Estimating Contemporary Effective Population Size on the Basis of Linkage Disequilibrium in the Face of Migration

Robin S. Waples[*,1] **and Phillip R. England**[†]

[*]National Oceanic and Atmospheric Administration, Northwest Fisheries Science Center, Seattle, Washington 98112, and
[†]Commonwealth Scientific and Industrial Research Organization Marine and Atmospheric Research and Wealth From Oceans
Flagship, Hobart, TAS 7001, Australia

ABSTRACT Effective population size (N_e) is an important genetic parameter because of its relationship to loss of genetic variation, increases in inbreeding, accumulation of mutations, and effectiveness of selection. Like most other genetic approaches that estimate contemporary N_e, the method based on linkage disequilibrium (LD) assumes a closed population and (in the most common applications) randomly recombining loci. We used analytical and numerical methods to evaluate the absolute and relative consequences of two potential violations of the closed-population assumption: (1) mixture LD caused by occurrence of more than one gene pool, which would downwardly bias $\hat{N}_e$, and (2) reductions in drift LD (and hence upward bias in $\hat{N}_e$) caused by an increase in the number of parents responsible for local samples. The LD method is surprisingly robust to equilibrium migration. Effects of mixture LD are small for all values of migration rate (m), and effects of additional parents are also small unless m is high in genetic terms. LD estimates of N_e therefore accurately reflect local (subpopulation) N_e unless $m > \sim$5–10%. With higher m, $\hat{N}_e$ converges on the global (metapopulation) N_e. Two general exceptions were observed. First, equilibrium migration that is rare and hence episodic can occasionally lead to substantial mixture LD, especially when sample size is small. Second, nonequilibrium, pulse migration of strongly divergent individuals can also create strong mixture LD and depress estimates of local N_e. In both cases, assignment tests, Bayesian clustering, and other methods often will allow identification of recent immigrants that strongly influence results. In simulations involving equilibrium migration, the standard LD method performed better than a method designed to jointly estimate N_e and m. The above results assume loci are not physically linked; for tightly linked loci, the LD signal from past migration events can persist for many generations, with consequences for N_e estimates that remain to be evaluated.

INTEREST in estimating the contemporary effective size (N_e) of natural populations using genetic methods is growing apace (reviewed by Leberg 2005; Wang 2005; Luikart *et al.* 2010), spurred by several major factors: (1) difficulty of collecting sufficient demographic information to calculate N_e directly, (2) rapidly increasing availability (and declining costs) of polymorphic genetic markers, and (3) increased development of software implementing new statistical methods. Until very recently, most genetically based estimates of contemporary N_e have used the temporal method, which requires at least two samples spaced in time (Nei and Tajima 1981; Waples 1989; Wang 2001; Anderson 2005). Notably, a recent review of genetic estimates of N_e (Palstra and Ruzzante 2008) included only the temporal method because so few published estimates were available for other methods. In the last few years, however, considerable interest has focused on estimators that require only a single sample (Nomura 2008; Tallmon *et al.* 2008; Waples and Do 2008; Pudovkin *et al.* 2009; Wang 2009). Underlying models for the one- and two-sample methods both typically involve a number of simplifying assumptions: selective neutrality, discrete generations, random samples, closed populations, and (in most cases) free recombination among loci. Although it is widely recognized that these assumptions are rarely completely satisfied, standard models nevertheless are routinely used to estimate N_e in nature.

In this article, we evaluate sensitivity of the most widely used single-sample method [that based on linkage disequilibrium (LD), defined as nonrandom associations of alleles

doi: 10.1534/genetics.111.132233
Manuscript received July 1, 2011; accepted for publication July 22, 2011
[1]Corresponding author: Northwest Fisheries Science Center, 2725 Montlake Blvd. E., Seattle, WA 98112. E-mail: robin.waples@noaa.gov

at different gene loci] to violations of the standard assumption that the focal population is closed to immigration. Difficulties in delineating population boundaries and quantifying contemporary dispersal make it important to consider the effect of this assumption being violated. Migration poses an interesting theoretical problem, as gene flow can have two opposing influences on LD. First, when used with unlinked markers, the LD method and other single-sample estimators provide an estimate of the effective number of parents that produced the cohort from which the sample was drawn (Waples 2005), and a sample that contains a number of immigrants is drawn from a larger total pool of parents than a sample derived only from local breeders. This suggests that migration could upwardly bias estimates of local N_e. On the other hand, immigrants that differ genetically from local individuals can create LD due to population mixture or admixture (Nei and Li 1973; Sinnock 1975), and this could downwardly bias estimates of local N_e (as suggested by Park 2011). We used both analytical and numerical methods to evaluate the relative importance of these two potential sources of bias under a variety of equilibrium and nonequilibrium scenarios (different population sizes, sample sizes, and migration rates).

Methods

Theory

The magnitude of disequilibrium (D) between alleles at two gene loci is defined as the difference between the observed frequency of a two-locus gamete and its expected frequency, based on population allele frequencies and assuming random assortment. D can be estimated directly from gametic frequencies. For most nonmodel species, however, only genotypic data are available, in which case gametic frequencies cannot be reconstructed with certainty because of ambiguity related to double heterozygotes. In that case, the most widely used method for estimating D is Burrows' composite delta (Δ) method (Weir 1979, 1996), which is simple to calculate and does not depend on the assumption of random mating. Because both D and Δ are sensitive to allele frequency, a standardized form of linkage disequilibrium (r) is often used, which can be interpreted as a correlation coefficient for alleles at different gene loci. Both D and r can be either positive or negative, so the squared terms D^2 and r^2 are often used when one is interested in the magnitude, rather than the direction, of linkage disequilibrium.

The premise of the LD method is that the magnitude of random association of alleles at different gene loci is determined by three variables: N_e, the number of individuals sampled (S), and the recombination rate between loci (c). For monoecious species or dioecious species with random mating and no permanent pair bonds,

$$E(\hat{r}^2) \approx \frac{(1-c)^2 + c^2}{2N_e c(2-c)} + \frac{1}{S} \quad (1)$$

(Weir and Hill 1980; Hill 1981). For most natural populations, the recombination fraction will not be known. However, unless the number of markers is large or the number of chromosomes is small (*e.g.*, as in *Drosophila* spp.), it might be reasonable to assume that the loci are unlinked ($c = 0.5$). Under that assumption, Equation 1 simplifies to

$$E(\hat{r}^2) \approx \frac{1}{3N_e} + \frac{1}{S}. \quad (2)$$

Equation 2 shows that the expectation of r^2 has two components: one due to drift created by reproduction of a finite effective number of parents ($1/(3N_e)$) and one due to sampling a finite number of individuals ($1/S$).

Equations 1 and 2 assume selective neutrality and a closed, panmictic population. Many (perhaps most) natural populations are connected at least sporadically to other populations through migration. At any point in time, therefore, a population of interest might contain individuals derived from more than one gene pool. Such a mixture creates the well-known Wahlund effect (Wahlund 1928), which is manifested as a deficiency of heterozygotes in comparison to the single-locus Hardy–Weinberg expected frequency. Mixtures also create a kind of two-locus Wahlund effect that is detectable as linkage disequilibrium (Nei and Li 1973; Sinnock 1975). The magnitudes of both the one-locus and two-locus Wahlund effects are determined by (a) mixture fraction and (b) allele frequency differences at the loci under consideration. Whereas the single-locus Wahlund effect disappears with a single generation of random mating, LD at unlinked loci decays only asymptotically at a rate of 50%/generation. Therefore, the two-locus Wahlund effect encompasses both population mixture in the current generation and population admixture from recent generations.

Simulated data

In the *Appendix*, we use analytical approximations to compare the expected magnitude of LD arising from both drift and population mixture/admixture, and this allows us to predict the relative influence of these two forces on estimates of local N_e. To test our predictions, we simulated genetic data for metapopulations of fixed size $N = 1000$ individuals, divided into either $n = 2$ subpopulations of $N = 500$ or $n = 10$ subpopulations of $N = 100$. We used a Wright–Fisher island model, so each subpopulation had a constant number $N = 100$ or 500 ideal individuals and also local $N_e = 100$ or 500. Migration rates (m) were symmetrical and assumed values of 0, 0.001, 0.01, 0.05, 0.1, 0.25 0.5, and 0.9 (the latter for $n = 10$ only). EasyPop (Balloux 2001) was used to simulate genotypic data for 20 loci in a K-allele model with a maximum of 10 allelic states. This produced data that were "microsat-like" in terms of the number and frequency of alleles, but we did not attempt to mimic the stepwise mutation model of microsatellites. Simulations were initialized with the maximal diversity option, and populations were followed for 200–300 generations before collecting data. This provided ample time to achieve migration–drift equilibrium and, with a mutation rate of 5×10^{-4}, produced a quasi-equilibrium between mutation, migration, and drift and levels of genetic variability (average heterozygosity = 0.4–0.8) comparable to those seen in most natural populations.

We also simulated nonequilibrium migration scenarios, which could involve sudden infusion of substantial numbers of genetically divergent individuals into a local population. This was accomplished by allowing, for the last generation in the simulation, migration rate to increase by a factor of either 2 (2× scenario) or 10 (10× scenario). Samples were taken in the same generation as the migration, so for these scenarios the samples included mixtures of pure F_0 individuals from two or more subpopulations, in addition to any residual admixture accrued from previous generations at the equilibrium migration rate. For each parameter set, we simulated two types of metapopulations with 1000 individuals each: 10 replicates of ($n = 10$, local $N = 100$) and 50 replicates of ($n = 2$, local $N = 500$). Each parameter set thus produced 100 replicate subpopulations for each metapopulation type, and we assessed bias by computing the harmonic mean $\hat{N}_e$ over all replicate subpopulations and comparing it to the number of ideal individuals in each subpopulation (N). Because the distribution of $\hat{N}_e$ can be strongly skewed with a long tail of high values, and because the drift signal is an inverse function of N_e, the harmonic mean is routinely used to evaluate bias in N_e estimators (*e.g.*, Nei and Tajima 1981; Waples 1989; Wang 2001, 2009; Jorde and Ryman 2007; Nomura 2008). For more details on this topic, see Waples and Do (2010).

Estimating N_e

At the end of each simulation, samples of S individuals were taken from each subpopulation, and the program LDNe (Waples and Do 2008) was used to estimate local effective size. In the derivation of Equations 1 and 2, second-order terms were ignored, which can lead to substantial bias in $\hat{N}_e$ depending on the ratio S/N_e (England *et al.* 2006). LDNe implements the bias correction method of Waples (2006) and uses the Burrows estimator as described by Weir (1996). Unless otherwise noted, we set P_{crit} in the program to screen out alleles at frequency <0.02; Waples and Do (2010) found that this criterion provides a generally good balance between maximizing precision and minimizing bias.

One published single-sample method (Vitalis and Couvet 2001) uses both one- and two-locus identity measures to jointly estimate N_e and m. We evaluated performance of this method and compared it to LDNe using simulated data as described above for two migration scenarios: $m = 0$ and 0.05. For both scenarios, we simulated 10 replicate island-model metapopulations with with $n = 10$ and $N_e = 100$ and took samples of $S = 50$ individuals. For each of the 100 sampled subpopulations, we estimated N_e using LDNe and estimated N_e and m using Vitalis and Couvet's program Estim.

Results

Analytical approximations

As discussed in the *Appendix*, if we ignore effects from sampling individuals, the expected value of r^2 has two components,

$$E(r^2) = \mathrm{Var}\,(r)_{\mathrm{drift}} + [E(r)]^2\,\mathrm{mix}, \qquad (3)$$

which represent the contributions to r^2 from drift and mixture, respectively. In a closed population at equilibrium with constant N, r will vary randomly in the range $[-1, 1]$ (or less, depending on allele frequencies), so that $E(r) = 0$ and there is no mixture LD. In that case, only the drift term is relevant and

$$E(r^2) = \mathrm{Var}(r) \approx \frac{1}{3N_e},$$

on the basis of Weir and Hill (1980) and Hill (1981). We use this standard-model expectation as a point of reference for evaluating the effects of migration on r^2 and $\hat{N}_e$.

Migration changes both the drift and mixture terms in Equation 1, in contrasting ways. First, migration expands the total number of parents that contribute to a local population, and this reduces the drift term. We quantify this effect by calculating how the effective pool of parents (EPP) changes as a function of m, n, and N: EPP $= N/[(1 - m)^2 + m^2/(n - 1)]$ (Equation A1). The expected magnitude of reduction in drift LD due to migration is calculated as $\Delta r^2{}_{\mathrm{drift}} = 1/(3\ \mathrm{EPP}) - 1/(3N)$. At the same time, migration brings together in the local population individuals that are progeny of parents with (potentially very) different suites of allele frequencies. This creates mixture disequilibrium, which will tend to increase overall LD. We quantify this effect by the term $\Delta r^2{}_{\mathrm{mix}}$ (Equation A10). Two primary factors determine the magnitude of mixture LD (Equation A6): population differentiation (all else being equal, genetically divergent populations create more mixture LD) and mixture fraction (LD is highest with equal mixture fractions). In an equilibrium model, these two factors act in opposing ways, as higher migration rates reduce levels of genetic divergence. As a result, under equilibrium conditions mixture LD is expected to be largest at relatively low levels of migration (Figure A1).

Table 1 summarizes results of applying the formulas developed in the *Appendix* to the two general metapopulation scenarios. Some general patterns can be noted. First, in all cases the expected contribution to overall r^2 from population mixture [$\Delta r^2{}_{\mathrm{mix}}$] is at least an order of magnitude smaller than the expected reduction in drift LD from recruiting additional parents [$\Delta r^2{}_{\mathrm{drift}}$]. This occurs because, under the equilibrium model assumed, the population mixture never involves large fractions of genetically divergent individuals; as population divergence increases (and with it the opportunity for creating large mixture LD), migration rate also drops sharply. As a consequence, we expect that in all cases the reductions in LD due to equilibrium migration will

Table 1 Theoretical expectations for contributions to r^2 and $\hat{N}_e$ from drift and population mixture, based on material in the *Appendix*

N	m	$E(r^2_N)^a$	EPP^b	$\Delta r^2_{\mathrm{drift}}{}^c$	$\Delta r^2_{\mathrm{mix}}{}^d$	$E(r^2_{\mathrm{Total}})^e$	$E(\hat{N}_e)^f$	$E(\hat{N}_e/N)$
500	0.001	0.00067	501.0	−0.000001	<0.000001	0.000666	500.9	1.002
500	0.01	0.00067	510.1	−0.000013	<0.000001	0.000654	509.9	1.020
500	0.05	0.00067	552.5	−0.000063	<0.000001	0.000604	552.3	1.105
500	0.1	0.00067	609.8	−0.000120	<0.000001	0.000547	609.5	1.219
500	0.25	0.00067	800.0	−0.000250	<0.000001	0.000417	799.7	1.599
500	0.5	0.00067	1000.0	−0.000333	<0.000001	0.000333	999.8	2.000
100	0.001	0.00333	100.2	−0.000007	0.000002	0.003328	100.1	1.001
100	0.01	0.00333	102.0	−0.000066	0.000011	0.003278	101.7	1.017
100	0.05	0.00333	110.8	−0.000324	0.000013	0.003023	110.3	1.103
100	0.1	0.00333	123.3	−0.000630	0.000012	0.002716	122.7	1.227
100	0.25	0.00333	175.6	−0.001435	0.000009	0.001907	174.8	1.748
100	0.5	0.00333	360.0	−0.002407	0.000004	0.000930	358.5	3.585

An equilibrium island model is assumed, with either $n = 2$ subpopulations with $N = 500$ ideal individuals each or $n = 10$, $N = 100$.

[a] $E(r^2_N) = 1/(3N)$ (*cf.* Equation 2).

[b] *EPP* = effective pool of parents = $N/[(1-m)^2 + m^2/(n-1)]$ (Equation A1).

[c] $\Delta r^2_{\mathrm{drift}} = 1/(3\ \mathrm{EPP}) - 1/(3N)$ (Equation A3).

[d] $\Delta r^2_{\mathrm{mix}}$ is from Equation A9.

[e] $E(r^2_{\mathrm{Total}}) = E(r^2 N) + \Delta r^2_{\mathrm{drift}} + \Delta r^2_{\mathrm{mix}}$.

[f] $E(\hat{N}_e) = 1/[3E(r^2_{\mathrm{Total}})]$.

outweigh any additional mixture LD. Second, the EPP rises only slowly with low levels of migration, so substantial upward biases in local $\hat{N}_e$ are not expected until migration rates are fairly high in genetic terms ($m > 5$–10%). Third, the two metapopulation scenarios are expected to produce generally similar results (indexed by the ratio $\hat{N}_e/N$) for low and moderate migration, but for $m > 0.1$ upward bias is expected to rise faster for $n = 10$, $N = 100$. This is expected because with high migration rates, $\hat{N}_e$ for both scenarios should converge on the overall metapopulation $N_e \sim 1000$, which is a larger multiple of local N_e for the scenario with $N = 100$.

Empirical results from simulations

Equilibrium migration: The main simulation results for equilibrium migration are plotted in Figures 1 and 2. Although our analyses here focus on bias (for an evaluation of precision of the LD method, see Waples and Do 2010), we have plotted empirical confidence intervals (C.I.'s) in Figure 1, and some general patterns are worth noting: (1) C.I.'s are tighter for the [10, 100] scenario because the variance of $\hat{N}_e$ increases with true N_e (Hill 1981); (2) C.I.'s are wider for $mN < 1$ because those scenarios have low genetic diversity in local populations and fewer allelic comparisons for calculating r^2; and (3) C.I.'s are tighter for moderate migration ($mN = 1$–10), because this level of migration is sufficient to maintain high levels of allelic diversity but not so high that $\hat{N}_e$ becomes substantially biased upward.

The simulation results generally agreed with the analytical predictions. For both metapopulation scenarios, the shape of the relationship between $\hat{N}_e/N$ and m was similar to that predicted. Little bias to local $\hat{N}_e$ was found for either scenario for low or moderate m, while $m \geq 0.1$ produced more substantial upward bias. As expected, this latter effect was stronger for $N = 100$ than $N = 500$. As also expected, for $N = 500$ we found no evidence for downward bias in $\hat{N}_e$ that could be attributed to population mixture (see below for discussion of results for $N = 100$). It appears that migration rate (m) is a more reliable indicator than the effective number of migrants (mN_e) of the likely consequences of migration on $\hat{N}_e$ (compare Figure 1A and 1B).

Two important deviations from the predicted patterns are also evident. First, although theoretical derivations in the *Appendix* capture the general pattern of the relationship between $\hat{N}_e$ and m, empirical results showed more upward bias than predicted under high migration rates (Figure 2). The second deviation is that for the scenario with $N = 100$, $n = 10$, we observed a downward bias in $\hat{N}_e$ at low migration rates (harmonic mean $\hat{N}_e = 92.9$ for $m = 0.01$ and 80.2 for $m = 0.001$). With $N = 100$, $m = 0.01$ means that a local population on average receives one immigrant per generation from the metapopulation as a whole, and the rate is one immigrant every 10 generations for $m = 0.001$. Since migration was stochastic, some generations can by chance receive an unusually large number of immigrants. Similarly, if one or a few migrants are unusually successful at reproducing, their offspring can contribute substantial admixture LD to the population for several generations before the associations decay through recombination. Furthermore, because the harmonic mean is strongly affected by occasional low values, and because of the nonlinear effects of m on mixture LD, we expect that the observed reduction in $\hat{N}_e$ for low migration rates was due to a few low values rather than a general across-the-board reduction in $\hat{N}_e$. This is supported by results shown in Figure 3, which compares the distribution of $\hat{N}_e$ for $m = 0.001$ with that under complete isolation. The distributions are generally similar, except that the scenario with rare migration produced four estimates with $\hat{N}_e < 40$ compared to none for $m = 0$. If those four values are omitted, harmonic mean $\hat{N}_e$ becomes 98.0, nearly identical to the value ($\hat{N}_e = 98.3$) for the scenario with no migration. In the rare-migration scenario, the frequency of relatively high estimates was also reduced slightly (Figure 3),

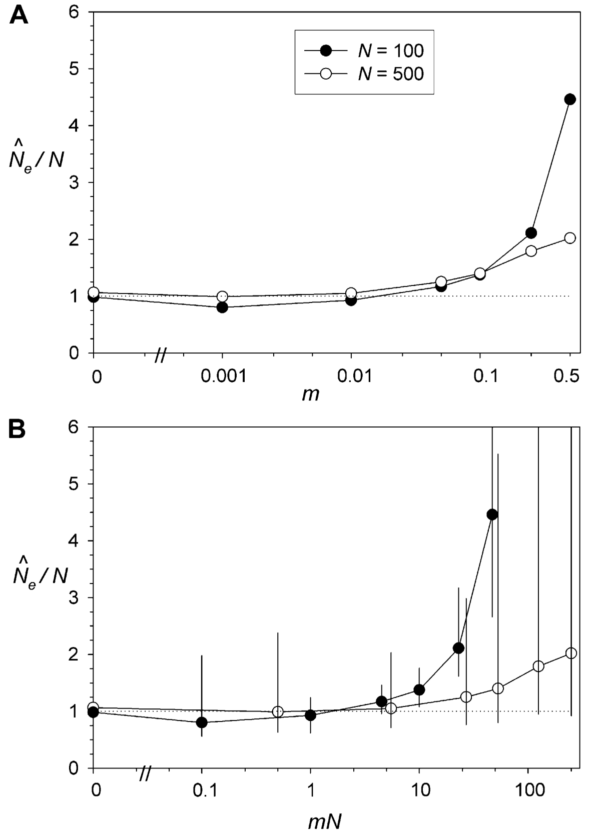

Figure 1 Bias in estimates of local N_e (indicated by the ratio $\hat{N}_e/N$) as a function of amount of migration among subpopulations. Migration is scaled by migration rate (*m*) (A) or number of migrants per generation (*mN*) (B). Local subpopulation size (*N*) was 100 or 500 ideal individuals. Values shown are based on harmonic mean $\hat{N}_e$ calculated using data for 20 loci assayed in *S* = 100 individuals. Vertical lines in B show the central 90% of the empirical distribution of $\hat{N}_e$.

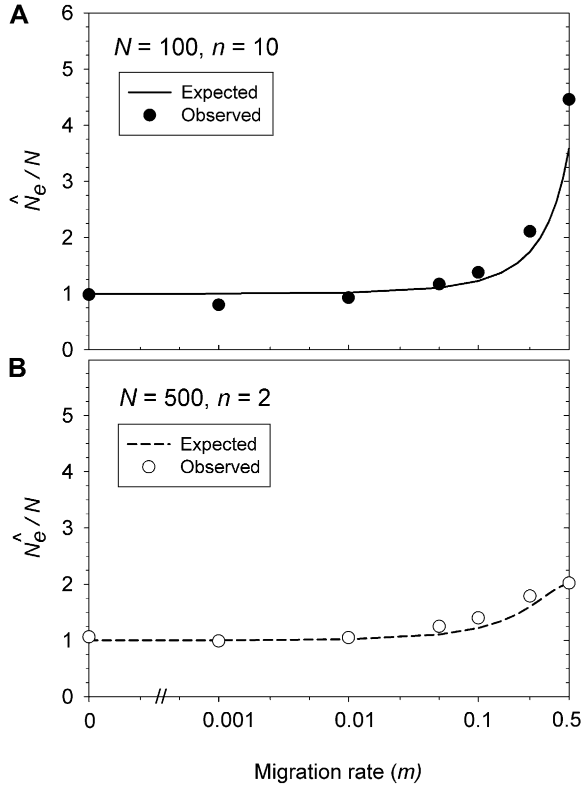

Figure 2 Comparison of observed $\hat{N}_e/N$ from simulations (same data that are plotted in Figure 1) with expected values based on theoretical considerations (from Table 1).

which could be due to a small amount of residual disequilibrium from migrants in previous generations.

To explore this issue further, we examined results for one of the metapopulations that produced one very low estimate ($\hat{N}_e = 13.8$ for population 10). We used Rannala and Mountain's (1997) method as implemented in GeneClass2 (Piry *et al.* 2004) to search for first-generation migrants in the entire metapopulation ($N = 1000$). Three migrants were identified at the $P < 0.001$ level (one each in populations 1, 5, and 9) and were detected with high certainty because the low migration rate produced very strong divergence ($F_{ST} = 0.48$) and essentially nonoverlapping sets of alleles in different populations. Surprisingly, no first-generation migrants were detected in population 10. However, when simulations were used to generate a "likely" range of multilocus genotypes that would be produced by each population (Paetkau *et al.* 2004), seven individuals from population 10 were estimated to have multilocus genotypes with a <1/1000 probability of being produced by a population with allele frequencies observed in population 10. Inspection of these seven individuals showed that in most cases they carried one allele that was rare and one that was common in population 10—the pattern that would be expected for F_1 or backcross progeny of first-generation immigrants. We concluded, therefore, that the low $\hat{N}_e$ for population 10 could be traced to one or a few immigrants in a recent generation that produced a number of descendants.

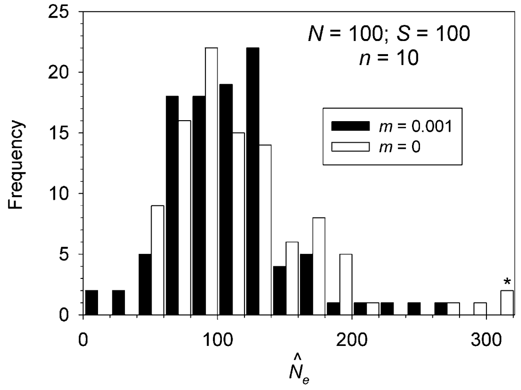

Figure 3 Distribution of $\hat{N}_e$ estimates for scenarios with true N_e = 100 in each local subpopulation and either metapopulations of *n* = 10 subpopulations connected by rare migration events (*m* = 0.001, solid bars) or completely isolated subpopulations (open bars). In both cases, each sample of *S* = 100 individuals was taken from a single subpopulation, and 20 loci were used for the estimate. The bin with the asterisk includes all estimates >300.

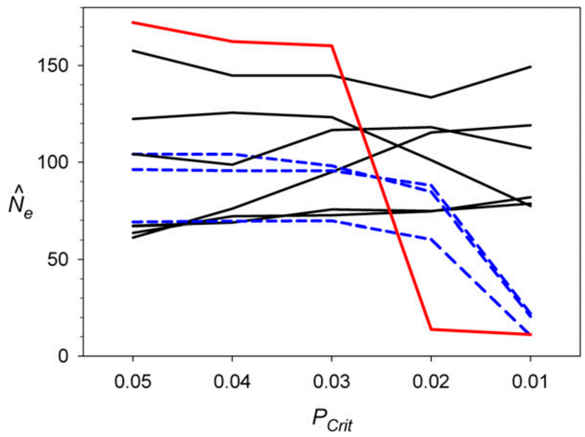

Figure 4 Changes in $\hat{N}_e$ as a function of the criterion for excluding rare alleles (P_{Crit}). Each line shows data for a sample of $S = 100$ from one of the 10 subpopulations in a single metapopulation connected by rare migration ($m = 0.001$, as shown in Figure 3). The three dashed blue lines are the populations in which exactly one first-generation immigrant was detected ($\hat{N}_e$ depressed only for $P_{Crit} = 0.01$). The red line is a population that appears to include a number of descendants of recent immigrants.

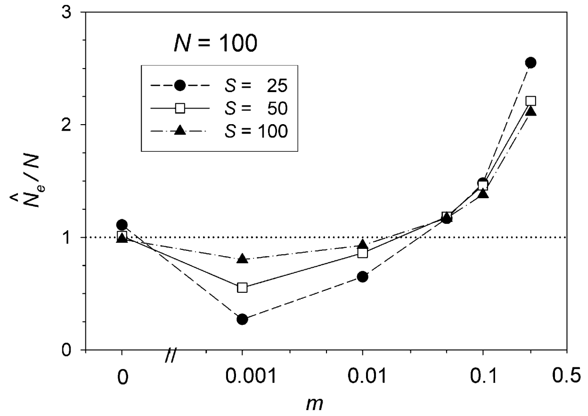

Figure 5 The ratio $\hat{N}_e/N$ as a function of the migration rate (m) among subpopulations. Local subpopulation size (N) was 100 ideal individuals. Values shown are based on harmonic mean $\hat{N}_e$ calculated using data for 20 loci assayed in $S = 25$–100 individuals.

Why did first-generation migrants in population 10 produce low estimates of N_e while those in populations 1, 5, and 9 did not? ($\hat{N}_e = 88.0$, 84.7, and 60.3, respectively, for the latter three populations—lower than average but well within the range expected). The primary reason appears to be an interaction with the criterion used for screening out rare alleles. We used $P_{CRIT} = 0.02$, which excludes alleles at frequency <0.02. Figure 4 shows how $\hat{N}_e$ for each of the 10 populations in the metapopulation varied as a function of P_{CRIT}. For 6 of the populations (Figure 4, black lines), $\hat{N}_e$ showed little variation for P_{CRIT} in the range [0.01–0.05]. The three populations with identified first-generation migrants (Figure 4, blue lines) all had "typical" $\hat{N}_e$ values for $P_{CRIT} = 0.02$–0.05 but sharply reduced values for $P_{CRIT} = 0.01$ ($\hat{N}_e \leq 22$). "Foreign" alleles that occur in only a single first-generation migrant cannot exceed frequency 0.01 in a sample of $S = 100$ individuals, so effects of lone migrants are screened out when $P_{CRIT} > 0.01$ is used. The red line in Figure 4 is for population 10, which shows a different pattern: high estimates ($\hat{N}_e \sim 150$–170) for $P_{CRIT} \geq 0.03$ and very low estimates ($\hat{N}_e = 11$–14) for $P_{CRIT} = 0.02$ or 0.01. When the seven individuals with highly unlikely genotypes were excluded from population 10, estimated effective size jumped dramatically to a value ($\hat{N}_e = 179$ using the $P_{CRIT} = 0.02$ criterion) comparable to the estimates found when rare (presumably mostly recent immigrant) alleles were screened out.

Results discussed so far used relatively large sample sizes ($S = 100$ individuals). Figure 5 shows that the biases discussed above are magnified with smaller samples: for low migration ($m \leq 0.01$), $\hat{N}_e$ is a smaller fraction of N as S decreases, and for high migration ($m \geq 0.1$) $\hat{N}_e$ rises more sharply compared to N for smaller S. It is worth noting that with $S = 50$, alleles carried in a homozygous state by a single immigrant will not be screened out at $P_{CRIT} = 0.02$, and with $S = 25$ the same criterion would include any allele that occurs in even a single copy in the sampled individuals. Waples and Do (2010) found that inclusion of singleton alleles was associated with upwardly biased estimates of N_e and suggested adjusting P_{CRIT} according to sample size to exclude alleles found in only a single copy. Application of this rule would reduce some of the biases seen in Figure 5.

Nonequilibrium migration: Pulse migration at 10 times the equilibrium rate led to substantial biases in $\hat{N}_e$, with the direction of bias depending on whether immigrants were genetically divergent (Figure 6). When background (equilibrium) migration was low enough to lead to strong genetic differences between populations, 10× pulse migration depressed $\hat{N}_e$ to a fraction of the local N_e. Conversely, when genetic differentiation was low due to high background migration, a sudden influx of large numbers of immigrants inflated the estimate of local N_e, reflecting the reality that parents from throughout the metapopulation contributed offspring to the sample. Pulse migration at twice the equilibrium rate had parallel but much more modest effects (Figure 6).

Joint estimates of m and N_e: With equilibrium migration at $m = 0.05$ in a $n = 10$, $N_e = 100$ metapopulation and sample sizes of $S = 50$, $\hat{N}_e$ from Estim was downwardly biased (harmonic mean $\hat{N}_e = 68$) and had a multimodal distribution, with 25% of the estimates below 50, 13% between 125 and 225, and 26% infinite (Figure 7). In contrast, LDNe estimates had a unimodal distribution with a moderate upward bias (harmonic mean $\hat{N}_e = 121$, range 62–790, 73% of estimates between 50 and 150). Simulations using the same parameters but allowing up to 40 alleles per locus and running for 2000 generations before collecting data produced nearly identical Estim results: harmonic mean $\hat{N}_e = 72$, 24% of estimates below 50, and 28% infinite. LDNe performed

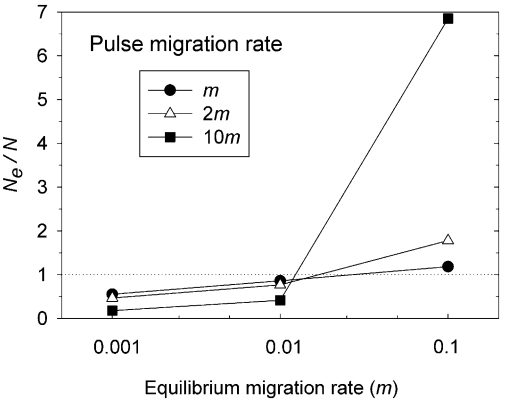

Figure 6 Effects of nonequilibrium (pulse) migration on estimates of local N_e for simulated "island model" metapopulations with $n = 10$ and true local $N_e = 100$. After simulations reached migration–drift equilibrium, a single generation of pulse migration occurred at a level 2 or 10 times the equilibrium rate m, after which samples of $S = 50$ individuals were taken for genetic analysis. Values shown are harmonic mean $\hat{N}_e$ across 100 replicate subpopulations.

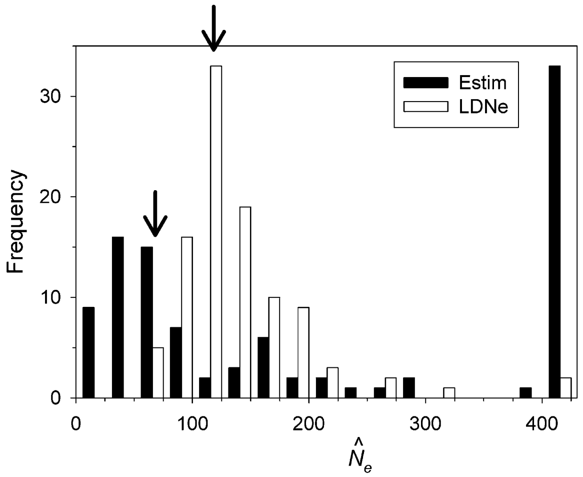

Figure 7 Distribution of $\hat{N}_e$ for simulated data using LDNe and Estim (Vitalis and Couvet 2001). An island model of equilibrium migration was simulated, with $n = 10$, local $N_e = 100$, $m = 0.05$, $S = 50$, and 20 loci. The Estim estimates assumed that the mutation rate was 5×10^{-4}, the value used in the simulations. The last bin on the right includes all estimates >400. The arrows indicate harmonic mean $\hat{N}_e$ for the two methods.

better with the 40-allele data sets, whose greater number of allelic comparisons provided enhanced precision: harmonic mean $\hat{N}_e = 116$, and 100% of estimates fell in the range [50–300] (data not shown). When the subpopulations were completely isolated ($m = 0$), the Estim estimates of N_e were strongly upwardly biased and sensitive to assumed mutation rate: harmonic mean $\hat{N}_e = 149$ assuming $u = 5 \times 10^{-4}$ (the value used in the simulations) and harmonic mean $\hat{N}_e = 360$ assuming $u = 10^{-6}$ (default value in Estim) (data not shown).

Estim also provides estimates of migration rate, which are not sensitive to assumed mutation rate. Mean $\hat{m}$ was 0.01 for the isolation scenario and 0.11 for the $m = 0.05$ scenario. These mean values omitted replicates for which m could not be estimated because $\hat{N}_e$ was infinite (this excluded 51% of the replicates for true $m = 0$ and 26% of the replicates for true $m = 0.05$) (data not shown).

Discussion

The LD method appears to be fairly robust to violations of the closed-population assumption: estimates are largely unbiased with respect to the local, subpopulation N_e unless equilibrium migration rates are high in genetic terms ($m \geq$ 5–10%). In addition, performance of the LD method in estimating N_e for populations connected by migration compared favorably to results for a method that jointly estimates effective size and migration rate (Figure 7). Theoretical and numerical results presented here agree on two major points:

1. The two contrasting effects of migration on linkage disequilibrium (reduced LD due to additional parents and increased LD due to population mixture/admixture) will both be small for $m < 0.05$.
2. For higher equilibrium m, mixture LD is negligible but reductions in LD due to a larger total pool of parents become increasingly important. As m increases, $\hat{N}_e$ converges on a value that represents the global (metapopulation) effective size.

For high migration rates ($m > 0.1$), empirical $\hat{N}_e$ from the simulations was somewhat higher than predicted from theory. Some discrepancy is not surprising, given that a number of rough approximations were used in the theoretical derivations (see *Appendix*). In particular, the algorithm to calculate the EPP might underestimate how this pool increases with migration, because Equation A1 considers only effects in the parental generation, whereas equilibrium levels of LD also reflect the number of parents in several previous generations (Sved 1971).

It should be noted that conclusions about the degree to which migration biases estimates of effective size depend on one's perspective and objectives. We have assumed that the goal is to estimate $\hat{N}_e$ in a local subpopulation, where sampling occurs, so bias has been assessed from that perspective. This is a common application, for example, for those interested in conservation or in studying evolutionary processes in small populations or demes. However, if one were primarily interested in estimating N_e for an entire metapopulation from samples taken in only a local area, conclusions about bias would be different: this approach would lead to a substantial underestimate of metapopulation N_e unless migration were very high in genetic terms.

Although we found little overall effect on harmonic mean $\hat{N}_e$ of low-level, equilibrium migration, if rates of gene flow are low ($mN_e < 1$), migration events are rare and episodic, and when immigrants do arrive they can be quite divergent genetically. When this occurs, immigrants can contribute substantial mixture LD that depresses $\hat{N}_e$ (see Figures 3 and 4). This effect is exacerbated by small samples, within which the occasional immigrant has a proportionally larger genetic effect (Figure 5). Fortunately, a variety of methods are available to help identify recent, genetically divergent immigrants (Rannala and Mountain 1997; Pritchard *et al.* 2000; Wilson and Rannala 2003; Paetkau *et al.* 2004), which could be removed from the analysis if one is interested in estimating local N_e (as was done above for population 10 in Figure 4). In addition, Figure 4 shows that adjusting the criterion for screening out rare alleles can effectively remove bias associated with recent immigrants.

In contrast to results for equilibrium migration, pulse migration can substantially bias estimates of local N_e. In particular, $\hat{N}_e$ can be biased downward if a substantial fraction of genetically divergent individuals suddenly enters the focal population (Figure 6). Note that actual migration might not be required: the same effect could occur if individuals from more than one local population are accidentally included in a single sample. This could happen, for example, if population boundaries are difficult to discern or if samples are collected on feeding grounds or migratory routes, where individuals from more than one breeding population regularly mix. These results emphasize the importance of understanding the biology of the target species (to develop an effective sampling design) and screening resulting samples for evidence that they contain individuals from more than one gene pool.

Our results appear to be consistent with those for the temporal method for estimating N_e: equilibrium migration has relatively little effect on estimates of local N_e unless $m >$ ~5–10% (Wang and Whitlock 2003; G. Luikart, unpublished data). What about effects of migration on other single-sample estimators? In the approximate Bayesian computation method proposed by Tallmon *et al.* (2008), a variety of genetic metrics are used, but the strongest signal comes from r^2. Therefore, we expect that migration would have similar effects on this method. Although the heterozygote excess and LD methods have some similarities (focusing on one- and two-locus disequilibria, respectively), we expect that effects of migration on N_e estimates would be qualitatively different. In the former, the signal is an excess of heterozygotes caused by random allele frequency differences between males and females, whereas the Wahlund effect associated with population mixture creates a deficit of heterozygotes. Thus, immigration would tend to erase the signal of small local N_e and should cause an upward bias in the heterozygote excess method. It would be interesting to examine this quantitatively. Unlike linkage disequilibrium, Hardy–Weinberg equilibrium is restored after a single generation of random mating, so migration in previous generations would not complicate estimates based on the heterozygote excess method.

We expect that the consequences of migration on estimates of N_e from the sibship-reconstruction method of Wang (2009) or the parentage-analysis-without-parents method of Waples and Waples (2011) would depend on the objectives. Presumably, immigrants would be determined to be unrelated to local individuals, which would tend to increase $\hat{N}_e$ for both methods. This might accurately reflect the larger number of parents producing the sampled individuals, but could be misleading if the primary interest was local N_e.

Our simulations produced data with numbers and frequencies of alleles comparable to those found for microsatellite studies of many natural populations. A detailed analysis of performance of the LD method with highly polymorphic markers, including consideration of numbers of loci and alleles, number of individuals sampled, true N_e, and effects of rare alleles, can be found in Waples and Do (2010). The LD method uses only information on allelic state and does not consider evolutionary relationships among alleles, and this has advantages as well as disadvantages. This enhances flexibility of the method, and we found no evidence that results depend on the mutation model used to generate the data (our unpublished data). On the other hand, the method does not take full advantage of information about N_e contained in allelic relationships. One single-sample N_e estimator does explicitly assume a stepwise mutation model and uses allele-size information (OneSamp) (Tallmon *et al.* 2008).

Like other methods for estimating N_e, the LD method makes a number of assumptions besides closed populations that are unlikely to be met entirely in nature. A brief summary of these assumptions follows, along with references to places where interested readers can find additional information.

Stable population size: With stable *N*, LD stabilizes when new disequilibria are generated each generation by drift at the same rate that existing disequilibria break down by recombination. With unlinked loci, the approach to quasi-equilibrium is rapid (only a few generations), although effects of strong bottlenecks might persist a bit longer (Sved 1971; Waples 2005, 2006).

Discrete generations: Age structure can affect most population-genetic estimators. Waples and Yokota (2007) evaluated effects of overlapping generations on temporal estimates of N_e, but comparable analyses have not been conducted for any single-sample estimator. LD estimates from single cohorts primarily estimate the effective number of breeders that reproduced in that year (Waples 2005). Waples and Do (2010) speculated that for the LD method, a mixed-age sample with the number of age classes approximately equal to the generation length might produce an estimate approximately equal to N_e per generation, but that conjecture remains to be evaluated quantitatively.

Unlinked loci: We assumed unlinked loci because linkage relationships are seldom known for nonmodel species, and it might be reasonable to assume that randomly chosen markers are unlinked. That assumption would become more tenuous if very large numbers of markers are used or if the target species has only a few chromosomes and sex-limited recombination. Linked markers actually provide more precision for estimating N_e, provided the recombination probability is known (Hill 1981). Linked markers also provide greater temporal dimension to inferences about historic population size. If next-generation sequencing technology becomes routine for nonmodel species, it might be feasible to resolve ambiguous haplotypes and gain more detailed information about a population's demographic history. Two recent studies that have used the LD method with human HapMap data demonstrate some of the possibilities. Park (2011) used data for SNPs on different chromosomes to estimate N_e in several human populations. Because the analysis was restricted to unlinked markers, resulting estimates provided information primarily about effective size in the recent past and (the author noted) could have been affected by recent migrations. In contrast, Tenesa *et al.* (2007) focused on pairs of SNPs on the same chromosome, separated by no more than 100 kb, and used a coalescent-based method to estimate recombination rates. Because more tightly linked markers retain historical signals of LD for longer periods of time, use of linked SNPs allowed Tenesa *et al.* to generate a temporal spectrum of estimates that show how human effective size has changed over the last ~5000 generations. Their data suggested a relatively constant N_e of ~2500–7000 for most of that time period, followed by a recent rapid expansion.

Acknowledgments

Peter Smouse and two anonymous reviewers provided useful suggestions on earlier drafts, and Gordon Luikart shared unpublished data. This work benefitted from discussions within the Genetic Monitoring Working Group jointly supported by the National Evolutionary Synthesis Center (Durham, NC) and the National Center for Ecological Analysis and Synthesis (Santa Barbara, CA). P.R.E. was assisted by an Australian Academy of Sciences Visit to North America Fellowship.

Literature Cited

Anderson, E. C., 2005 An efficient Monte Carlo method for estimating N_e from temporally spaced samples using a coalescent based likelihood. Genetics 170: 955–967.

Balloux, F., 2001 *EasyPop* (version 1.7): a computer program for population genetics simulations. J. Hered. 92: 301–302.

Crow, J. F., and K. Aoki, 1984 Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. Proc. Natl. Acad. Sci. USA 81: 6073–6077.

England, P. R., J.-M. Cornuet, P. Berthier, D. A. Tallmon, and G. Luikart, 2006 Estimating effective population size from linkage disequilibrium: severe bias using small samples. Conserv. Genet. 7: 303–308.

Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38: 209–216.

Jorde, P. E., and N. Ryman, 2007 Unbiased estimator for genetic drift and effective population size. Genetics 177: 927–935.

Leberg, P., 2005 Genetic approaches for estimating the effective size of populations. J. Wildl. Manage. 69: 1385–1399.

Luikart, G., N. Ryman, D. A. Tallmon, M. K. Schwartz, and F. W. Allendorf, 2010 Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. Conserv. Genet. 11: 355–373.

Nei, M., and W. Li, 1973 Linkage disequilibrium in subdivided populations. Genetics 75: 213–219.

Nei, M., and F. Tajima, 1981 Genetic drift and estimation of effective population size. Genetics 98: 625–640.

Nomura, T., 2008 Estimation of effective number of breeders from molecular coancestry of single cohort sample. Evol. Appl. 1: 462–474.

Paetkau, D., R. Slade, M. Burden, and A. Estoup, 2004 Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Mol. Ecol. 13: 55–65.

Palstra, F. P., and D. E. Ruzzante, 2008 Genetic estimates of contemporary effective population size: What can they tell us about the importance of genetic stochasticity for wild population performance? Mol. Ecol. 17: 3428–3447.

Park, L., 2011 Effective population size of current human population. Genet. Res. Camb. 93: 105–114.

Piry, S., A. Alapetite, J.-M. Cornuet, D. Paetkau, L. Baudouin *et al.*, 2004 Geneclass2: a software for genetic assignment and first-generation migrant detection. J. Hered. 95: 536–539.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Pudovkin, A. I., O. L. Zhdanova, and D. Hedgecock, 2009 Sampling properties of the heterozygote-excess estimator of the effective number of breeders. Conserv. Genet. 11: 759–771.

Rannala, B., and J. L. Mountain, 1997 Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94: 9197–9201.

Sinnock, P., 1975 The Wahlund effect for the two-locus model. Am. Nat. 109: 565–570.

Sved, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2: 125–141.

Tallmon, D. A., A. Koyuk, G. Luikart, and M. A. Beaumont, 2008 ONeSamp: a program to estimate effective population size using approximate Bayesian computation. Mol. Ecol. Res. 8: 299–301.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.

Vitalis, R., and D. Couvet, 2001 Estimation of effective population size and migration rate from one- and two-locus identity measures. Genetics 157: 911–925.

Wahlund, S., 1928 Zuzammensetzung von populationen und korrelation-serscheiunungen von standpunkt der vererbungslehre aus betrachtet. Hereditas 11: 65–106.

Wang, J., 2001 A pseudo-likelihood method for estimating effective population size from temporally spaced samples. Genet. Res. 78: 243–257.

Wang, J., 2005 Estimation of effective population sizes from data on genetic markers. Philos. Trans. R. Soc. Ser. B 360: 1395–1409.

Wang, J., 2009 A new method for estimating effective population size from a single sample of multilocus genotypes. Mol. Ecol. 18: 2148–2164.

Wang, J. L., and M. C. Whitlock, 2003 Estimating effective population size and migration rates from genetic samples over space and time. Genetics 163: 429–446.

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

Waples, R. S., 2005 Genetic estimates of contemporary effective population size: To what time periods do the estimates apply? Mol. Ecol. 14: 3335–3352.

Waples, R. S., 2006 A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv. Genet. 7: 167–184.

Waples, R. S., and C. Do, 2008 *LdNe*: a program for estimating effective population size from data on linkage disequilibrium. Mol. Ecol. Res. 8: 753–756.

Waples, R. S., and C. Do, 2010 Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. Evol. Appl. 3: 244–262.

Waples, R. S., and P. E. Smouse, 1990 Gametic disequilibrium analysis as a means of identifying mixtures of salmon populations. Am. Fish. Soc. Symp. 7: 439–458.

Waples, R. S., and R. K. Waples, 2011 Inbreeding effective population size and parentage analysis without parents. Mol. Ecol. Res. 11(Suppl. 1): 162–171.

Waples, R. S., and M. Yokota, 2007 Temporal estimates of effective population size in species with overlapping generations. Genetics. 175: 219–233.

Weir, B. S., 1979 Inferences about linkage disequilibrium. Biometrics 35: 235–254.

Weir, B. S., 1996 *Genetic Data Analysis*, Ed. 2. Sinauer Associates, Sunderland, MA.

Weir, B. S., and W. G. Hill, 1980 Effect of mating structure on variation in linkage disequilibrium. Genetics 95: 447–488.

Wilson, G. A., and B. Rannala, 2003 Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163: 1177–1191.

Communicating editor: N. A. Rosenberg

Appendix

We are interested in the magnitude of LD in a single focal subpopulation that is connected by migration to other supopulations. The metapopulation conforms to a finite island model at migration–drift equilibrium, with n subpopulations each having N ideal individuals (so local $N_e = N$). In the present case, n and N can take the values [2500] or [10,100], so the total metapopulation size is always $nN = 1000$. Here, we use m to represent the fraction of individuals that are born in one subpopulation and migrate to another subpopulation before reproduction.

A rough idea of the joint effects of drift and migration on LD can be obtained by considering analytical approximations for the effects of finite population size and population mixture/admixture on expected values for r and r^2. On the basis of the simple relationship

$$E(r^2) = \mathrm{Var}(r) + [E(r)]^2$$

we see that $E(r^2)$ has two components: the variance of r [$\mathrm{Var}(r)$] and the square of the expected value of r. As discussed below, these two components represent the contributions to r^2 from drift and mixture, respectively.

Drift

In a closed population at equilibrium with constant N and no evolutionary forces except drift, the correlation of allele frequencies among loci (r) will vary randomly in the range $[-1, 1]$, so that $E(r) = 0$. However, under drift $E(r^2) = \mathrm{Var}(r)$ will be greater than zero, with its magnitude being an inverse function of effective size and the recombination fraction between loci. Assuming the loci are independent, and ignoring sampling and considering only population parameters in a closed, ideal, random mating population, $\mathrm{Var}(r) = E(r^2{}_{\mathrm{drift}}) \approx 1/(3N_e) = 1/(3N)$ (Hill 1981). [This approximation is biased because it ignores second-order terms in N_e (England *et al.* 2006; Waples 2006), but the effect is relatively small compared to other factors considered here.] In a metapopulation with $m > 0$, the total pool of parents is larger than the local size N, which should tend to reduce drift variance in r. We want an expression for how the effective pool of parents (EPP) and hence $E(r^2{}_{\mathrm{drift}})$ change as a function of m, n, and N.

Intuitively, EPP should reach a maximum when each parent in the metapopulation is equally likely to contribute to the N current individuals in the focal population; this occurs when $m = (n - 1)/n$—that is, when the entire metapopulation is panmictic. Conversely, EPP should reach a minimum when only the local subpopulation is a potential source of parents ($m = 0$). An analogous situation occurs with respect to effective size of a single population: $N_e = N$ when each parent has an equal opportunity to contribute to the next generation, and N_e is reduced if successful reproduction is dominated by a small number of parents. In the present case, for a given N and n, we are interested in how EPP changes as m increases from 0 (maximum skewness in contributions by the different subpopulations) to $(n - 1)/n$ (equality of contributions by each subpopulation). For a single population, inbreeding N_e is related to the inverse of f, where f is the probability that two randomly chosen genes in the progeny generation are identical by descent. For a metapopulation, with respect to the current census of N individuals in a single focal subpopulation, an analogous measure is the probability (P) that two randomly chosen individuals were born in the same subpopulation the previous generation. Our simulated data involve migration of individuals, not gametes, so P must be the sum of two mutually exclusive probabilities: (1) the probability that both individuals were born in the local subpopulation [probability $= (1 - m)^2$] and (2) the probability that both individuals are migrants *and* migrated from the same subpopulation [probability $= m^2/(n - 1)$]. Putting these together leads to

$$P = (1-m)^2 + \frac{m^2}{n-1}$$

and

$$\mathrm{EPP} = \frac{N}{P} = \frac{N}{\left[(1-m)^2 + m^2/(n-1)\right]}. \tag{A1}$$

It is easy to verify that Equation A1 produces the expected result for some simple cases. With $m = 0$, the system collapses to a series of completely isolated populations of size N, and Equation A1 yields N as expected. With panmixia ($m = (n-1)/n$), $P = 1/n$ and EPP $= nN$, the size of the entire metapopulation. Finally, with $m = 1$, $P = 1/(n-1)$ and EPP $= (n-1)N = nN - N$. In this case, everyone migrates away from the local population each generation, so the pool of parents is the remaining $(n-1)N$ individuals in the metapopulation. This is only an approximation because in calculating EPP we have considered only the parental generation, whereas drift LD is also influenced by the effective number of parents in preceding generations. However, for unlinked loci (as considered here), drift LD decays rapidly so that $r^2{}_{\mathrm{drift}}$ is determined primarily by the effective number in the parental generation (Waples 2005, 2006), so the approximation should be fairly good.

After accounting for migration, the expected magnitude of LD due to drift is

$$E\left(r^2_{\mathrm{drift}}\right) \approx \frac{1}{3\ \mathrm{EPP}} = \frac{\left[(1-m)^2 + m^2/(n-1)\right]}{3N}. \tag{A2}$$

The expected change in r^2 in a focal subpopulation that arises from contributions by other parents in the metapopulation can be expressed as

$$\Delta r^2_{\mathrm{drift}} = \frac{1}{3\ \mathrm{EPP}} - \frac{1}{3N} = \frac{\left[(1-m)^2 + m^2/(n-1) - 1\right]}{3N}. \tag{A3}$$

It is apparent that $\Delta r^2{}_{\mathrm{drift}}$ is 0 for $m = 0$ and negative if $m > 0$; that is, all else being equal, migration should reduce LD due to drift and hence increase $\hat{N}_e$.

Migration

Nei and Li (1973) studied LD generated by population mixture and showed that the amount of mixture disequilibrium is a function of the mixture fraction and the magnitude of allele frequency difference between populations. On the basis of this work, Waples and Smouse (1990) and P. Smouse (personal communication) developed the following expression for r for a two-population mixture,

$$r_{\mathrm{mix}} = \frac{m(1-m)(P_1 - P_2)(Q_1 - Q_2)}{\sqrt{\bar{P}_{\mathrm{w}}(1-\bar{P}_{\mathrm{w}})\bar{Q}_{\mathrm{w}}(1-\bar{Q}_{\mathrm{w}})}},$$

where m is the fraction of the mixture derived from population 2, P_1 and P_2 are frequencies of an allele at locus A in populations 1 and 2, respectively, Q_1 and Q_2 are comparable frequencies for locus B, and $\bar{P}_{\mathrm{w}}$ and $\bar{Q}_{\mathrm{w}}$ are weighted mean frequencies in the mixture [$\bar{P}_{\mathrm{w}} = mP_1 + (1-m)P_2$, and $\bar{Q}_{\mathrm{w}}$ is defined similarly]. We are interested in the squared correlation coefficient, r^2, which has expectation $E(r^2) = \mathrm{Var}(r) + [E(r)]^2$. In the previous section we focused on the drift term Var(r); here, we are interested in the nonrandom component of $E(r^2)$, which is captured by directional deviations of r from 0 caused by migration. Therefore, ignoring the drift term,

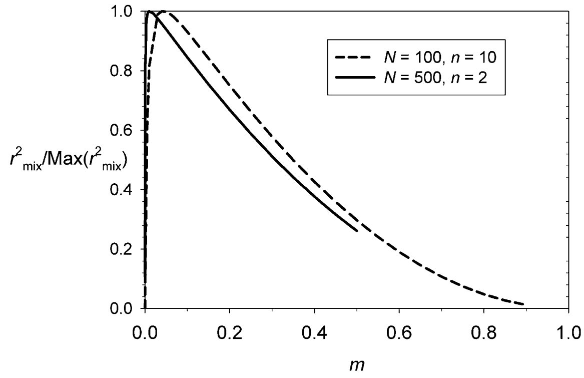

Figure A1 Relationship between mixture LD ($r^2{}_{\mathrm{mix}}$) and migration rate for the two metapopulation scenarios considered here. Max($r^2{}_{\mathrm{mix}}$) is the maximum value of $r^2{}_{\mathrm{mix}}$ over the range $0 \le m \le 1$, based on Equation A9.

$$\begin{aligned} E(r^2_{\mathrm{mix}}) \approx [E(r_{\mathrm{mix}})]^2 &\approx \frac{[m(1-m)]^2(P_1-P_2)^2(Q_1-Q_2)^2}{\bar{P}_{\mathrm{w}}(1-\bar{P}_{\mathrm{w}})\bar{Q}_{\mathrm{w}}(1-\bar{Q}_{\mathrm{w}})} \\ &= [m(1-m)]^2 \frac{(P_1-P_2)^2}{\bar{P}_{\mathrm{w}}(1-\bar{P}_{\mathrm{w}})} \frac{(Q_1-Q_2)^2}{\bar{Q}_{\mathrm{w}}(1-\bar{Q}_{\mathrm{w}})}. \end{aligned} \tag{A4}$$

Note that the two quantities on the right are similar to the standardized variance of allele frequency between populations, F_{ST}:

$$F_{\mathrm{ST}} = \frac{\mathrm{Var}(P)}{\bar{P}(1-\bar{P})}. \tag{A5}$$

For a two-population model, $(P_1 - P_2)^2 = 4\,\mathrm{Var}(P)$; more generally, as n becomes large, $E(P_i - P_j)^2 \Rightarrow 2\,\mathrm{Var}(P)$, where P_i and P_j are allele frequencies in two subpopulations. Considering the two metapopulation scenarios considered here, therefore, Equation A4 can be written as

$$E(r^2_{\mathrm{mix}})_{n=2} \approx [m(1-m)]^2 \frac{4\,\mathrm{Var}(P)}{\bar{P}_{\mathrm{w}}(1-\bar{P}_{\mathrm{w}})} \frac{4\,\mathrm{Var}(Q)}{\bar{Q}_{\mathrm{w}}(1-\bar{Q}_{\mathrm{w}})}$$

$$E(r^2_{\mathrm{mix}})_{n=10} \approx [m(1-m)]^2 \frac{2\,\mathrm{Var}(P)}{\bar{P}_{\mathrm{w}}(1-\bar{P}_{\mathrm{w}})} \frac{2\,\mathrm{Var}(Q)}{\bar{Q}_{\mathrm{w}}(1-\bar{Q}_{\mathrm{w}})}. \tag{A6}$$

For $n > 2$, m can be interpreted as the fraction of immigrants from all other populations combined into the focal subpopulation (population 1). If we ignore for the moment that the mean allele frequencies in the denominator of Equation A4 are weighted by a mixture fraction while those in Equation A5 are unweighted, Equation A4 can be rewritten as a function of F_{ST},

$$E(r^2_{\mathrm{mix}}) \approx [m(1-m)]^2 \alpha F_{\mathrm{ST(A)}} \alpha F_{\mathrm{ST(B)}},$$

where $\alpha = 4$ for $n = 2$ and $\alpha \approx 2$ for $n = 10$. For neutral alleles at unlinked loci, $E(F_{\mathrm{ST}})$ is the same for both loci, leading to

$$\begin{aligned} E(r^2_{\mathrm{mix}})_{n=2} &\approx [m(1-m)]^2 16F^2_{\mathrm{ST}}; \\ E(r^2_{\mathrm{mix}})_{n=10} &\approx [m(1-m)]^2 4F^2_{\mathrm{ST}}. \end{aligned} \tag{A7}$$

Now assume that the mixture process leading to Equation A3 continues until migration–drift equilibrium, with individuals in each population having a constant probability m of migrating to another population each generation. In any given generation, then, after migration the individuals in focal population 1 can be viewed as a mixture composed of a fraction m of individuals that migrated in the current generation from other populations and a fraction $(1 - m)$ of individuals that were born in population 1. We want to find the amount of mixture disequilibrium in population 1 attributable to current generation migrants from other populations.

In Wright's finite island model (as considered here), the expectation of F_{ST} is also a function of migration rate,

$$E(F_{\mathrm{ST}}) \approx \frac{1}{1 + 4NmX},$$

where $X = [n/(n - 1)]^2$ (Crow and Aoki 1984). Our examples involve $n = 2$ or 10, leading to

$$\begin{aligned} E(F_{\mathrm{ST}})_{n=2} &\approx \frac{1}{1 + 16Nm} \\ E(F_{\mathrm{ST}})_{n=10} &\approx \frac{1}{1 + 5Nm}. \end{aligned} \tag{A8}$$

Substituting the expected values from Equation A8 into Equation A7 yields

$$\begin{aligned} E(r^2_{\mathrm{mix}})_{n=2} &\approx \frac{16[m(1-m)]^2}{[1 + 16mN]^2}; \\ E(r^2_{\mathrm{mix}})_{n=10} &\approx \frac{4[m(1-m)]^2}{[1 + 5mN]^2}. \end{aligned} \tag{A9}$$

A plot of the relationships described in Equation A9 (Figure A1) shows that we expect mixture LD to be largest at intermediate migration rates. At higher migration rates, the contribution from a larger mixture fraction is outweighed by a reduction in the genetic distinctiveness of the immigrants.

Putting the two components (Equations A2 and A9) together leads to the following expectations for r^2 in an island-model metapopulation:

$$\begin{aligned} E(r^2)_{(n=2)} &\approx \frac{(1-m)^2 + m^2}{3N}\ (\mathrm{drift}) + \frac{16[m(1-m)]^2}{[1 + 16mN]^2}\ (\mathrm{mix}); \\ E(r^2)_{(n=10)} &\approx \frac{(1-m)^2 + m^2/9}{3N}\ (\mathrm{drift}) + \frac{4[m(1-m)]^2}{[1 + 5mN]^2}\ (\mathrm{mix}). \end{aligned} \tag{A10}$$

These formulas should be regarded as only rough approximations, as they involved many simplifying assumptions. However, the relative importance of the drift and mixture terms is apparent from the form of the equations. The maximum possible value of r^2_{mix} is $<1/N^2$, so the contribution of mixture to r^2 will be small unless N is very small. Conversely, regardless what N is, high levels of migration substantially reduce r^2_{drift} compared to the value that would occur $(1/(3N))$ in a single isolated subpopulation.