On the Potential for Estimating the Effective Number of Breeders From Heterozygote-Excess in Progeny

A. I. Pudovkin,* D. V. Zaykin*,† and D. Hedgecock‡

**Institute of Marine Biology, Vladivostok 690041, Russia, †Department of Statistics, North Carolina State University, Raleigh, North Carolina 27625-8203 and ‡Bodega Marine Laboratory, University of California, Davis, Bodega Bay, California 94923-0247*

Manuscript received July 24, 1995
Accepted for publication May 20, 1996

ABSTRACT

The important parameter of effective population size is rarely estimable directly from demographic data. Indirect estimates of effective population size may be made from genetic data such as temporal variation of allelic frequencies or linkage disequilibrium in cohorts. We suggest here that an indirect estimate of the effective number of breeders might be based on the excess of heterozygosity expected in a cohort of progeny produced by a limited number of males and females. In computer simulations, heterozygote excesses for 30 unlinked loci having various numbers of alleles and allele-frequency profiles were obtained for cohorts produced by samples of breeders drawn from an age-structured population and having known variance in reproductive success and effective number. The 95% confidence limits around the estimate contained the true effective population size in 70 of 72 trials and the Spearman rank correlation of estimated and actual values was 0.991. An estimate based on heterozygote excess might have certain advantages over the previous estimates, requiring only single-locus and single-cohort data, but the sampling error among individuals and the effect of departures from random union of gametes still need to be explored.

THE effective size of a population, N_e, is one of the most important factors for understanding evolutionary processes (WRIGHT 1931) and conserving diversity of genetic resources (LANDE and BARROWCLOUGH 1987; HEDRICK and MILLER 1992). However, the demographic data needed to calculate N_e (see CROW and DENNISTON 1988) are often not available. As an alternative to the calculation of N_e from demographic data, indirect estimation of N_e from genetic data has attracted much attention in recent years (KRIMBAS and TSAKAS 1971; LEWONTIN and KRAKAUER 1973, 1975; PAMILO and VARVIO-AHO 1980; NEI and TAJIMA 1981; POLLACK 1983; MUELLER *et al.* 1985; WAPLES 1989, 1991; HEDGECOCK and SLY 1990; HEDGECOCK *et al.* 1992; BARTLEY *et al.* 1992; HEDGECOCK 1994; JORDE and RYMAN 1995). Two different approaches have been used in these studies, one based on temporal variation of allelic frequencies (NEI and TAJIMA 1981; POLLACK 1983; WAPLES 1989) and the other based on average linkage disequilibrium between pairs of segregating loci (HILL 1981).

Thus far overlooked is another way in which genetic data might be used to estimate the effective number of breeders, which is equivalent to N_e in populations with non-overlapping generations and can be related to N_e for populations with overlapping generations if the age-structure is known (JORDE and RYMAN 1995). When the number of breeders forming the next generation or cohort of progeny is small, allelic frequencies in male and female parents will differ because of binomial sampling error. The consequence of this difference is an excess of heterozygotes, with respect to HARDY-WEINBERG equilibrium proportions, in the cohort of progeny produced by a random union of gametes from the breeders (ROBERTSON 1965; RASMUSSEN 1979). Thus, H', the proportion of heterozygotes expected at a diallelic locus, in a cohort of progeny produced by a small and equal number of males and females, is given by the following expression (FALCONER 1989, p. 67):

$$H' = 2pq + pq/n = 2pq(1 + 1/(2n)), \quad (1)$$

where n is the number of genes in the mothers or fathers and p and q are the frequencies of alleles at this locus, in the population from which parents were drawn.

We suggest to use this excess of heterozygosity expected in progeny, which is dependent on the actual number of parents, for estimation of the effective number of breeders. After rearrangement, (1) becomes

$$n = pq/(H' - 2pq).$$

As n is the number of alleles in male or female parents (which are considered here equal in number), the number of males or females is $n/2$ and the effective number of breeders, N_{eb}, is

$$N_{eb} = n/2 + n/2 = n.$$

Thus, we can write

Corresponding author: Dennis Hedgecock, Bodega Marine Laboratory, University of California, Davis, Bodega Bay, CA 94923-0247.
E-mail: dehedgecock@ucdavis.edu

TABLE 1

Ratios of estimated to actual effective numbers of breeders in 72 simulations

			$\hat{N}_{eb}/\tilde{N}_{eb}$					
			2		3		5	
V_p	N	N_{eb}	eq	tri	eq	tri	eq	tri
1.5	4	3.4	0.84	0.78	0.95	1.10	0.85	0.85
1.9	20	19.5	0.92	1.03	1.23	1.23	1.24	1.01
2.0	100	99.0	0.86	1.03	0.82	1.17	1.01	1.05
2.0	500	499.0	0.94	0.74	1.23	0.94	1.25	0.91
75.7	50	2.5	1.10	0.97	1.11	0.88	1.08	1.21
75.7	250	12.8	0.80	0.93	1.27	0.91	1.19	0.99
75.7	1250	64.3	1.08	0.99	0.95	1.03	1.04	1.08
75.7	6250	321.6	1.19	0.91	1.00	0.94	0.89	1.05
299.6	100	1.3	1.12	1.08	1.08	1.13	1.16	1.13
299.6	500	6.6	1.17	1.03	0.74	0.93	0.83	1.02
299.6	2500	33.1	1.06	0.67	1.35	1.05	0.77	0.91
299.6	12500	165.8	0.70	0.96	0.74	1.10	0.93	1.14

V_p is the variance in number of progeny per parent; N is the number of parents per generation; N_{eb} is the actual effective number of breeders, calculated as $N_{eb} = (4N - 4)/(V_p + 2)$; $\hat{N}_{eb}$ is the estimate calculated from genotypic data for 30 loci by (4); and $\tilde{N}_{eb}$ is the harmonic mean of the actual effective numbers of breeders in the samples drawn from the population for each of the 30 loci (calculated according to CROW and KIMURA 1970). Ratios are grouped by number of alleles per locus (2, 3 and 5) and allele-frequency profile, eq being equally frequent and tri being traingular: 0.667, 0.333; 0.5, 0.333, 0.167; and 0.33, 0.27, 0.2, 0.13, 0.07 for 2, 3, and 5 alleles, respectively.

$$N_{eb} = pq/(H' - 2pq). \quad (2)$$

As H' is the proportion of heterozygotes that is expected to be *observed* in the progeny (given that progeny were derived from a limited number of parents), we may denote it as H_{obs}. On the other hand, the expected proportion of heterozygotes in the base population under HARDY-WEINBERG equilibrium, $2pq$, is denoted as H_{exp}. Now, (2) can be expressed as

$$N_{eb} = H_{exp}/(2\cdot(H_{obs} - H_{exp})).$$

The ratio $H_{exp}/(H_{obs} - H_{exp})$ is the reciprocal of SELANDER's (1970) index, D, for excess or deficiency of heterozygotes; thus, an estimate of N_{eb} is

$$\hat{N}_{eb} = 1/(2D). \quad (3)$$

FALCONER (1989) considers only tangentially the excess of heterozygotes in progeny produced by a line. For the situation we discuss and simulate below, the dependance of $\hat{N}_{eb}$ on observed and HARDY-WEINBERG expected heterozygosities in progeny is given more exactly by

$$\hat{N}_{eb} = 1/(2D) + 1/(2(D + 1)). \quad (4)$$

Derivation of (4) is given in APPENDIX A.

The above expression is given for two alleles. For a multiallelic locus one should average D over all k alleles:

$$D = (1/k)\cdot\left(\sum D_i\right),$$

where D_i is the excess of heterozygotes for the ith allele,

$$D_i = (H_{obs[i/j]} - H_{exp[i/j]})/H_{exp[i/j]},$$

$H_{obs[i/j]}$ and $H_{exp[i/j]}$ being observed and expected heterozygosities, respectively, for the ith allele (*i.e.*, $i \neq j$).

The potential to estimate the effective number of breeders from heterozygote excesses in cohorts of progeny was investigated using a computer simulation model, which was derived from a model created for another purpose but capable of generating the necessary data. The model simulates genetic drift in a population with a specified schedule of age-specific reproduction. Each generation of progeny is formed in the model by random union of gametes obtained from only a few individuals sampled from a very large adult population. The original model was designed to elucidate the hypothesis that variance in individual reproductive success might be large in certain marine animal populations (HEDGECOCK 1994) and will be discussed fully elsewhere. It is pertinent to the present discussion that the model specifies an exact value for the effective number of breeders, N_{eb}, from the known numbers of males and females forming progeny and the known variance of their reproductive contributions, and yields observed and expected frequencies of heterozygotes in the progeny, among other data.

Simulations were run with different values and combinations of the following population parameters: number of alleles per locus, allele-frequency profile, variance of reproductive contribution, and number of breeders. A total of 72 combinations of parameter settings were simulated (Table 1). For each combination

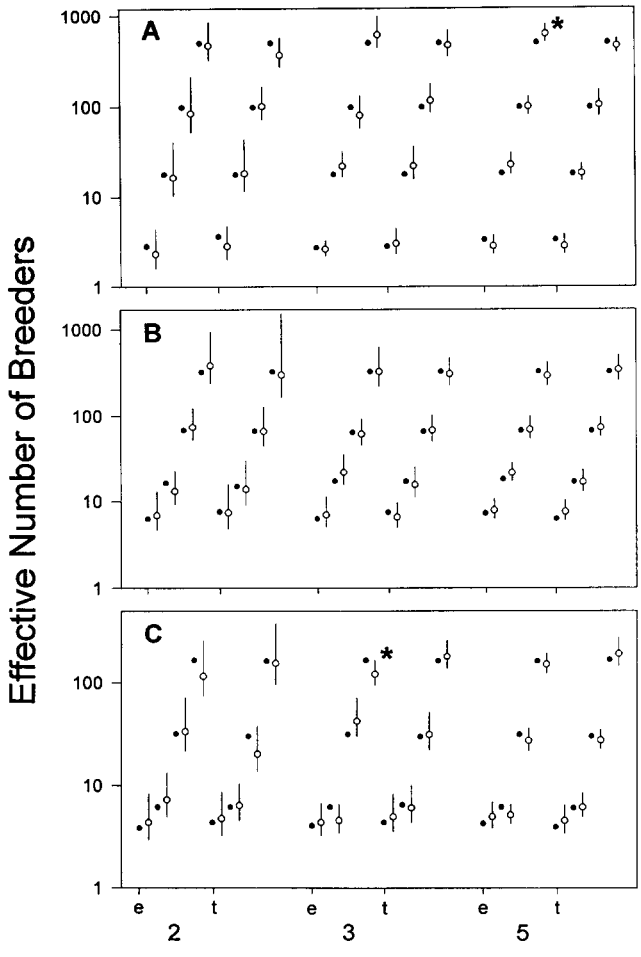

FIGURE 1.—Effective numbers of breeders (N_{eb}) in computer simulations of an age-structured population in which small numbers of males and females produce cohorts of progeny. Actual effective numbers of breeders (●) are calculated from $N_{eb} = (4N - 4)/(V_p + 2)$, where N, the number of breeders (both males and females), and V_p, the variance in number of progeny per parent, are set for each simulation. Estimated effective numbers (○) are means calculated according to (4), using an index of heterozygote excess, D, averaged over 30 loci in each simulation (see text); 95% confidence limits on $\hat{N}_{eb}$ (vertical bars behind open circles) are calculated from the Student's t distribution. Panels A, B and C represent populations with V_p = 2.0, 75.7 and 299.6, respectively. Within each panel, results are arranged along the abscissa for three pairs of simulations in which loci had 2, 3 and 5 alleles, respectively; for a given number of alleles, two allele-frequency profiles were simulated, equal and triangular (see Table 1). For each combination of allele-number and allele-frequency profile within a panel, four different numbers of breeders were simulated (see Table 1). Asterisks denote the two cases in which the actual N_{eb} was not contained within the 95% confidence interval for the estimated N_{eb}.

of model parameters, we made 30 repeats, each corresponding to an individual locus. A mean N_{eb} for each simulation was then estimated by averaging D over the 30 independant loci (open circles in Figure 1). Note that heterozygote excesses were calculated from genotypic proportions in the progeny, which were in turn obtained from allelic frequencies in their parents. Thus,

TABLE 2

Average ratios of estimated to actual effective numbers of breeders for different levels of model parameters

Parameter (levels)[a]	Level 1	2	3	4
Allele-frequency profile (eq, tri)	1.01	1.00		
V_p (2.0, 75.7, 299.6)	1.00	1.02	0.99	
Number of alleles (2, 3, 5)	0.96	1.04	1.02	
N (four values for each V_p)	1.02	1.03	1.00	0.98

[a] Parameters and levels as described in Table 1; V_p is the variance in number of progeny per parent; N is the number of parents per generation.

there were no sampling errors in these calculations because we, in effect, considered *all* progeny, an infinite number. How sampling error in estimating progeny genotypic frequencies affects the precision and accuracy of an estimate of N_{eb} based on heterozygote excess requires further study.

The "true" N_{eb} values for each simulated population (filled circles in Figure 1) were calculated independently from the model's demographic parameters, according to CROW and KIMURA (1970):

$$N_{eb} = 4(N_{eb.m})(N_{eb.f})/(N_{eb.m} + N_{eb.f}),$$

where $N_{eb.m} = (4N_m - 4)/(V_{p.m} + 2)$ and $N_{eb.f} = (4N_f - 4)/(V_{p.f} + 2)$, N_m and N_f being the numbers of male and female parents and $V_{p.m}$ and $V_{p.f}$ being the variances of reproductive contributions by male and female parents, respectively.

To render an idea of variation in $\hat{N}_{eb}$, we calculated 95% confidence intervals for sample mean $\hat{N}_{eb}$s. The distribution of $\hat{N}_{eb}$ is certainly not normal, but "even if the distribution in the original population is far from normal, the distribution of the sample means x tends to become normal under random sampling as the size of sample increases" (SNEDECOR and COCHRAN 1987, p. 41). We therefore calculated 95% confidence intervals from averages of $\hat{N}_{eb}$ over 30 loci, using Student's t distribution. These confidence intervals contain the exact values of N_{eb} calculated from the model's demographics in 70 of 72 (97%) of the simulations (Figure 1). More importantly, the nonparametric Spearman rank correlation between our estimates and the actual N_{eb}s, r_S = 0.991, shows a very strong correspondence between "true" effective numbers of breeders in these computer simulations and estimates of N_{eb} based on excesses of heterozygotes in cohorts of progeny. Performance of our estimate of the effective number of breeders may also be judged from the ratio of the estimate to the actual value, under the various combinations of model parameters (Tables 1 and 2). The grand average of this ratio is 1.005, suggesting that there is no bias in our estimate. The estimate shows no pattern of dependence upon the population parameters investigated, performing equally well under all parameter settings used.

Some of the chief limitations of the suggested approach to estimating effective population size are shared by the other two indirect, genetical methods of estimating N_e: allelic variation is assumed to be selectively neutral, mating within the population is at random, and immigration from other populations is absent. Additional limitations with the heterozygote-excess method may become apparent, however, when departures from random union of gametes and sampling variation among individuals, two important factors ignored in the simulations presented here, are taken into account. If progeny are produced not by random union of gametes but instead by pairwise matings among randomly chosen parents, excesses of heterozygotes can be cancelled in the cohort of progeny by the WAHLUND effect. Thus, the effects of the mating system can potentially mask the excess of heterozygotes expected from small N_{eb} and require more study. At this point our estimate appears to be most appropriate for mating systems that approach the ideal, such as those of marine invertebrates that free-spawn gametes *en masse* for external fertilization. Future development of a statistical estimator of N_{eb} based on heterozygote excess will have also to consider: sampling variation among both individuals and loci, dependence of the estimate on numbers of alleles per locus and allele-frequency profiles, how data from multiple loci with different numbers of alleles should be averaged, and the reliability and power of this estimate compared to the other two genetical methods.

Advantages of this approach to estimating effective population size compared to other estimates may be several. First, an estimate based on heterozygote excess requires genotypic data for a single cohort only, while the temporal genetic-change method requires data on at least two generations (separated by some period of time) and yields an estimate of the average N_e over the interval. On the other hand, if data are available for different years, the heterozygote-excess method may permit tracking of N_{eb} from year to year. Second, the heterozygote-excess method requires only simple computations based on single-locus genotypic data, while the method based on linkage disequilibrium requires more difficult calculations based on multilocus genotypic data. The estimate of N_e based on average linkage disequilibrium, moreover, is influenced by genotypic distributions in the grandparental as well as the parental generation. Third, while temporal genetic change and linkage disequilibria caused by a small N_e can have any sign (decrease or increase in the frequency of an allele; positive or negative values of disequilibrium coefficients), the change in heterozygosity, caused by a small N_{eb}, is usually positive, very seldom zero, and never negative, at least when gametes unite at random. Also, the excess of heterozygosity is seen (if sampling disturbances are ignored) in every locus and in every heterozygote class. This feature may make the suggested method a sensitive measure for hypothesis testing. All three methods are quite independant and may be used in concert to produce more reliable estimates of effective population size.

We thank S. JACKSON for help in preparing the figure. This work was made possible by a grant from the National Science Foundation (OCE-9301416 to D.H.), especially a supplement to this grant for collaboration with A.I.P.; by grants from the Russian State program, "Frontiers in Genetics," the Russian Foundation for Basic Research, the Bodega Marine Laboratory Program for Distinguished Research Faculty (to A.I.P.), by a gift from the Eugene Garfield Foundation and by National Institutes of Health grant GM-43544 to North Carolina State University.

LITERATURE CITED

BARTLEY, D., M. BAGLEY, G. GALL and B. BENTLEY, 1992 Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. Conserv. Biol. **6:** 365–375.

CROW, J. F., and C. DENNISTON, 1988 Inbreeding and variance effective population numbers. Evolution **42:** 482–495.

CROW, J. F., and M. KIMURA, 1970 *An Introduction to Population Genetics Theory.* Harper and Row, New York.

FALCONER, D. S., 1989 *Introduction to Quantitative Genetics,* Ed. 3, Longman Scientific & Technical with J. Wiley & Sons, Inc., New York.

HEDGECOCK, D., 1994 Does variance in reproductive success limit effective population sizes of marine organisms? pp. 122–134 in *Genetics and Evolution of Aquatic Organisms,* edited by A. R. BEAUMONT. Chapman and Hall, London.

HEDGECOCK, D., and F. L. SLY, 1990 Genetic drift and effective population sizes of hatchery-propagated stocks of the Pacific oyster *Crassostrea gigas.* Aquaculture **88:** 21–38.

HEDGECOCK, D., V. CHOW and R. S. WAPLES, 1992 Effective population numbers of shellfish broodstock estimated from temporal variance in allelic frequencies. Aquaculture **108:** 215–232.

HEDRICK, P. W., and P. S. MILLER, 1992 Conservation genetics: techniques and fundamentals. Ecol. Appl. **3:** 30–46.

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, 1995 Temporal allele frequency change and estimation of effective size in populations with overlapping generations. Genetics **139:** 1077–1090.

KRIMBAS, C. B., and S. TSAKAS, 1971 The genetics of *Dacus oleae.* V. Changes of esterase polymorphism in a natural population following insecticide control. Selection or drift? Evolution **25:** 454–462.

LANDE, R., and G. F. BARROWCLOUGH, 1987 Effective population size, genetic variation, and their use in population management, pp. 87–124 in *Viable Populations for Conservation,* edited by M. SOULÉ. Cambridge University Press, Cambridge, England.

LEWONTIN, R. C., and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175–195.

LEWONTIN, R. C., and J. KRAKAUER, 1975 Testing the heterogeneity of F values. Genetics **80:** 397–398.

MUELLER, L. D., B. A. WILCOX, P. R. EHRLICH, D. G. HECKEL and D. D. MURPHY, 1985 A direct assessment of the role of genetic drift in determining allele frequency variation in populations of *Euphydryas editha.* Genetics **110:** 495–511.

NEI, M., and F. TAJIMA, 1981 Genetic drift and estimation of effective population size. Genetics **98:** 625–640.

PAMILO, P., and S. L. VARVIO-AHO, 1980 On the estimation of population size from allele frequency changes. Genetics **95:** 1055–1057.

POLLAK, E., 1983 A new method for estimating the effective population size from allele frequency changes. Genetics **104:** 531–548.

RASMUSSEN, D. I., 1979 Sibling clusters and gene frequencies. Am. Nat. **113:** 948–951.

ROBERTSON, A., 1965 The interpretation of genotypic ratios in domestic animal populations. Anim. Prod. **7:** 319–324.

SELANDER, R. K., 1970 Behavior and genetic variation in natural populations. Am. Zool. **10:** 53–66.

SNEDECOR, G. W., and W. G. COCHRAN, 1987 *Statistical Methods,* Ed. 7, The Iowa State University Press, Ames.

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., 1991 Genetic methods for estimating the effective size of cetacean populations, pp. 279–300 in *Genetic Ecology of Whales and Dolphins*, edited by A. R. HOELZEL. Intl. Whaling Comm., Special Issue 13, London.

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

Communicating editor: G. B. GOLDING

APPENDIX A

When progeny are formed by the random union of gametes from mothers and fathers that differ in allelic frequencies, the proportion of heterozygotes in the progeny is

$$H_{pr} = 2p'q' + \tfrac{1}{2}d^2,$$

where $p' = (p_m + p_f)/2$ and $q' = (q_m + q_f)/2$ are allelic frequencies in the progeny, d is the difference in allelic frequencies between mothers and fathers, $d = (p_m - p_f) = -(q_m - q_f)$, and p_m, q_m and p_f, q_f are frequencies of alleles in male and female breeders, respectively. This relation can be easily obtained from multiplication of gametic frequencies (ROBERTSON 1965). When parents are randomly drawn from an infinite base population with allelic frequencies p_0 and q_0, $\bar{d}$ averaged over all possible parental groups is 0, and $\bar{d}^2$ equals σ_d^2. This, σ_d^2, is the variance of the difference in allelic frequencies between two binomial samples of size n, or $2p_0q_0/n$ (FALCONER 1989), where n is the number of alleles in male or female parents. Thus,

$$H_{pr} = 2p'q' + \tfrac{1}{2}\sigma_d^2 = 2p'q' + p_0q_0/n$$

Denoting p' and q', allelic frequencies in the progeny, as p and q and noting that $N_e = n/2 + n/2 = n$, we can write

$$H_{pr} = 2pq + p_0q_0/N_e. \qquad \text{(A1)}$$

Moreover, as $2pq$ and $2p_0q_0$ are the proportions of heterozygotes expected under random-mating (HARDY-WEINBERG equilibrium), in the first generation of progeny (H_t), and in the base population (H_{t-1}), respectively, they are approximately related in the following way (CROW and KIMURA 1970, p. 104):

$$\lambda = H_t/H_{t-1} = 2pq/(2p_0q_0) \qquad \text{(A2)}$$

where

$$\lambda = (N_e - 1 + \sqrt{(N_e^2 + 1)})/(2N_e). \qquad \text{(A3)}$$

Thus, after substitution of (A2) into (A1) and rearrangement

$$H_{pr} = 2pq + \frac{pq}{\lambda \cdot N_e} = 2pq \cdot \left(1 + \frac{1}{2N_e \cdot \lambda}\right). \qquad \text{(A4)}$$

After substitution of λ, the expression in parentheses becomes

$$\frac{N_e + \sqrt{(N_e^2 + 1)}}{N_e - 1 + \sqrt{(N_e^2 + 1)}}.$$

Denoting $2pq$ as H_{exp}, the expected (HARDY-WEINBERG) heterozygosity in progeny, we can rewrite (A4) as

$$H_{pr} = H_{exp} \cdot \frac{N_e + \sqrt{(N_e^2 + 1)}}{N_e - 1 + \sqrt{(N_e^2 + 1)}}.$$

This can be simplified to

$$\sqrt{(N_e^2 + 1)} + N_e = \frac{H_{pr}}{H_{pr} - H_{exp}}. \qquad \text{(A5)}$$

Let us denote

$$x = \frac{H_{pr}}{H_{pr} - H_{exp}} = \frac{H_{pr} - H_{exp} + H_{exp}}{H_{pr} - H_{exp}}$$

$$= 1 + \frac{H_{exp}}{H_{pr} - H_{exp}}. \qquad \text{(A6)}$$

Here, again, H_{pr} stands for the heterozygosity in progeny, which should be observed under the situation considered (inequality of allelic frequencies in males and females) and H_{exp} is the HARDY-WEINBERG expected heterozygosity in progeny. Then,

$$x = 1 + \frac{1}{D} = \frac{D + 1}{D},$$

where D is SELANDER's index of heterozygote excess (SELANDER 1970):

$$D = \frac{H_{obs} - H_{exp}}{H_{exp}}.$$

From (A5) and (A6) we have

$$\sqrt{(N_e^2 + 1)} = x - N_e$$

or

$$N_e^2 + 1 = x^2 + N_e^2 - 2xN_e.$$

Cancelling N_e^2 and rearranging, we have

$$N_e = \frac{x}{2} - \frac{1}{2x} \qquad \text{(A7)}$$

Finally, substituting x into (A7) we have

$$N_e = \frac{D + 1}{2D} - \frac{1}{2} \cdot \frac{D}{D + 1} = \frac{D^2 + 1 + 2D - D^2}{2D(D + 1)}$$

$$= \frac{1 + D + D}{2D(D + 1)} = \frac{1}{2D} + \frac{1}{2(D + 1)},$$

as given in Equation 4 of the main text.