Polymorphism and Divergence for Island-Model Species

John Wakeley[1]

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received August 6, 2002
Accepted for publication October 14, 2002

ABSTRACT

Estimates of the scaled selection coefficient, γ of Sawyer and Hartl, are shown to be remarkably robust to population subdivision. Estimates of mutation parameters and divergence times, in contrast, are very sensitive to subdivision. These results follow from an analysis of natural selection and genetic drift in the island model of subdivision in the limit of a very large number of subpopulations, or demes. In particular, a diffusion process is shown to hold for the average allele frequency among demes in which the level of subdivision sets the timescale of drift and selection and determines the dynamic equilibrium of allele frequencies among demes. This provides a framework for inference about mutation, selection, divergence, and migration when data are available from a number of unlinked nucleotide sites. The effects of subdivision on parameter estimates depend on the distribution of samples among demes. If samples are taken singly from different demes, the only effect of subdivision is in the rescaling of mutation and divergence-time parameters. If multiple samples are taken from one or more demes, high levels of within-deme relatedness lead to low levels of intraspecies polymorphism and increase the number of fixed differences between samples from two species. If subdivision is ignored, mutation parameters are underestimated and the species divergence time is overestimated, sometimes quite drastically. Estimates of the strength of selection are much less strongly affected and always in a conservative direction.

ONE of the primary goals of population genetics has been to measure and to understand the role of natural selection in shaping variation within and between species. Now that molecular technologies allow genetic variation to be assayed with relative ease, this goal seems within reach. A number of different approaches to studying selection have been proposed (Hudson and Kaplan 1988; Neuhauser and Krone 1997; Yang 1998; Donnelly *et al.* 2001; Slatkin and Bertorelle 2001), and a multitude of neutrality tests, reviewed by Nielsen (2001), can be applied if appropriate genetic data are gathered. This work considers Sawyer and Hartl's (1992) method, which belongs to a class of methods that use overall levels of polymorphism and divergence at two or more categories of sites in samples of DNA sequences from a pair of species. Hudson *et al.* (1987) were the first to propose such a method, in which the categories were different loci, followed by McDonald and Kreitman (1991), who categorized sites within a locus as being either synonymous or nonsynonymous with respect to changes in the amino acid sequence of the protein product. Both methods assumed no intralocus recombination and allowed the hypothesis of strict selective neutrality to be tested. Shortly afterward, by assuming Kimura's (1969) infinite-sites mutation model, *i.e.*, with free recombination between sites, Sawyer and Hartl (1992) showed that McDonald-Krietman test data could be used not only to test neutrality but also to estimate selection, mutation, and divergence-time parameters.

Nielsen (2001) pointed out that McDonald-Kreitman and related tests, in which sites can be classified *a priori*, provide a very powerful framework for inferences about natural selection, in contrast to tests like Tajima's (1989) and Fu and Li's (1993), which measure deviations from the highly variable process of neutral coalescence. It is likely that McDonald-Kreitman and related methods will become the mainstay of genomic analyses of the role of selection. In two recent works, modified McDonald-Kreitman tests were applied to genomic data from Drosophila, suggesting that 45% of the amino acid differences between *Drosophila simulans* and *D. yakuba* resulted from positive selection (Smith and Eyre-Walker 2002) and that positive selection at a relatively small number of genes is responsible for the divergence of *D. simulans* and *D. melanogaster* (Fay *et al.* 2002). Bustamante *et al.* (2002) used a modified Sawyer-Hartl method to show that Arabidopsis species have experienced a higher proportion of deleterious amino acid substitutions than Drosophila species, in which positive selection is common, and attributed the difference to high levels of inbreeding in Arabidopsis.

An obvious shortcoming of these methods is that they assume the species under study are panmictic, *i.e.*, not geographically or otherwise subdivided. It is well known that this assumption is incorrect for many species (Slatkin 1985). When there is no intralocus recombination, McDonald and Kreitman (1991) point out that shared

[1]*Address for correspondence:* 2102 Biological Laboratories, 16 Divinity Ave., Cambridge, MA 02138. E-mail: wakeley@fas.harvard.edu

genealogical history should control for the effects of demography when sites can be categorized *a priori*. It is less clear that this should be the case when collections of unlinked sites are used to estimate selection, mutation, and divergence-time parameters as in Sawyer and Hartl's (1992) method. It is possible that the effects of subdivision on the numbers of polymorphisms and fixed differences at synonymous and nonsynonymous sites could lead to errors in inferences. Therefore, the goal of this work is to extend the Poisson random field (PRF) theory of polymorphism and divergence developed by Sawyer and Hartl (1992) to include subdivided species. To do this, it is first shown that in the limit of a large number of subpopulations or demes allele-frequency dynamics at a single locus in a population with island-model migration (Wright 1931; Moran 1959; Maruyama 1970; Latter 1973) are governed by a diffusion process that has the same form as the usual Wright-Fisher diffusion, *e.g.*, see Ewens (1979), but with a timescale different from that of the panmictic case. Then, the assumption of free recombination between sites allows the PRF model to be used to predict the patterns of variation in samples from a pair of island-model species.

The diffusion result is obtained using Theorem 3.3 in Ethier and Nagylaki (1980) and relies upon the fact that the process of migration and drift within subpopulations occurs on a much faster timescale than changes in allele frequency by drift and selection in the total population. The result thus depends on a stochastic equilibrium of allele frequencies within demes with respect to migration and drift, which is also described. This follows some recent work (Cherry and Wakeley 2003) in which simulations supported the existence of such a diffusion under the additional assumption that demes are very large and migration rates correspondingly small. The present analysis shows that this additional assumption is unneccessary. The assumption of infinite deme sizes and infinitesimal migration rates was also made in the recent coalescent work on neutral large-number-of-demes models (Wakeley 1998, 2001), and it is made below in *The expected number of neutral segregating sites*, when the forward and backward results are compared. Otherwise, here it is assumed that the demes are finite in size and the migration rates are unconstrained.

This work makes a connection between the PRF theory and work on the robustness of the coalescent process to population structure (Nordborg 1997; Möhle 1998), in particular for the case of geographic structure (Wakeley 1998, 2001). The two are related by showing that the effective size of the ancestral, coalescent process is the same as that of the forward-time diffusion of allele frequencies and that the forward- and backward-derived predictions for the expected number of segregating sites in a sample are the same under neutrality. We expect such connections between forward and backward approaches to exist, a fact that is well established in the case of panmictic populations (Ewens 1990; Möhle 2001). Like the forward (Nagylaki 1980) and backward (Notohara 1993) strong-migration limits, these results and those of Wakeley (1998, 2001) for the coalescent process are based on a "separation of timescales." In this case, the fast processes are migration and drift within demes and the slow process is drift and possibly selection in the total population, which is mediated by migration. The effective size of the population is rescaled and patterns of genetic variation depend on how a sample is distributed among demes. In contrast, under the usual strong-migration limit, the only effect of structure is to rescale the effective size of the population (Nagylaki 1980; Notohara 1993; Nordborg 1997; Charlesworth 2001).

The main result presented here, besides the existence of the diffusion (9) below, is that, if mutations are introduced at a constant rate per generation and sites segregate independently of one another, the PRF results of Sawyer and Hartl (1992) can be applied, but with a correction that depends on how samples are taken among demes. If each sample is taken from a different deme, then Sawyer and Hartl's (1992) results apply directly, but with slightly different mutation and divergence-time parameters. If some or all of the samples come from the same deme, the PRF results must be corrected for the effect of drift and migration within demes. Failure to recognize this can cause serious errors in the estimation of mutation rates and divergence times, but not, surprisingly, of selection coefficients.

THEORY

A population or species is assumed to be subdivided into D demes of equal size N. The organisms are assumed to be haploid, but the results will hold for diploid organisms if N is replaced with $2N$, if selection is additive, and if migration is gametic. The island model of migration (Wright 1931; Moran 1959) is assumed: a fraction m of each deme is replaced by migrants every generation and all migrants are randomly sampled from a migrant pool to which all demes contribute equally. In each generation, migration occurs first, followed by selection, and then resampling (drift) within demes according to the Wright-Fisher model (Fisher 1930; Wright 1931). In the next two sections, two alleles are assumed to be segregating at a single locus, and *Many independently segregating loci* considers their introduction by mutation. The wild-type or nonmutant allele has relative fitness equal to 1, and the mutant allele has fitness $1 + s_D$, where $s_D \geq -1$. The next section establishes the diffusion approximation for the frequency of the mutant allele as $D \to \infty$, but Ds_D remains finite. The migration rate can vary between 0 and 1 ($0 < m \leq 1$) and N is assumed to be finite. This is in contrast to the usual assumption that Nm is finite as N goes to infinity.

Considering the number of mutants within each deme,

it is apparent that there are exactly $N + 1$ kinds of demes. Each deme that begins a generation with i copies of the mutant will have mutant frequency

$$q_i = (1 - m)\frac{i}{N} + mx + s_D\left[(1 - m)\frac{i}{N} + mx\right]\left[1 - (1 - m)\frac{i}{N} - mx\right] + o(s_D) \quad (1)$$

after migration and selection, where x is the frequency of the mutant in the total population. The next generation within the deme will be produced by randomly sampling N haploid individuals from this distribution. Thus, a deme that contains i copies of the mutant now has probability

$$P_{ij} = \binom{N}{j} q_i^j (1 - q_i)^{N-j} \quad (2)$$

of having j copies at the start of the following generation. Because $\lim_{D\to\infty} s_D = 0$, it is often necessary to consider only one part of P_{ij}:

$$P_{ij}^* = \binom{N}{j}\left[(1 - m)\frac{i}{N} + mx\right]^j\left[1 - (1 - m)\frac{i}{N} - mx\right]^{N-j}. \quad (3)$$

The notation $o(s_D)$ used in Equation 1 and below means that $\lim_{D\to\infty} o(s_D)/s_D = 0$. Thus $P_{ij} = P_{ij}^* + o(1)$. The process of drift, described by Equation 2, happens independently within each deme.

Limiting allele frequency dynamics at a single locus: Let $Z_i^D(t)$ record the fraction of demes that contain i copies of the mutant and $z_i(t)$ be a particular realization of this random variable. Thus, $Z^D(t)$ is a Markov chain whose state space consists of all possible configurations of the D demes among the $N + 1$ mutant-count classes. appendix a proves a diffusion result for $Z^D(t)$ as D goes to infinity and Ds_D remains finite. Briefly, this is done by using Equation 1 of Ethier and Nagylaki (1980)—see also Equation 22 of Nagylaki (1980). Define $X^D(t) = \sum_{i=0}^N i Z_i^D(t)/N$. The random variable $X^D(t)$ records the frequency of the mutant in the total population or the average frequency of the mutant among demes (x above). Next, let $Y^D(t) = Z_i^D(t) - \nu_i(t)$ be the deviation of $Z_i^D(t)$ from the equilibrium prediction $\nu_i(t)$. For a given $P_{ij}(t)$, this equilibrium satisfies

$$\nu_j(t) = \sum_{i=0}^{N} \nu_i(t) P_{ij}^*(t). \quad (4)$$

It exists because $\boldsymbol{P}^* = \langle P_{ij}^* \rangle$ is ergodic and has a finite number of states. We can set $\sum_{i=0}^N \nu_i(t) = 1$, and $\nu_i(t)$ becomes the equilibrium prediction for $Z_i^D(t)$.

The nature of the diffusion approximation (9) below is that the migration and drift within demes equilibrate quickly in comparison to the rate of drift and selection in the total population. The results show that, to a sufficient order of approximation, demes can be considered to always be at a stochastic equilibrium ν_j ($0 \le j \le N$) with respect to migration and drift for a given x. The fraction of demes that have j copies of the mutant converges on $\nu_j = \sum_{i=1}^N \nu_i P_{ij}^*$, where P_{ij}^* is given by Equation 3 with x constant. The distribution ν is very well approximated by the hypergeometric distribution

$$\nu_j = \frac{\binom{-Nm(2-m)x/(1-m)^2}{j}\binom{-Nm(2-m)(1-x)/(1-m)^2}{N-j}}{\binom{-Nm(2-m)/(1-m)^2}{N}}, \quad (5)$$

which is a special case of the multivariate Poly(A) distribution; see Equation 40.13 in Johnson *et al.* (1997). Equation 5 is also identical to the two-allele case of the compound multinomial Dirichlet distribution, which Rannala (1996) proved to hold for the frequencies of multiple alleles within a deme in the infinite-island or continent-island model, *i.e.*, where allele frequencies among migrants are assumed to be constant. Rannala (1996) did not assume Wright-Fisher reproduction, but rather that a birth-death-immigration process occurred within demes. Thus, Rannala's (1996) model is similar to the Moran model, in which such distributions are known to arise: see pages 131–133 in Moran (1962). Rothman *et al.* (1974) argued for the use of the compound multinomial Dirichlet distribution in the case of Wright-Fisher reproduction within demes.

The form of Equation 5 was obtained by selecting parameters of a hypergeometric distribution that gave the same mean and variance of allele counts among demes as Equation 4, namely

$$E_\nu[j] = Nx \quad (6)$$

$$\mathrm{Var}_\nu[j] = \frac{N^2 x(1 - x)}{m(2 - m) + (1 - m)^2}, \quad (7)$$

which were obtained using (4) together with the moments of the binomial distribution (P^*). Equation 5 is the exact solution to (4) when $N \le 2$. In addition, as required by (4): when m approaches 1, ν_i becomes a binomial distribution with parameters N and x; and as m approaches 0, we have $\nu_0 = 1 - x$, $\nu_N = x$, and $\nu_j = 0$ for $1 \le j \le N - 1$. Finally, if $x_i = j/N$ is the frequency in some deme i, then as N grows but $2Nm = M$ remains constant, (5) converges on the well-known β-distribution result

$$g(x_i|x)\,dx_i = \frac{\Gamma(M)}{\Gamma(Mx)\Gamma(M(1 - x))} x_i^{Mx-1}(1 - x_i)^{M(1-x)-1}dx_i, \quad (8)$$

which Wright (1931) obtained under the assumption that x was constant among migrants. To derive (8) from (5), it is necessary to use the limit result 6.1.46 in Abramowitz and Stegun (1965) for ratios of gamma functions and to let $dx_i = 1/N$. Figure 1 plots the distribution (5) when $N = 10$ and $x = 0.75$ over the full range of migration rates. With these parameter values, the absolute error of using (5) to approximate the solution of (4) is never $> \sim 0.007$ and the relative error is never $> \sim 5\%$.

appendix a shows that, in the limit as D goes to infin-

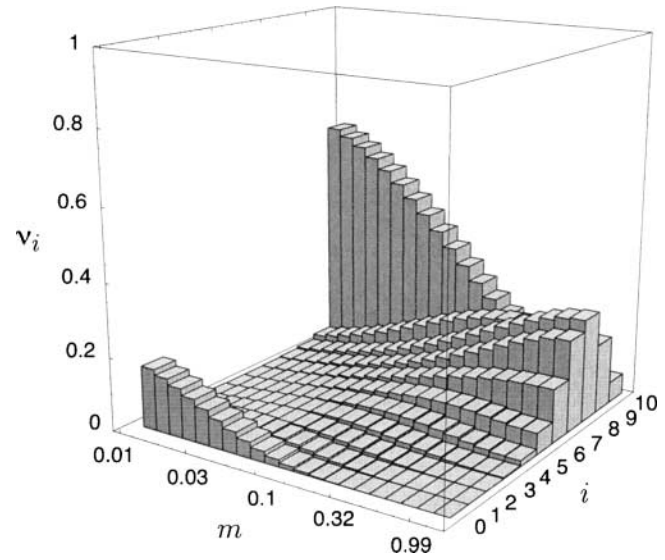

FIGURE 1.—The approximation (5) for the distribution of mutant allele counts among demes assuming that $N = 10$ and $x = 0.75$, shown as a function of the per-generation migration rate m.

ity, the change in x by drift and selection is so much slower than that by migration and drift within demes that the collection of demes is always at the equilibrium ν_i, which depends on N and m, and of course x. By Theorem 3.3 of ETHIER and NAGYLAKI (1980), as D goes to infinity the above system reduces to a diffusion $x(\cdot)$ with generator

$$\mathcal{L} = \frac{1}{2}x(1 - x)\frac{d^2}{dx^2} + \gamma x(1 - x)\frac{d}{dx}, \tag{9}$$

in which $\gamma = N \lim_{D\to\infty} Ds_D$. Time is measured in units of $ND/(1 - F)$ generations, where F is the fixation coefficient, in this case given by Equation (A13) in APPENDIX A. Thus, the diffusion of x is identical to the usual Wright-Fisher diffusion with genic selection, with the exception that it occurs on a timescale longer than that of the panmictic case by the factor $1/(1 - F)$. Thus, all the well-known predictions of that model apply; *e.g.*, see chapter 5 of EWENS (1979).

CHERRY and WAKELEY (2003) assumed (8) to hold and showed that simulations agreed well with the predictions of the implied diffusion process, such as the time to fixation or loss of the mutant type. Without giving a proof, we can guess that this diffusion should be given by the results of the section above and APPENDIX A if $N_D \to \infty$ when $D \to \infty$ and $\lim_{D\to\infty} 2N_D m_D = M$, so that $F = 1/(M + 1)$, but with $\lim_{D\to\infty} N_D/D = 0$ (ETHIER and NAGYLAKI 1980). CHERRY and WAKELEY (2003) also showed that the distribution of allele frequencies among demes in simulations with $N = 100$ and $m = 0.01$ (and $D = 1000$ and $s_D = 0.001$) conformed well to the predictions of Equation 8 in a particular generation when x was equal to 0.611. Further support for the existence of this large-D, large-N diffusion is given in *The expected number of neutral segregating sites* by comparing its predictions under neutrality to those of the corresponding coalescent model (WAKELEY 1998). Otherwise, N is assumed here to be finite.

Many independently segregating loci: If we posit an infinite number of loci, *i.e.*, nucleotide sites, which can sustain mutations and which each evolve according to the diffusion of the previous section independently, then the PRF results of SAWYER and HARTL (1992) hold for x. Because of the way time is measured in the diffusion, the appropriate mutation parameters are also scaled:

$$\theta_a = \frac{2NDu_a}{1 - F} \quad \text{and} \quad \theta_s = \frac{2NDu_s}{1 - F}. \tag{10}$$

The subscripts in Equation 10 refer to "amino acid replacement" and "synonymous" following BUSTAMANTE *et al.* (2002), and u_a and u_s are the per-generation rates. Thus, one effect of restricted migration is to increase the apparent mutation rates over the panmictic case since $0 \le F \le 1$. The other effect, of course, is to distribute variation among demes as described in the previous section. In addition, the parameter t_{div} in SAWYER and HARTL (1992) must here be measured in units of $ND/(1 - F)$ generations. With these modifications, Equations 13 and 14 in SAWYER and HARTL (1992) apply here to x.

Rewriting SAWYER and HARTL's (1992) Equations 13 and 14 in terms of the present notation gives

$$\theta_s t_{div} \tag{11}$$

$$\theta_a t_{div}\frac{2\gamma}{1 - e^{-2\gamma}} \tag{12}$$

$$d\phi_s(x) = \theta_s\frac{dx}{x} \tag{13}$$

$$d\phi_a(x) = \theta_a\frac{1 - e^{-2\gamma(1-x)}}{1 - e^{-2\gamma}}\frac{dx}{x(1 - x)} \tag{14}$$

for the expected numbers of fixed and polymorphic, synonymous, and replacement differences in two species. When a sample is taken from the two species, as in SAWYER and HARTL (1992), we need to consider the chance that a polymorphic site appears fixed in a sample from the species. Here, in contrast to the panmictic case, the distribution of the sample among demes becomes important.

Assume that we have taken a random sample of n sequences from d different demes in one of the species, such that $n_1, n_2, \ldots, n_d$ are the sample sizes from each deme. We can write in general that the expected number of sites that show $i_1, i_2, \ldots, i_d$ copies of the mutant base in the sample ($0 \le i_k \le n_k$) is given by

$$E[S_j(i_1, \ldots, i_d)] = \int_0^1 \prod_{k=1}^{d} h(i_k|x, n_k)\, d\phi_j(x), \tag{15}$$

where $j = a, s$. The probability $h(i_k|x, n_k)$, that i_k copies of the mutant base are in the sample of n_k items from the kth sampled deme, is an average over the within-deme distribution of allele frequencies:

$$h(i_k|x, n_k) = \sum_{j=i_k}^{N} \frac{\binom{j}{i_k}\binom{N-j}{n_k - i_k}}{\binom{N}{n_k}} \nu_j. \quad (16)$$

If N is large and m correspondingly small, we may wish to use the large-deme approximation:

$$h^*(i_k|x, n_k) = \int_0^1 \binom{n_k}{i_k} x_k^{i_k}(1 - x_k)^{n_k - i_k} g(x_k|x)\, dx_k. \quad (17)$$

That is, when N is large we can approximate the hypergeometric probability that the sample contain i_k copies of the mutant allele (present in j copies in the deme) with a binomial distribution and the allele count distribution ν_j with Wright's (1931) continous β-distribution of allele frequences, $g(x_k|x)$.

Because we have assumed an infinite number of independently segregating sites with collective mutation rates given by (10), the PRF model (Sawyer and Hartl 1992) shows that $S_j(i_1, \ldots, i_d)$ is Poisson distributed with expected value equal to (15). The numbers of sites segregating at various frequencies within each deme contain information about migration rates, and the numbers of sites segregating at various frequencies in the total population contain information about the selection coefficient. Note that (15) can also be used to compute the expected number of apparent fixed differences, *i.e.*, polymorphisms where the entire sample has the mutant base, as required in Sawyer and Hartl's (1992) analysis. This provides a framework for estimating selection coefficients (and migration rates) in the context of a subdivided population. As illustrated in results, we use Equations 11–14 in conjunction with Equation 15 to obtain predictions about the numbers of fixed-synonymous, fixed-replacement, polymorphic-synonymous, and polymorphic-replacement sites in a sample from two species. Further, Equation 15 gives the joint frequencies among demes of segregating polymorphisms. In the panmictic case, Hartl *et al.* (1994), Akashi (1999), and Bustamante *et al.* (2001) showed that allele frequencies at polymorphic sites contain substantial information about selection.

RESULTS

The first result to note is that if each sample is taken from a different deme, the methods of Sawyer and Hartl (1992) can be applied wihout modification. It is necessary only to realize that the inferred mutation parameters and the divergence time are scaled in terms of $ND/(1 - F)$ generations instead of the usual ND generations. This result follows from the fact that each sample drawn in this way has probability x of showing the mutant base. That is, $h(1|x, 1) = x$ and $h(0|x, 1) = 1 - x$, and similarly for $h^*(i_k|x, n_k)$. Summing Equation 15, for each species, over all $i_1, i_2, \ldots, i_d$ such that $0 < \Sigma_{k=1}^{d} i_k < \Sigma_{k=1}^{d} n_k$ gives Sawyer and Hartl's (1992) Equations 15 and 19 but with the scaled mutation rates that apply here: θ_s and θ_a. Similarly, Sawyer and Hartl's (1992) Equations 17 and 18 are derived by considering the chance that $i_k = 1$ for all k. In sum, inferences about selection coefficients, mutation rates, and divergence times are entirely robust to (island-model) population subdivision when each sample is taken from a different deme.

Inferences from single-deme samples: At the opposite extreme, consider the case in which all samples are drawn from the same deme within each species. Note that we assume, as in Sawyer and Hartl (1992), that the two species are identical (here in terms of N, m, and γ). Let n_1 and n_2 denote the sample sizes from the two species. For this sample, the expected numbers of fixed-synonymous (K_s), fixed-replacement (K_a), polymorphic-synonymous (S_s), and polymorphic-replacement (S_a) sites are given by

$$E(K_s) = \theta_s\left\{t_{div} + \int_0^1 [h(n_1|x, n_1) + h(n_2|x, n_2)]\frac{dx}{x}\right\} \quad (18)$$

$$E(K_a) = \theta_a\frac{2\gamma}{1 - e^{-2\gamma}}\left\{t_{div} + \int_0^1 [h(n_1|x, n_1) + h(n_2|x, n_2)]\frac{1 - e^{-2\gamma(1-x)}}{2\gamma x(1 - x)}dx\right\} \quad (19)$$

$$E(S_s) = \theta_s\int_0^1 [H(x, n_1) + H(x, n_2)]\frac{dx}{x} \quad (20)$$

$$E(S_a) = \theta_a\frac{2\gamma}{1 - e^{-2\gamma}}\int_0^1 [H(x, n_1) + H(x, n_2)]\frac{1 - e^{-2\gamma(1-x)}}{2\gamma x(1 - x)}dx \quad (21)$$

in which $H(x, n) = 1 - h(n|x, n) - h(0|x, n)$. The results from *Limiting allele frequency dynamics at a single locus* are used to compute $h(n|x, n)$ and $h(0|x, n)$. Namely,

$$h(n|x, n) = \sum_{j=1}^{N} \frac{j!(N - n)!}{(j - n)!N!}\nu_j. \quad (22)$$

This same equation can be used to compute $h(0|x, n) = h(n|1 - x, n)$.

Figure 2 plots the expected values of K_s, K_a, S_s, and S_a as functions of the migration rate when $n_1 = n_2 = 10$ and $N = 100$ and for three different values of γ: -2, 0, and 2. The results are as expected for single-deme samples. When $m = 1$, they are the same as in a panmictic population. As m decreases, samples from single demes tend to be closely related, so the numbers of polymorphisms will decrease and the numbers of (apparent) fixation events will increase. This is true regardless of whether γ is positive, zero, or negative, although the relative magnitudes of the four quantities depend

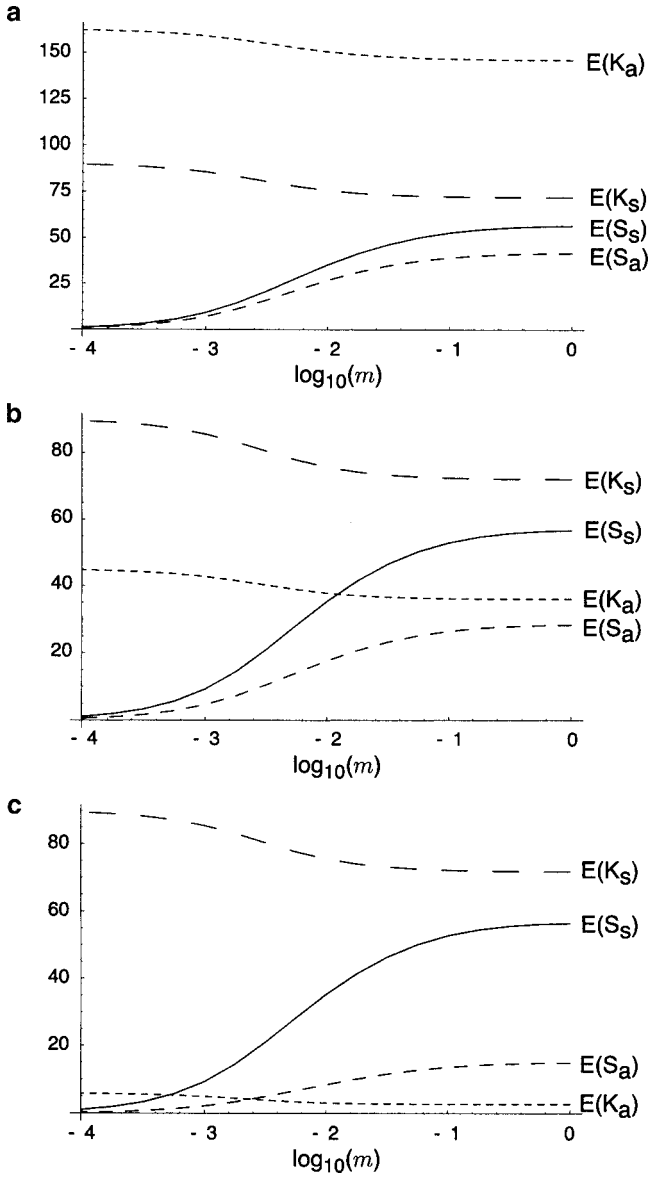

FIGURE 2.—The dependence on migration rate (m) of the expected values of K_s, K_a, S_s, and S_a computed using Equations 18–21, assuming $n_1 = n_2 = 10$ and $N = 100$. In addition, $\theta_s = 10$, $\theta_a = 5$, and $t_{div} = 7$. (a) $\gamma = 2$; (b) $\gamma = 0$; (c) $\gamma = -2$.

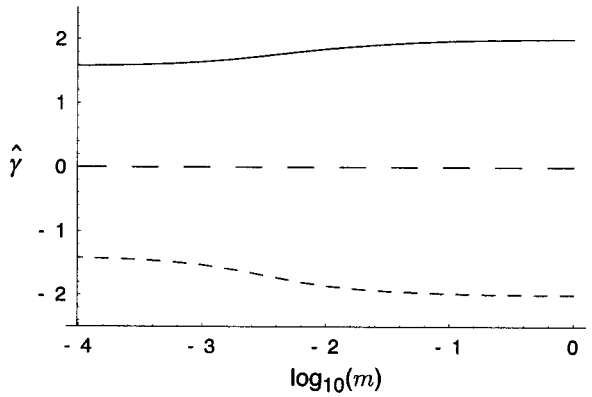

FIGURE 3.—The dependence on migration rate (m) of the estimated values of γ using the values of K_s, K_a, S_s, and S_a plotted in Figure 2 and assuming SAWYER and HARTL's (1992) panmictic PRF model. At the right ($m = 1$) the population is in fact panmictic, and γ is estimated accurately in all three cases.

strongly on γ. The curves for $E(K_s)$ and $E(S_s)$ are, of course, identical for all values of γ. The results that would be obtained by assuming $\lim_{N\to\infty} 2Nm = M$ and using Equations 8 and 17 would be similar to what is shown in Figure 2 if M were varied from 0.02 to 200.

To understand the effects of (island-model) population subdivision for the extreme case of single-deme samples, we can use the "data" of Figure 2 to fit the parameters of SAWYER and HARTL's (1992) panmictic model. Figure 3 shows that estimates of γ are remarkably robust to subdivision, even in this case, where the effects of subdivision should be strongest. Again, if samples were taken singly from different demes, there would be no error in using the panmictic model. For single-deme samples there is some error when the migration rate is low, but even in the extreme case of $m = 10^{-4}$ ($2Nm = 0.02$) the estimates are off only by ~25%. However, the level of error will be greater for larger samples (see DISCUSSION) and when the absolute value of γ is larger. An additional effect is that the error in estimating γ is conservative in that the bias is toward neutrality regardless of whether γ is positive or negative. Figure 4 shows the effect on the other parameters: t_{div}, θ_s, and θ_a. As should be expected from Figure 2, mutation rates are underestimated and the divergence time is overestimated when the migration rate is small. The error in estimating these other parameters is much more extreme than that for γ. In addition, there is a small effect of γ on estimates of θ_a.

The expected number of neutral segregating sites: Under neutrality, the results presented here agree with those found using a coalescent approach in WAKELEY (1998), and later in WAKELEY (1999, 2001), which were derived under the assumption that $\lim_{N\to\infty} 2Nm = M$. We make the same assumption here and further assume that this occurs in such a way that the diffusion result still holds (see *Limiting allele frequency dynamics at a single locus*). Then we can use $g(x_k|x)$ and $h^*(i_k|x, n_k)$ in expression (15) to show that the expected number of synonymous segregating sites is equal to $\theta_s \sum_{i=1}^{n-1} 1/i$ when all n sampled are taken from separate demes. This was found in WAKELEY (1998) to hold for the samples from the neutral genetic locus in the large-D island model, under the assumption of no intralocus recombination. We expect this agreement under the infinite-sites model of mutation, because the marginal distribution of genealogies at a single site must be the same as that of an entire nonrecombining locus under neutrality. It is important to note that the variances and other moments of the numbers of segregating sites do depend on the recombination rate.

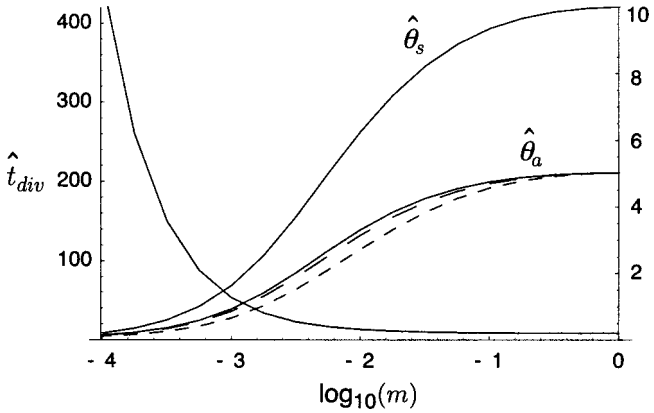

FIGURE 4.—The dependence on migration rate (m) of the estimated values of θ_s, θ_a, and t_{div} using the values of K_s, K_a, S_s, and S_a plotted in Figure 2 and assuming SAWYER and HARTL's (1992) panmictic PRF model. Estimates of θ_s and t_{div} depend only on neutral variation, but estimates of θ_a show some effect of selection. The three curves are, from the top, $\gamma = 2$, $\gamma = 0$, and $\gamma = -2$.

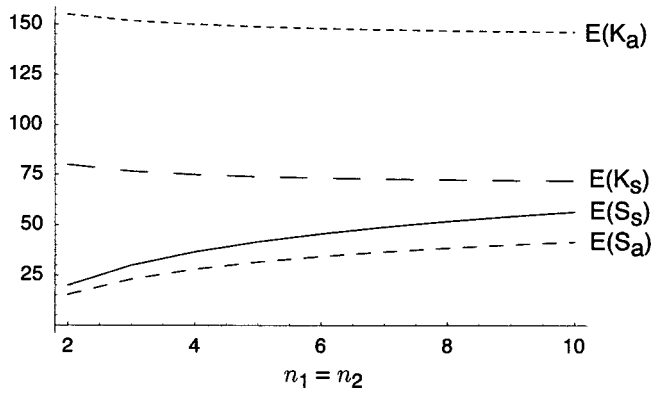

FIGURE 5.—An illustration that the overestimation of fixation events and underestimation of polymorphism levels result from a sample-size effect. Except for n_1 and n_2, parameters are the same as in Figure 2c, and the curves plot Equations 18–21 as a function of sample size.

Consider the number of segregating sites in a sample of n sequences, all from the same deme. From the coalescent approach we have

$$E[S] = \theta_s \sum_{n'=2}^{n} \frac{|S_1(n, n')|M^{n'}}{M_{(n)}} \sum_{i=1}^{n'-1} \frac{1}{i} \qquad (23)$$

(WAKELEY 1998), in which $S_1(i, j)$ are Stirling numbers of the first kind (ABRAMOWITZ and STEGUN 1964) and $M_{(n)} = M(M + 1) \ldots (M + n - 1)$. Here, Equation 15 becomes

$$E[S] = \int_0^1 \int_0^1 [1 - x_1^n - (1 - x_1)^n] g(x_1|x)\, dx_1\, d\phi_s(x) \qquad (24)$$

and this is shown in APPENDIX B to be equivalent to (23).

DISCUSSION

The results presented above can be understood in terms of a sample-size effect of subdivision, one that depends on how the sample is distributed among demes. In the limit of a large number of demes, the history of a sample under neutrality has two distinct phases: the scattering phase and the collecting phase described in WAKELEY (1999). Although in this analysis incorporating selection was not phrased in these terms, it is clear from Figure 2 that the same effect is at work, namely, that a scattering phase, which is a stochastic sample size adjustment that begins with a sample of size n and ends with n' lineages each in a separate deme, where $1 \leq n' \leq n$ (WAKELEY 1999), induces a downward sample-size adjustment to single-deme samples. In the case of large N and correspondingly small m, the scattering phase for a sample from a single deme is given by $P[n'|n] = |S_1(n, n')|M^{n'}/M_{(n)}$, which appears in Equation 23. Figure 5 shows how the expected values of K_s, K_a, S_s, and S_a depend on $n_1 = n_2$ under panmixia with $\gamma = 2$. Thus, the values on the right-hand side of Figure 5 are identical to those on the right-hand side of Figure 2a. Although scales of the horizontal axes are not the same, the effect of smaller migration rate is qualitatively similar to that of smaller sample size. The reason that the values on the left-hand sides of the two panels are different is that the average value of n' at the left in Figure 2a is equal to 1.06, which is considerably smaller than the practical lower limit of 2 in Figure 5. Instead, the values on the left-hand side of Figure 5 can be compared to those in Figure 2a for $\log_{10}(m) = -2.67$, or $m = 0.00215$, which (with $N = 100$) gives $E[n'] \approx 2$.

This work shows that inferences about natural selection made from DNA polymorphism and divergence data are robust to population subdivision (Figure 3) as long as the migration rate is not too low. This is remarkable in view of the strong effects subdivision has on numbers of polymorphisms, shown in Figure 2, but is understandable in terms of the effect of subdivision on θ_s, θ_a, and t_{div}. Except for the weak dependence of θ_a estimates on γ (Figure 4), subdivision and migration act equally on selected and neutral variation. In both cases, fixation events are overestimated and polymorphisms underestimated when the migration rate is small. This causes mutation rates to be substantially underestimated and divergence times grossly overestimated if subdivision is ignored, but these effects compensate one another and allow relatively accurate estimates of selection even if subdivision is ignored. Often γ will be the focus of study, but if θ_s, θ_a, and t_{div} are also of interest, it would be useful to have a framework for simultaneous inferences about migration rates, selection coefficients, and these other parameters. The theory presented above is a first step toward this goal.

It is important to note that inferences about natural selection made from allele frequencies at polymorphic sites will be robust to subdivision only in the case of

samples taken singly from different demes. Otherwise, the distribution of samples among demes will cause some frequency classes to be overrepresented, resulting in biased inferences. Even when all the samples are taken from the same deme, restricted migration can mimic the effect of positive γ on allele frequencies (Wakeley and Aliacar 2001). While allele frequencies at polymorphic sites provide an additional source of information about natural selection (Hartl *et al.* 1994), this illustrates that they are also greatly influenced by nonselective demographic factors; see also Nielsen (2001). In addition, allele frequency patterns are quite sensitive to levels of recombination (Bustamante *et al.* 2001). Thus, it is especially important to account for subdivision when making inferences from allele-frequency data.

I thank Dan Hartl, Thomas Nagylaki, Stanley Sawyer, and Clifford Taubes for helpful discussions of the work. I am also grateful to Sabin Lessard for seeing that deme sizes need not be large for the large-number-of-demes coalescent to hold. This work was supported by grants DEB-9815367 and DEB-0133760 from the National Science Foundation.

LITERATURE CITED

Abramowitz, M., and I. A. Stegun, 1965 *Handbook of Mathematical Functions.* Dover, New York.

Akashi, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151:** 221–238.

Bustamante, C. D., J. Wakeley, S. Sawyer and D. L. Hartl, 2001 Directional selection and the site-frequency spectrum. Genetics **159:** 1779–1788.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan *et al.*, 2002 The cost of inbreeding in *Arabidopsis.* Nature **416:** 531–534.

Charlesworth, B., 2001 Effect of life history and mode of inheritance on neutral genetic variation. Genet. Res. **77:** 153–166.

Cherry, J. L., and J. Wakeley, 2003 A diffusion approximation for selection and drift in a subdivided population. Genetics **163:** 421–428.

Donnelly, P., M. Nordborg and P. Joyce, 2001 Likelihoods and simulation methods for a class of nonneutral population genetic models. Genetics **159:** 853–867.

Ethier, S. N., and T. Nagylaki, 1980 Diffusion approximations of Markov chains with two timescales and applications to population genetics. Adv. Appl. Prob. **12:** 14–49.

Ewens, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, Berlin.

Ewens, W. J., 1990 Population genetics theory—the past and the future, pp. 177–227 in *Mathematical and Statistical Developments of Evolutionary Theory*, edited by S. Lessard. Kluwer Academic Publishers, Amsterdam.

Fay, J. C., G. J. Wyckoff and C.-I Wu, 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila.* Nature **415:** 1024–1026.

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

Fu, X.-Y., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Hartl, D. L., E. N. Moriyama and S. A. Sawyer, 1994 Selection intensity for codon bias. Genetics **138:** 227–234.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Johnson, N. L., S. Kotz and N. Balakrishnan, 1997 *Discrete Multivariate Distributions.* Wiley, New York.

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.

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

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

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *adh* locus in *Drosophila.* Nature **351:** 652–654.

Möhle, M., 1998 Robustness results for the coalescent. J. Appl. Prob. **35:** 438–447.

Möhle, M., 2001 Forward and backward diffusion approximations for haploid exchangeable population models. Stoch. Proc. Appl. **95:** 133–149.

Moran, P. A. P., 1959 The theory of some genetical effects of population subdivison. Austr. J. Biol. Sci. **12:** 109–116.

Moran, P. A. P., 1962 *Statistical Processes of Evolutionary Theory.* Clarendon Press, Oxford.

Nagylaki, T., 1980 The strong-migration limit in geographically structured populations. J. Math. Biol. **9:** 101–114.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics **145:** 519–534.

Nielsen, R., 2001 Statistical tests of neutrality in the age of genomics. Heredity **86:** 641–647.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501–1514.

Notohara, M., 1993 The strong migration limit for the genealogical process in geographically structured populations. J. Math. Biol. **31:** 115–122.

Rannala, B., 1996 The sampling theory of neutral alleles in an island population of fluctuating size. Theor. Popul. Biol. **50:** 91–104.

Rothman, E. D., C. F. Sing and A. R. Templeton, 1974 A model for the analysis of population structure. Genetics **78:** 934–960.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132:** 1161–1176.

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

Slatkin, M., and G. Bertorelle, 2001 The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics **158:** 865–874.

Smith, N. G., and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila.* Nature **415:** 1022–1024.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Wakeley, J., 1998 Segregating sites in Wright's island model. Theor. Popul. Biol. **53:** 166–175.

Wakeley, J., 1999 Non-equilibrium migration in human history. Genetics **153:** 1863–1871.

Wakeley, J., 2001 The coalescent in an island model of population subdivision with variation among demes. Theor. Popul. Biol. **59:** 133–144.

Wakeley, J., and N. Aliacar, 2001 Gene genealogies in a metapopulation. Genetics **159:** 893–905 (corrigendum: **160:** 1263–1264).

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

Yang, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15:** 568–573.

Communicating editor: N. Takahata

APPENDIX A

Again, $Z_i^D(t)$ is the fraction of demes that contain i copies of the mutant. Let $Z_{ij}^D(t)$ record the fraction of demes that contain i copies of the mutant and are descended from a deme that contained j copies of the mutant in the previous generation. Of course, $Z_i^D(t) = \sum_{j=0}^{N} Z_{ij}^D(t)$, and we can study the behavior of $Z_i^D(t)$ by

considering the simpler behavior of $Z_{ij}^D(t)$. In particular, conditional on the state of the system $\mathbf{z}(t)$ at time t,

$$(Z_{0j}^D(t+1), \ldots, Z_{Nj}^D(t+1)) \sim \frac{1}{Dz_j(t)}\text{multinomial}(Dz_j(t), P_{j0}(t), \ldots, P_{jN}(t)), \quad \text{(A1)}$$

and $Z_{ij}^D(t+1)$ and $Z_{kl}^D(t+1)$ are independent for all i and k, and $j \neq l$. Thus, we have conditional moments

$$E[Z_i^D(t+1)|\mathbf{z}(t)] = \sum_{j=0}^{N} z_j(t)P_{ji}(t) \quad \text{(A2)}$$

$$\text{Var}[Z_i^D(t+1)|\mathbf{z}(t)] = \frac{1}{D}\sum_{j=0}^{N} z_j(t)P_{ji}(t)(1 - P_{ji}(t)) \quad \text{(A3)}$$

$$\text{Cov}[Z_i^D(t+1), Z_j^D(t+1)|\mathbf{z}(t)] = -\frac{1}{D}\sum_{j=0}^{N} z_j(t)P_{ki}(t)P_{kj}(t). \quad \text{(A4)}$$

All the higher central moments of the $Z_i^D(t+1)$ are $o(1/D)$.

Now let $X^D(t) = \Sigma_{i=0}^N iZ_i^D(t)/N$, and $Y_i^D(t) = Z_i^D(t) - \nu_i(t)$. The diffusion result follows from these results (derived below) for changes over one generation:

$$E[X^D(1) - x|\mathbf{z}] = b(x, y) + o\left(\frac{1}{D}\right) \quad \text{(A5)}$$

$$E[\{X^D(1) - x\}^2|\mathbf{z}] = a(x, y) + o\left(\frac{1}{D}\right) \quad \text{(A6)}$$

$$E[\{X^D(1) - x\}^4|\mathbf{z}] = o\left(\frac{1}{D}\right) \quad \text{(A7)}$$

$$E[Y_i^D(1) - y_i|\mathbf{z}] = c_i(x, y) + o(1) \quad \text{(A8)}$$

$$\text{Var}[Y_i^D(1)|\mathbf{z}] = o(1) \quad \text{(A9)}$$

in which t has been suppressed, $x = \Sigma_{i=0}^N iz_i/N$, and $y_i = z_i - \nu_i$, and

$$b(x, y) = s_D(1 - F)x(1 - x) + s_D(1 - m)^2\sum_{i=0}^{N}\left(\frac{i}{n} - x\right)^2 y_i \quad \text{(A10)}$$

$$a(x, y) = \frac{1}{ND}(1 - F)x(1 - x) + \frac{(1 - m)^2}{ND}\sum_{i=0}^{N}\left(\frac{i}{n} - x\right)^2 y_i \quad \text{(A11)}$$

$$c_i(x, y) = \sum_{j=0}^{N} y_j P_{ji}^* - y_i. \quad \text{(A12)}$$

The fixation index is given by

$$F = \frac{(1 - m)^2}{Nm(2 - m) + (1 - m)^2}. \quad \text{(A13)}$$

It is clear from Equation A12 that $c(x, 0) = 0$ for all $x \in (0, 1)$. If, in addition, the zero solution of the difference equation

$$Y(k + 1, x, y) - Y(k, x, y) = c(x, Y(k, x, y)), \quad Y(0, x, y) = y, \quad \text{(A14)}$$

is globally asymptotically stable, then the diffusion (9) holds (Ethier and Nagylaki 1980). Note that $y = 0$ is equivalent to $z_i = \nu_i$ and that Equation A14 is equivalent to $Y(k + 1, x, y) = Y(k, x, y)\boldsymbol{P}^*$. Proof of Equation A14 follows from the ergodicity of the stochastic matrix $\boldsymbol{P}^*$, *i.e.*, that $\lim_{k\to\infty} P_{ij}^{*(k)} = \nu_j$, along the same lines as the proof in Nagylaki (1980, pp. 111–112).

The derivation of Equations A5–A9 follows from Equations A2–A4. For Equation A5 we have

$$E[X^D(1) - x|\mathbf{z}] = E\left[\sum_{i=0}^{N}\frac{i}{N}Z_i^D(1)\right] - x \quad \text{(A15)}$$

$$= \sum_{i=0}^{N}\frac{i}{N}\sum_{j=0}^{N} z_j P_{ji} - x \quad \text{(A16)}$$

$$= \sum_{j=0}^{N} z_j q_j - x. \quad \text{(A17)}$$

Putting in q_j from Equation 1 and simplifying give

$$E[X^D(1) - x|\mathbf{z}] = s_D\left[x(1 - x) - (1 - m)^2\sum_{i=0}^{N}\left(\frac{i}{N} - x\right)^2 z_i\right] + o(s_D), \quad \text{(A18)}$$

which gives (A5) if we put $z_i = y_i + \nu_i$ on the right and simplify using Equation 7.

For Equation A6 we have

$$E[\{X^D(1) - x\}^2|\mathbf{z}] = E\left[\left\{\sum_{i=0}^{N}\frac{i}{N}Z_i^D(1) - \sum_{i=0}^{N}\frac{i}{N}z_i\right\}^2\right] \quad \text{(A19)}$$

$$= E\left[\left\{\sum_{i=0}^{N}\frac{i}{N}(Z_i^D(1) - E[Z_i^D(1)]) + \sum_{i=0}^{N}\frac{i}{N}(E[Z_i^D(1)] - z_i)\right\}^2\right] \quad \text{(A20)}$$

$$= E\left[\left\{\sum_{i=0}^{N}\frac{i}{N}(Z_i^D(1) - E[Z_i^D(1)])\right\}^2\right] + o(s_D) \quad \text{(A21)}$$

$$= \sum_{i=0}^{N}\left(\frac{i}{N}\right)^2\text{Var}[Z_i^D(1)] + \sum_{i=0}^{N}\sum_{\substack{k=0\\k\neq i}}^{N}\frac{i}{N}\frac{k}{N}\text{Cov}[Z_i^D(1), Z_k^D(1)] + o(s_D) \quad \text{(A22)}$$

$$= \frac{1}{D}\sum_{j=0}^{N} z_j\left\{\sum_{i=0}^{N}\left(\frac{i}{N}\right)^2 P_{ji} - \left[\sum_{i=0}^{N}\frac{i}{N}P_{ji}\right]^2\right\} + o(s_D) \quad \text{(A23)}$$

$$= \frac{1}{D}\sum_{j=0}^{N} z_j\frac{q_j(1 - q_j)}{N} + o(s_D). \quad \text{(A24)}$$

Again, putting in q_j and simplifying, this becomes Equation A6.

For equation A7 we have

$$E[\{X^D(1) - x\}^4|\mathbf{z}] = E\left[\left\{\sum_{i=0}^{N} \frac{i}{N} Z_i^D(1) - \sum_{i=0}^{N} \frac{i}{N} z_i\right\}^4\right] \quad \text{(A25)}$$

$$= E\left[\left\{\sum_{i=0}^{N} \frac{i}{N}(Z_i^D(1) - E[Z_i^D(1)]) + \sum_{i=0}^{N} \frac{i}{N}(E[Z_i^D(1)] - z_i)\right\}^4\right]. \quad \text{(A26)}$$

As in (A20) above, the second sum on the right in (A26) is equal to $E[X^D(1) - x|\mathbf{z}]$, which, from (A5), is $o(1)$. Expanding and considering the third and fourth central moments of $Z_i^D(1)$ gives the result (A7).

In the derivations of (A8) and (A9) below I assume that the exact solution of (4) is sufficiently close to (5) that the latter can be used in place of the exact solution. More precisely, I assume that

$$\nu_i(1) = \nu_i + \sum_{k=1}^{N} r_k (X^D(1) - x)^k, \quad \text{(A27)}$$

where the coefficients r_k depend on N, i, m, and x. This is certainly true for Equation 5, and because (5) and the exact solution of (4) are nearly identical in form (see Figure 1 and associated text), it should also be true of the exact solution although the coefficients r_k will be different.

For Equation A8 we have

$$E[Y_i^D(1) - y_i|\mathbf{z}] = E[Z_i^D(1) - \nu_i(1)|\mathbf{z}] - y_i \quad \text{(A28)}$$

$$= \sum_{j=0}^{N} z_j P_{ji} - E[\nu_i(1)|\mathbf{z}] - y_i. \quad \text{(A29)}$$

Using (A27), the second term on the right in Equation A29 becomes

$$E[\nu_i(1)|\mathbf{z}] = \nu_i + \sum_{k=1}^{N} r_k E[(X^D(1) - x)^k|\mathbf{z}]. \quad \text{(A30)}$$

Then by the same argument that gave (A7), using (A1), it can be shown that these higher moments are also $o(1/D)$. Because of this, and putting in $\nu_i = \Sigma_{j=0}^{N} \nu_j P_{ji}^*$, Equation A29 becomes

$$E[Y_i^D(1) - y_i|\mathbf{z}] = \sum_{j=0}^{N} z_j P_{ji} - \sum_{j=0}^{N} \nu_j P_{ji}^* - y_i + o(1) \quad \text{(A31)}$$

$$= \sum_{j=0}^{N} y_j P_{ji}^* - y_i + o(1), \quad \text{(A32)}$$

which is equal to (A8).

For Equation A9 we have

$$\mathrm{Var}[Y_i^D(1)|\mathbf{z}] = \mathrm{Var}[Z_i^D(1) - \nu_i(1)|\mathbf{z}] \quad \text{(A33)}$$

$$\leq 2\,\mathrm{Var}[Z_i^D(1)|\mathbf{z}] + 2\,\mathrm{Var}[\nu_i(1)|\mathbf{z}], \quad \text{(A34)}$$

using (3.12) in Ethier and Nagylaki (1980). From (A3), we have $\mathrm{Var}[Z_i^D(1)|\mathbf{z}] = o(1)$. From Equation A27 we can see that, like (A30), the second term in (A34) ultimately depends on the moments of Z_i^D and so is also $o(1)$. Therefore, $\mathrm{Var}[Y_i^D(1)|\mathbf{z}] = o(1)$ as required in Equation A9.

This completes the derivation of (A5–A9), showing that Theorem 3.3 in Ethier and Nagylaki (1980) can be applied and that the diffusion $x(\cdot)$ with generator (9) in the text holds as D goes to infinity.

APPENDIX B

Beginning with Equation 24, and then putting in $g(x|x)$ and $\phi_s(x)$, we have

$$E[S] = \int_0^1 \int_0^1 [1 - x_1^n - (1 - x_1)^n] g(x_1|x)\, dx_1 d\phi_s(x) \quad \text{(B1)}$$

$$= \theta_s \int_0^1 \frac{1}{x} \int_0^1 [1 - x_1^n - (1 - x_1)^n] g(x_1|x)\, dx_1 dx \quad \text{(B2)}$$

$$= \theta_s \int_0^1 \frac{1}{x} \left[1 - \frac{(Mx)_{(n)}}{M_{(n)}} - \frac{(M(1-x))_{(n)}}{M_{(n)}}\right] dx. \quad \text{(B3)}$$

Using the identity

$$M_{(n)} = \sum_{n'=1}^{n} |S_1(n, n')| M^{n'} \quad \text{(B4)}$$

we obtain

$$E[S] = \theta_s \sum_{n'=1}^{n} \frac{|S_1(n, n')| M^{n'}}{M_{(n)}} \int_0^1 \frac{1}{x} [1 - x^{n'} - (1 - x)^{n'}] dx \quad \text{(B5)}$$

$$= \theta_s \sum_{n'=1}^{n} \frac{|S_1(n, n')| M^{n'}}{M_{(n)}} \sum_{i=1}^{n'} \frac{1}{i}, \quad \text{(B6)}$$

which is the same as Equation 23 since the first term ($n' = 1$) is equal to zero.