Section 2.2) were derived under the assumption of an infinite number of demes. In fact, these formulas appear very accurate even for moderate values of D. Coalescent simulations, using the same programs as Hudson *et al.* (1992)—Dick Hudson has kindly made these available—demonstrate this. Figure 5 shows the mean over simulation replicates of the number of segregating sites in a sample of 10 sequences, either taken singly from each of 10 different demes or all sampled from a single deme. The average value of S is plotted as the total number of demes, D, varies between one and 100. It appears from Fig. 5 that the error of the large-D approximation depends on

$$E(n')=\sum_{i=1}^{n} n_i \sum_{j=0}^{i-1} \frac{2M}{2M+j}, \tag{39}$$

which can be obtained using (37). As M approaches zero, $E(n')$ approaches d, and as M gets large, $E(n')$ approaches n. In Fig. 5a, $E(n')$ is equal to 10 and in Fig. 5b, (39) gives $E(n')=4$. In both 5a and 5b the relative error of the large-D approximation is less than 1%, as long as D is about three times $E(n')$.

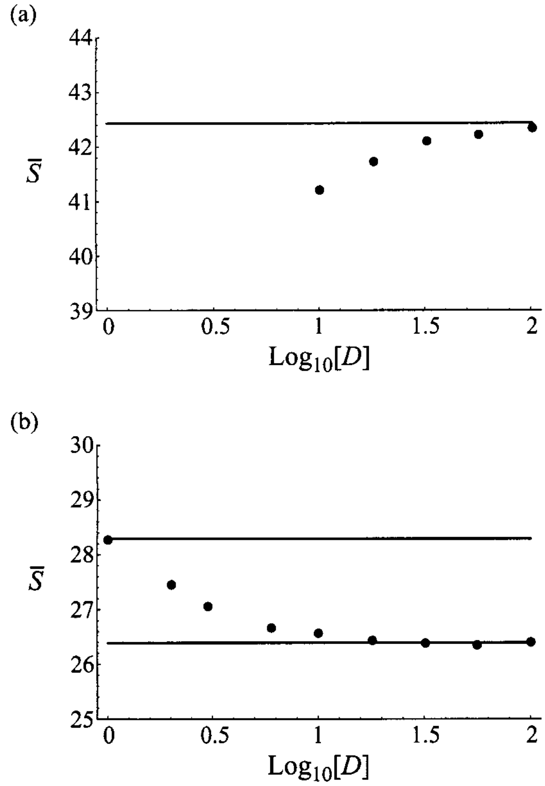

FIG. 5. The average over simulation replicates of the number of segregating sites in samples of 10 sequences, plotted against the total number of demes in the population. Two configurations were modelled: (a) 10 sequences sampled singly from 10 different demes; and (b) 10 sequences sampled from a single deme. In both cases, M was equal to one. On 100,000 replicates were done for each value of D: 1, 2, 3, 6, 10, 18, 32, 56, and 100, provided $D \geqslant d$. The line in (a) shows expectation given by (26). In (b), the top line shows the expectation given by (21) and the bottom line shows the expectation given by (33).

The present results provide a better way to estimate M than the currently used F_{ST}-based, pairwise method. In the pairwise method, the average numbers of pairwise differences within, H_w, and between, H_b, populations are calculated from the data. The expectations of these quantities for any sample are identical to $E[S(0,1)]=\theta$ and $E[S(2,0)]=\theta(1+1/(2M))$, so that

$$\hat{M}_F=\frac{1}{2(H_b/H_w-1)}. \tag{40}$$

Thus, it is assumed that $E[H_b/H_w]=E[H_b]/E[H_w]$, which does not depend on θ, and $\hat{M}_F$ is the value that equates theoretical predictions with the observed ratio.

Equations (26) and (33) allow for a multisequence generalization of the within versus among deme comparison embodied by H_b/H_w. Let

$$S_w=\sum_{i=1}^{d} S^{[i]}, \tag{41}$$

where $S^{[i]}$ is the number of segregating sites in the sample from the ith deme, and

$$S_a=\frac{1}{n^{[1]}n^{[2]}\cdots n^{[d]}}\sum_{i=1}^{n^{[1]}}\sum_{j=1}^{n^{[2]}}\cdots\sum_{k=1}^{n^{[d]}} S^{[ij\cdots k]}, \tag{42}$$

where $n^{[i]}$ is the number of sequences sampled from the ith deme and $S^{[ij\cdots k]}$ is the number of segregating sites among d genes, $ij\cdots k$, one from each deme's sample. Thus, S_w is the number of segregating sites within demes and S_a is the average number of segregating sites among demes. Then $\hat{M}_S$ solves

$$\frac{E[S_a]}{E[S_w]}=\frac{S_a}{S_w}, \tag{43}$$

where, using (26) and (33) together with (41) and (42),

$$\frac{E[S_a]}{E[S_w]} = \frac{\sum_{i=1}^{d-1} 1/i}{\sum_{i=1}^{d} (1/(2M)_{(n^{[i]})}) \sum_{j=2}^{n^{[i]}} |s(n^{[i]}, j)| \sum_{k=1}^{j-1} 1/k}. \tag{44}$$

Given the complexity of (44), $\hat{M}_S$ must be obtained numerically.

Figure 6 compares the distributions of $\hat{M}_S$ and $\hat{M}_F$ when applied to simulated data. In the simulations, five DNA sequences were assumed to have been sampled from each of five different demes. Thus, $d = 5$ and $n^{[i]} = 5$, $1 \leqslant i \leqslant d$, are used in (44). The total population consisted of 100 demes. One hundred thousand replicates were done for each of two different values of M: 0.1 and 1.0. Figure 6 shows that $\hat{M}_S$ performs better than $\hat{M}_F$; more of the distribution is centered around the true value of M and, consequently, less is found in the long upper tail. This is analogous to the improvement in estimates of θ made using segregating sites versus pairwise differences in samples from a single, Wright–Fisher population. Both $\hat{M}_F$ and $\hat{M}_S$ fail some of the time: $\hat{M}_F$ when $H_w \geqslant H_b$, and $\hat{M}_S$ when, loosely speaking, S_w is too big, relative to S_a. This occurs with greater frequency as M increases, but it appears that $\hat{M}_S$ may be calculable slightly more often than $\hat{M}_F$. Although the differences may be slight, the results shown in Fig. 6 recommend the use of $\hat{M}_S$ over $\hat{M}_F$. A computer program that calculates $\hat{M}_S$ is available upon request.

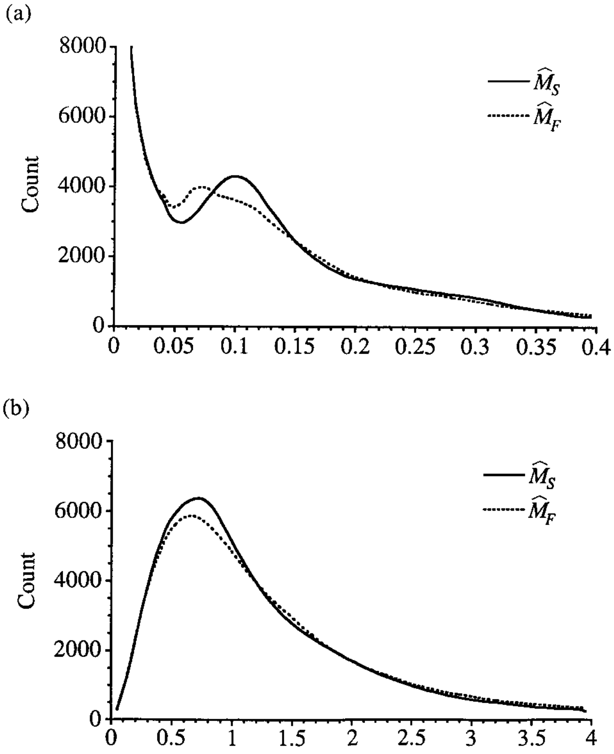

FIG. 6. The distribution of $\hat{M}_S$ and $\hat{M}_F$ in 100,000 simulated datasets for each of two different values of M: 0.1 in (a) and 1.0 in (b). In both cases, θ/D was equal to one. In (a), 6.7% of the distribution of $\hat{M}_F$ lies above 0.4, compared to 5.6% for $\hat{M}_S$, and $\hat{M}_F$ and $\hat{M}_S$ failed zero and six times, respectively. In (b), 9.1% of the distribution of $\hat{M}_F$ lies above 4.0, compared to 7.9% for $\hat{M}_S$, and $\hat{M}_F$ and $\hat{M}_S$ failed 1513 and 1350 times, respectively. The medians of $\hat{M}_F$ and $\hat{M}_S$ were 0.091 and 0.096 in (a) and 1.104 and 1.035 in (b). The smoothed curves trace the numbers of values of $\hat{M}_S$ and $\hat{M}_F$ in 40 evenly spaced intervals along the horizontal axis.

ACKNOWLEDGMENTS

I thank Monty Slatkin, Jody Hey, Magnus Nordborg, and two anonymous reviewers for helpful comments. Special thanks go to Bob Griffiths for his suggestion to use (34), which greatly improved the work. This work was supported by National Research Service Award, GM 17745-02, from the National Institutes of Health.

REFERENCES

Abramowitz, M., and Stegun, I. A. 1964. "Handbook of Mathematical Functions," Dover, New York.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles, *Theoret. Popul. Biol.* **3**, 87–112.

Hudson, R. R. 1983a. Testing the constant-rate neutral allele model with protein sequence data, *Evolution* **37**, 203–217.

Hudson, R. R. 1983b. Properties of a neutral allele model with intragenic recombination, *Theoret. Popul. Biol.* **23**, 183–201.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, *in* "Oxford Surveys in Evolutionary Biology" (D. J. Futuyma and J. Antonovics, Eds.), Vol. 7, Oxford Univ. Press, Oxford.

Hudson, R. R., Slatkin, M., and Maddison, W. P. 1992. Estimation of levels of gene flow from dna sequence data, *Genetics* **132**, 583–589.

Kaplan, N. L., Darden, T., and Hudson, R. R. 1988. Coalescent process in models with selection, *Genetics* **120**, 819–829.

Karlin, S., and McGregor, J. 1972. Addendum to a paper of W. Ewens, *Theoret. Popul. Biol.* **3**, 113–116.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to the steady flux of mutations, *Genetics* **61**, 893–903.

Kingman, J. F. C. 1982a. The coalescent, *Stochastic Process. Appl.* **13**, 235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations, *J. Appl. Prob.* **19A**, 27–43.

Latter, B. D. H. 1973. The island model of population differentiation: a general solution, *Genetics* **73**, 147–157.

Li, W.-H. 1976. Distribution of nucleotide difference between two randomly chosen cistrons in a subdivided population: the finite island model, *Theoret. Popul. Biol.* **10**, 303–308.

Maruyama, T. 1970. Effective number of alleles in a subdivided population, *Theoret. Popul. Biol.* **1**, 273–306.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population, *J. Math. Biol.* **29**, 59–75.

Slatkin, M. 1985. Gene flow in natural populations, *Ann. Rev. Ecol. Syst.* **16**, 393–430.

Slatkin, M. 1987a. The average number of sites separating DNA sequences drawn from a subdivided population, *Theoret. Popul. Biol.* **32**, 42–49.

Slatkin, M. 1987b. Gene flow and the geographic structure of natural populations, *Science* **236**, 787–792.

Strobeck, C. 1987. Average number of nucleotide differences in an sample from a single subpopulation: a test or population subdivision, *Genetics* **117**, 149–153.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations, *Genetics* **105**, 437–460.

Tajima, F. 1989. DNA polymorphism in a subdivided population: The expected number of segregating sites in the two-population model, *Genetics* **123**, 229–240.

Takahata, N. 1991. Genealogy of neutral genes and spreading of selected mutations in a geographically structured population, *Genetics* **129**, 585–595.

Takahata, N., and Nei, M. 1984. F_{ST} and G_{ST} statistics in the finite island model, *Genetics* **107**, 501–504.

Tavaré, S. 1984. Lines-of-descent and genealogical processes, and their application in population genetic models, *Theoret. Popul. Biol.* **26**, 119–164.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination, *Theoret. Popul. Biol.* **7**, 256–276.

Wright, S. 1931. Evolution in Mendelian populations, *Genetics* **16**, 97–159.

Wright, S. 1977. “Evolution and the Genetics of Populations. Vol. 3. Experimental Results and Evolutionary Deductions,” Univ. of Chicago Press, Chicago.

Printed in Belgium