LDNE: a program for estimating effective population size from data on linkage disequilibrium

Robin Waples
NOAA, robin.waples@noaa.gov

Chi Do
Northwest Fisheries Science Center

Molecular Ecology Resources (2008) **8**, 753–756 doi: 10.1111/j.1755-0998.2007.02061.x

COMPUTER PROGRAMS

LDNE: a program for estimating effective population size from data on linkage disequilibrium

ROBIN S. WAPLES and CHI DO
Northwest Fisheries Science Center, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

Abstract

LDNE is a program with a Visual Basic interface that implements a recently developed bias correction for estimates of effective population size (N_e) based on linkage disequilibrium data. The program reads genotypic data in standard formats and can accommodate an arbitrary number of samples, individuals, loci, and alleles, as well as two mating systems: random and lifetime monogamy. LDNE calculates separate estimates using different criteria for excluding rare alleles, which facilitates evaluation of data for highly polymorphic markers such as microsatellites. The program also introduces a jackknife method for obtaining confidence intervals that appears to perform better than parametric methods currently in use.

Keywords: bias, genetics, jackknife, N_e/N ratio, precision

Received 4 August 2007; revision accepted 1 November 2007

Genetic methods are increasingly being used to estimate effective population size (N_e) in natural populations. By far the most common approach for estimating N_e is called the temporal method (Nei & Tajima 1981; Waples 1989; Wang 2001) because it requires two or more samples, separated in time, from the same population. In contrast, methods for estimating effective size that require only a single sample [the linkage disequilibrium (LD) method (Hill 1981) and the heterozygote excess method (Pudovkin *et al*. 1996)] have seen relatively little use. This is curious, because any implementation of the temporal method involves two samples, each of which could be used to estimate N_e using a single-sample estimator. The standard linkage disequilibrium method (Hill 1981) was recently shown to be biased when sample size is less than the true (unknown) effective size (England *et al*. 2006). Waples (2006) developed an empirical correction that effectively eliminates the bias, but this bias correction is not implemented in currently available software for estimating N_e (Peel *et al*. 2004). Furthermore, the LD method has seen relatively few practical applications, and its performance has not been evaluated with highly polymorphic genetic markers (such as microsatellites) that are commonly used today (but see Russell & Fewster in press and Tallmon *et al*. 2007). Here, we describe a computer program LDNE that implements the bias-correction method of Waples (2006). LDNE reads genotypic data in standard formats (GENEPOP, Raymond & Rousset 1995; FSTAT, Goudet 2001) and can accommodate an unlimited number of populations, individuals, loci, and alleles.

Calculation of linkage disequilibrium and estimation of N_e

LDNE uses Burrows' Δ, the most common method for estimating linkage disequilibrium, which has several attractive features: it is simple to calculate; it does not depend on the assumption of random mating; and it does not require haplotype data, which is not routinely available for most natural populations. We used Weir's (1979) unbiased estimator of Δ, $\hat{\Delta} = \Delta S/(S-1)$, which adjusts for effects of sampling a finite number (S) of individuals. Δ can be standardized to adjust for the effect of allele frequencies, yielding a correlation coefficient (r_Δ), which forms the basis for estimating N_e. Separate values of r_Δ are calculated for each pair of alleles at each pair of loci. For a comparison of allele A at locus i with allele B at locus j, the estimator of r_Δ is (Weir 1996; Waples 2006)

$$\hat{r}_\Delta = \frac{\hat{\Delta}}{\sqrt{[\hat{p}(1-\hat{p})+(h_i-\hat{p}^2)][\hat{q}(1-\hat{q})+(h_j-\hat{q}^2)]}}, \qquad \text{(eqn 1)}$$

Correspondence: Robin Waples, Fax: (206) 860 3335; E-mail: robin.waples@noaa.gov

Table 1 Parameters for estimating N_e for large and small sample sizes (S) under two mating systems. $\hat{r}^{2'} = \bar{r}^2 - E(\hat{r}^2_{\text{sample}})$ is the empirical $\bar{r}^2$ after subtracting the expected contribution from sampling for that model and sample size. Modified from Waples (2006).

	$S \geq 30$	$S < 30$
Random mating		
$E(\hat{r}^2_{\text{sample}})$	$1/S + 3.19/S^2$	$0.0018 + 0.907/S + 4.44/S^2$
$\hat{N}_e$	$\dfrac{1/3 + \sqrt{1/9 - 2.76\hat{r}^{2'}}}{2\hat{r}^{2'}}$	$\dfrac{0.308 + \sqrt{0.308^2 - 2.08\hat{r}^{2'}}}{2\hat{r}^{2'}}$
Monogamy		
$E(\hat{r}^2_{\text{sample}})$	$1/S + 3.19/S^2$	$0.0018 + 0.907/\text{S} + 4.44/S^2$
$\hat{N}_e$	$\dfrac{2/3 + \sqrt{4/9 - 7.2\hat{r}^{2'}}}{2\hat{r}^{2'}}$	$\dfrac{0.618 + \sqrt{0.618^2 - 5.24\hat{r}^{2'}}}{2\hat{r}^{2'}}$

where h_i and h_j are the observed frequencies of AA and BB homozygotes at loci i and j, respectively, and $\hat{p}$ and $\hat{q}$ are sample frequencies of alleles A and B.

The expected value of $\hat{r}^2_\Delta$ is a function of N_e, S, the recombination rate between loci, and the mating system. Although it is possible to estimate N_e from data for physically linked markers (Hill 1981), such information is rarely available for natural populations. LDNE therefore assumes that the loci under consideration are freely recombining. Waples (2006) empirically derived bias-corrected estimators for N_e for two mating systems (random mating and permanent pair bonds = monogamy), and those formulae (Table 1) are used in LDNE.

N_e is estimated from the overall mean $\hat{r}^2_\Delta$ averaged across multiple loci and alleles, which is computed as follows. For each pair of loci i and j, with k_i and k_j alleles, respectively, $\hat{r}^2_\Delta$ is computed (equation 1) for each of the $k_i{*}k_j$ allelic combinations, and a mean of these allele-pair estimates ($\hat{r}^2_\Delta i, j$) is calculated for that pair of loci. If L loci are used, there are $L(L{-}1)/2$ different $\hat{r}^2_\Delta i, j$ values. Next, two factors are considered in determining the proper weights to give to each $\hat{r}^2_\Delta i, j$ value in calculating the overall mean $\bar{r}^2_\Delta$: the number of independent alleles and the sample size. Since a locus with k alleles has the equivalent of $k - 1$ independent alleles, each $\hat{r}^2_\Delta i, j$ is based on the equivalent of $n_{i,j} = (k_i - 1) * (k_j - 1)$ independent comparisons. With missing data, the sample size $S_{i,j}$ can differ among locus pairs (see next section), and $\hat{r}^2_\Delta i, j$ values based on larger sample sizes should receive greater weight. LDNE uses weights that are inversely proportional to variances. Hill (1981) provided an approximate formula for the coefficient of variation of $\hat{N}_e$ based on linkage disequilibrium data. For unlinked loci, and using the current notation, this can be rearranged to provide an approximate variance for $\hat{N}_e$ associated with $\hat{r}^2_\Delta i, j$:

$$Var(\hat{N}_{e(i,j)}) \approx \frac{2N_e^2}{n_{i,j}}\left[1 + \frac{3N_e}{S_{i,j}}\right]^2. \qquad \text{(eqn 2)}$$

Equation 2 is a function of the true (unknown) N_e. However, if we assume that sample size is small compared to effective size, the last term in brackets dominates, leading to

$$Var(\hat{N}_{e(i,j)}) \approx \frac{18N_e^4}{n_{i,j}(S_{i,j})^2}. \qquad \text{(eqn 3)}$$

Inverting and ignoring the constant leads to the following proportional weights for each $\hat{r}^2_\Delta i, j$ value: $w_{i,j} \propto n_{i,j}\,(S_{i,j})^2$.

Missing data

Since $E(\bar{r}^2)$ and $\hat{N}_e$ depend on sample size, the effective size for each sample has to be adjusted to account for missing data. For each pair of loci i and j, the sample size $S_{i,j}$ was computed as the number of individuals with scored genotypes for both loci. The overall effective sample size was computed as the weighted harmonic mean of the $S_{i,j}$, with the weights proportional to the $n_{i,j}$. This weighted-harmonic-mean sample size was used in the formulae in Table 1 to estimate N_e. Although LDNE can handle arbitrary amounts of missing data, as a quality control measure users might want to eliminate loci or individuals that cannot be consistently and reliably scored.

Allele frequency

Allele frequencies close to 0 or 1 can affect $\hat{r}^2_\Delta$ and hence $\hat{N}_e$ (Waples 2006), but this topic has not been studied in any comprehensive way. A feature of LDNE facilitates evaluation of the effects of allele frequency: as a default, the program returns separate estimates after excluding all alleles with frequencies less than three different critical values ($P_{\text{crit}} = 0.05, 0.02, 0.01$). The user can choose additional or different P_{crit} values (up to six total) as an option.

Confidence intervals

Parametric confidence intervals for $\hat{N}_e$ can be computed based on the premise that $\Phi \approx 2/n$ (Hill 1981), where $\Phi = Var(\bar{r}^2)/(\bar{r}^2)^2$ is the squared coefficient of variation of $\bar{r}^2$ and $n = \Sigma n_{i,j}$ is the total number of independent comparisons the estimate is based upon. LDNE computes parametric 95% CIs for $\hat{N}_e$ using equation 12 in Waples (2006). However, the locus pairs are not entirely independent (Hill 1981), which means that n overestimates the number of independent comparisons, and as a consequence these parametric CIs are too narrow (Waples 2006). In an attempt to correct this problem, we implemented a jackknife option in LDNE that provides an empirical estimate of the effective number of independent comparisons (n') associated with an overall mean $\hat{r}^2$. New values of $\hat{r}^2$ were computed after eliminating in turn each of the $L(L\text{–}1)/2$ pairs of loci, and these new data were used to estimate $Var(\hat{r}^2)$ (as described in Efron & Gong 1983) and hence Φ. We calculated the effective number of independent comparisons as $n' = 2/\Phi$ and used the result to calculate adjusted parametric CIs using n' and the original $\hat{r}^2$.

An example

We evaluated performance of LDNE with simulated data generated using a different model than the one used by Waples (2006), which provided an independent test of the efficacy of the empirical bias correction. EASYPOP (Balloux 2001) was used to simulate genotypic data for ideal populations of fixed size, with discrete generations, equal sex ratio, and random mating. The mutational model approximated that of microsatellites (mutation rate $\mu = 5 \times 10^{-4}$; 10 possible allelic states). Each simulation was initiated with maximal diversity and $N_e = 500$ and run for 128 generations before taking a sample of S individuals for genetic analysis, at which point average heterozygosity was about 0.8 (comparable to that seen for microsatellites in many natural populations).

Figure 1 shows results of 1000 replicate simulated populations with true $N_e = 500$, sample sizes of $S = 50$ or 200, and data for 20 gene loci. The mean $\hat{r}^2_\Delta$ was computed across all replicates and the result was used to estimate N_e using the equations for random mating and $S > 30$ in Table 1. The top panel shows that bias, measured as the ratio of $\hat{N}_e$ to true N_e, was negligible when $\hat{r}^2_\Delta$ was based on alleles with frequency ≥ 0.1 and increased as lower frequency alleles were included in the analysis. The upward bias in $\hat{N}_e$ was more pronounced for $S = 50$, reaching 30% when alleles at frequency as low as 0.01 were used. The second panel depicts the coefficient of variation of the replicate $\hat{r}^2_\Delta$ values, which is a measure of precision. As expected, precision increased as more alleles were included in the analysis. It is apparent, therefore, that users of the LD method are faced with a tradeoff between bias and precision in dealing with highly polymorphic markers. Lowering P_{crit} allows more allelic combinations and increases precision of $\hat{r}^2_\Delta$ but also increases the upward bias of $\hat{N}_e$.

Another measure of performance is the percentage of CIs that contain the true N_e. Table 2 shows the fraction of 95% CIs of $\hat{N}_e$ that contained the true value ($N_e = 500$) for the simulations shown in Fig. 1. In agreement with Waples (2006), parametric CIs were consistently too narrow and included the true N_e only 84–91% of the time. Adjusted CIs using the jackknife method performed as well as or better than the parametric method in all scenarios considered; however, the jackknife CIs still contained the true N_e less than the nominal 95% of the time (Table 2). For both methods, performance declined slightly with lower P_{crit}, presumably because the point estimate is slightly biased when low frequency alleles are used.

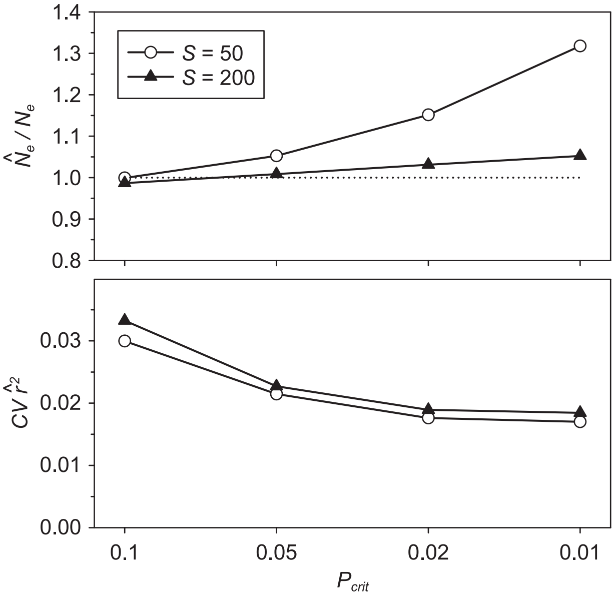

Fig. 1 Results of estimating N_e from simulated data for 1000 replicate populations, each with true $N_e = 500$, data for 20 'microsatellite-like' gene loci, and samples of $S = 50$ or 200 individuals. P_{crit} is the minimum frequency for alleles to be included in the analysis. Top panel: dotted line is the expected result for an unbiased estimator of N_e. Results shown are based on the harmonic mean $\hat{N}_e$ across replicates. Bottom panel: CV is the coefficient of variation of $\hat{r}^2_\Delta$ values across replicates.

Table 2 Percentage of putative 95% confidence intervals that contained the true N_e (500) for the simulations depicted in Figure 1. P_{crit} is the minimum frequency for alleles to be included in the analysis

	P_{crit}	Parametric	Jackknife
$S = 50$	0.1	91.4	94.0
	0.05	89.0	91.5
	0.02	90.3	90.7
	0.01	88.5	88.5
$S = 200$	0.1	87.2	92.1
	0.05	86.5	91.9
	0.02	86.4	90.6
	0.01	84.1	86.9

Collectively, these results are encouraging for several reasons and suggest that the LD method can be generally useful for estimating N_e with highly polymorphic markers. First, including uncommon alleles at highly polymorphic markers considerably enhances precision with relatively modest increases in bias. Second, the jackknife method for generating CIs appears to be partially effective in dealing with nonindependence of $\hat{r}^2_\Delta i, j$ values for different overlapping pairs of loci; however, the method does not eliminate this effect entirely. It must be stressed that these results represent only a small subset of the wide range of conditions that the LD method might be applied to. General conclusions about performance of the method (and the jackknife option) must await a more extensive set of evaluations that considers a wider range of values of S and N_e, as well as different mutation models and numbers of loci and alleles. We are presently conducting some of these evaluations as part of another project (Waples and Do, in preparation).

Assumptions

The LD method and the standard temporal method are both based on some simplifying assumptions (selective neutrality; closed populations; discrete generations) that might not apply to many natural populations. The consequences of violating these assumptions have not been rigorously evaluated for the LD method, but this topic is discussed in Waples (2006). The LD method estimates N_e in the parental generation for the individuals sampled (Waples 2005), although the estimate can also be affected by N_e in the recent past if population size has changed. The method for weighting pairwise $\hat{r}^2_\Delta i, j$ assumes that N_e is large compared to S, which will not always be the case. We expect that violation of this assumption is not likely to have a substantial effect on $\hat{N}_e$, but this should be evaluated with a range of S and N_e values.

LDNE is a FORTRAN 95 program with a Visual Basic interface, written for a personal computer. The FORTRAN code was compiled with the Lahey FORTRAN 95 compiler, version 7.1. The LDNE program, User's Manual, and example data sets can be downloaded from http://fish.washington.edu/xfer/LDNE/.

Acknowledgement

We are grateful to James Russell for useful discussions about calculation of mean $\bar{r}^2_\Delta$ with missing data and variable numbers of alleles per locus.

References

Balloux F (2001) EASYPOP version 1.7: a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, **37**, 36–48.

England PR, Cornuet J-M, Berthier P, Tallmon DA, Luikart G (2006) Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Conservation Genetics*, **7**, 303–308.

Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices. Version 2.9.3. Available from http://www2.unil.ch/popgen/softwares/fstat.htm.

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, **38**, 209–216.

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

Peel D, Ovenden JR, Peel SL (2004) *NEESTIMATOR*: software for estimating effective population size, Version 1.3. Queensland Government, Department of Primary Industries and Fisheries. Available at http://www2.dpi.qld.gov.au/fishweb/13887.html.

Pudovkin AI, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, **144**, 383–387.

Raymond M, Rousset F (1995) GENEPOP version 1.2.: population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Russell JC, Fewster RM Inferences on estimating linkage disequilibrium effective population size. *Environmental and Ecological Statistics*, in press.

Tallmon DA, Koyuk A, Luikart G, Beaumont MA (in press) ONESAMP: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources*, in press.

Wang J (2001) A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical Research*, **78**, 243–257.

Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics*, **121**, 379–391.

Waples RS (2005) Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? *Molecular Ecology*, **14**, 3335–3352.

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, **7**, 167–184.

Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics*, **35**, 235–254.

Weir BS (1996) *Genetic Data Analysis*, 2nd edn. Sinauer Associates, Sunderland, Massachusetts.