

Effective Size and Polymorphism of Linked Neutral Loci in Populations Under Directional Selection

Enrique Santiago* and Armando Caballero†

*Departamento Biología Funcional, Universidad de Oviedo, 33071 Oviedo, Spain and †Departamento Bioquímica, Genética e Inmunología, Universidad de Vigo, 36200 Vigo, Spain

Manuscript received August 19, 1997

Accepted for publication April 30, 1998

ABSTRACT

The general theory of the effective size (N_e) for populations under directional selection is extended to cover linkage. N_e is a function of the association between neutral and selected genes generated by finite sampling. This association is reduced by three factors: the recombination rate, the reduction of genetic variance due to drift, and the reduction of genetic variance of the selected genes due to selection. If the genetic size of the genome (L in Morgans) is not extremely small the equation for N_e is

$$N_e = N \exp\left(-\frac{C^2}{(1-Z)L}\right),$$

where N is the number of reproductive individuals, C^2 is the genetic variance for fitness scaled by the squared mean fitness, $(1-Z) = V_m/C^2$ is the rate of reduction of genetic variation per generation and V_m is the mutational input of genetic variation for fitness. The above predictive equation of N_e is valid for the infinitesimal model and for a model of detrimental mutations. The principles of the theory are also applicable to favorable mutation models if there is a continuous flux of advantageous mutations. The predictions are tested by simulation, and the connection with previous results is found and discussed. The reduction of effective size associated with a neutral mutation is progressive over generations until the asymptotic value (the above expression) is reached after a number of generations. The magnitude of the drift process is, therefore, smaller for recent neutral mutations than for old ones. This produces equilibrium values of average heterozygosity and proportion of segregating sites that cannot be formally predicted from the asymptotic N_e , but both parameters can still be predicted by following the drift along the lineage of genes. The spectrum of gene frequencies in a given generation can also be predicted by considering the overlapping of distributions corresponding to mutations that arose in different generations and with different associated effective sizes.

DIRECTIONAL selection generates differences in the reproductive success of individuals, increasing the variance of change in gene frequency and reducing the genetic diversity of neutral alleles. A population, then, behaves for these parameters like an ideal unselected population of size N_e , the effective population size (Wright 1931), which is in general smaller than the actual number of reproductive individuals. If selection acts on a noninherited trait, N_e is simply a function of the variance of the number of progeny per parent, and predictions have been developed for a variety of cases (see review by Caballero 1994).

When differences in fitness are inherited, the effective population size cannot be predicted from the variance of progeny number at a given generation. The drift process is amplified over generations because the random association that originated in a given generation between neutral and selected genes remains in descen-

dants for a number of generations until it is eliminated by segregation and recombination. This problem was first addressed by Robertson (1961), and recently, adequate solutions were given by Woolliams *et al.* (1993) and Santiago and Caballero (1995) for directional selection on quantitative traits determined by an unlinked system of additive loci. But the drift process is larger when selection acts on a linked set of loci as random associations last longer under linkage. Although previous formulations predict the inbreeding coefficient that is calculated by tracing the paths in a genealogy independently of the existence of linkage, this may be different from the real inbreeding coefficient that represents the probability of identity by descent of genes carried by individuals. The reason for this is that both copies of a neutral gene in the same individual do not have identical probabilities of being transmitted to the following generations, as they are embedded in different chromosomes with different selective genes.

Hudson and Kaplan (1995) and, more generally, Nordborg *et al.* (1996) have derived expressions for predicting the nucleotide diversity at neutral loci under

Corresponding author: Enrique Santiago, Departamento de Biología Funcional, Universidad de Oviedo, 33071 Oviedo, Spain.
E-mail: esr@sauron.quimica.uniovi.es

the background selection model (Charlesworth *et al.* 1993), which is based on the continuous appearance of linked deleterious mutations in the population. Barton (1995) made similar derivations for the fixation probability of a favorable allele. In parallel, models dealing with the hitchhiking of neutral genes caused by the spread of selectively favorable mutations at linked loci (Maynard-Smith and Haigh 1974) have been refined in recent years (Wiehe and Stephan 1993, and references therein).

Here, we develop a prediction of the effective population size under linkage, extending the argument of Robertson (1961) and Santiago and Caballero (1995). They have basically shown that the effective size of a population is a function of the variance of the cumulative selective values associated with neutral genes. Simplifying for second-order terms, under random mating and Poisson distribution of family size, the equation for the effective population size becomes

$$N_e = N / (1 + Q^2 C^2),$$

where N is the number of reproductive individuals, C^2 is the genetic variance of fitness of individuals (these are measured relatively to the mean fitness), and Q is the sum of a series of relative terms, the first one being the change of one unit in neutral gene frequency because of new associations created in a given generation and the rest being the remaining fractions of this change in the following generations. For example, for unlinked genes and weak selection, $Q \approx 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 2$ (Robertson 1961), because the average selective advantages of individuals (and, therefore, the changes in gene frequency of neutral alleles) are expected to be reduced by one-half each generation in its descendants, because of segregation and recombination. The complete argument can be found in the derivations leading to Equation 16 in Santiago and Caballero (1995).

Thus, the term $Q^2 C^2$ is the variance of the long-term selective values, and with no linkage and weak selection it approximates $4C^2$. This does not hold, however, under linkage, but the argument can be rebuilt to consider the decline of the association between a neutral gene and selected genes on chromosomes. The problem of the reduction of the effective size under linkage is reduced to the problem of finding the appropriate value of Q^2 , and the same argument used by Santiago and Caballero (1995) can be followed.

Initially, to make predictions independently of gene frequencies and effects, an infinitesimal model (an infinite number of genes of small effect) is considered, but predictive equations are also valid for the background selection model of Charlesworth *et al.* (1993), and we connect our equations with those of Nordborg *et al.* (1996) for this model. Moreover, we apply the principles of the theory to the selective sweep model (Maynard-Smith and Haigh 1974) in considering a continuous flux of weakly advantageous mutations, instead of rare

TABLE 1

Summary of most common notations

| | |
|------------|---|
| N | Number of reproductive individuals |
| N_e | Asymptotic effective population size (with subscript i refers to generation i) |
| $N_{e,HT}$ | Harmonic mean of the effective size between generations 1 and τ |
| l | Chromosome length Morgans |
| ν | Number of chromosomes in the genome |
| L | Total length of the genome in Morgans ($= l\nu$) |
| r | Recombination fraction |
| x | Genetic distance in Morgans |
| t | Heterozygous effect of selected gene |
| μ | Mutation rate per locus, gamete and generation |
| U | Total genomic (diploid) mutation rate |
| π | Equilibrium heterozygosity (with subscript 0 refers to no selection) |
| s | Equilibrium proportion of segregating sites (subscript 0 refers to no selection) |
| Z | Reduction in genetic variance per generation due to selection and drift |
| Z_i | Reduction in genetic variance between generations 1 and i |
| C^2 | Variance of relative fitnesses of individuals |
| c_j^2 | Contribution of selected locus j to the genetic variance for fitness |
| V_m | Mutational input of variance per generation for fitness |
| Q' | Long term selective values of the chromosome initially carrying the neutral allele |
| Q'' | Long term selective values of the homologous chromosome |

mutations of strong favorable effect passing quickly through the population and further recovery of variation by neutral mutation (Wiehe and Stephan 1993). Finally, we show that the two categories of estimators of polymorphism, basically mean heterozygosity per site and proportion of segregating sites, are related to the effective size to different extent, but predictions can still be made from effective size theory.

DERIVATION OF EXPRESSIONS

The general model: We consider a monoecious diploid population with random mating and a constant number of reproductive individuals, N . Every individual in the population is made up of two haploid homologous complements with ν chromosomes l Morgans (M) long each. (Table 1 shows the most common notation used in this article.) Each complement is referred to as a "gamete" (*i.e.*, there are $2N$ gametes in the population). It is assumed that there is no genetic correlation between gametes in parents. The mapping function of Haldane (1919), $r = [1 - \exp(-2x)]/2$, is assumed to relate the recombination fraction r and the genetic distance x in Morgans. A large number n of loci uniformly distributed on the chromosomes determines fit-

ness. Allelic effects can be different for different loci, but gene action is additive within loci; that is, the fitness value of the heterozygote is the average value of the corresponding homozygotes. This latter assumption, however, can be removed for some models (see below). Gene effects are multiplicative between loci. This genetic system is at mutation-selection-drift equilibrium with a mean fitness of one; *i.e.*, each parent has two descendants on average, and the genetic variance for fitness of individuals is C^2 , which is assumed to be small.

As the effects of loci are multiplicative, the relationship between variance of individuals and the contributions of the n selective loci to variation is $C^2 = \prod_{j=1}^n (1 + c_j^2) - 1$, where c_j^2 is the square of the coefficient of variation contributed by locus j , *i.e.*, the variance for fitness of the locus, scaled such that the average fitness is 1.

Consider a neutral locus in the middle of a chromosome. We assume that the neutral alleles at this locus are initially produced by mutation, but this is not a necessary assumption of the model. Due to the finite size of the population, the sampling process generates random associations between the neutral alleles and selected loci. The expected change in gene frequency of the neutral allele (S) is the covariance between the frequency of the allele in gametes (p) and the selective value (f) of individuals carrying the gametes, $S = \text{cov}(p, f)$ (see Santiago and Caballero 1995, p. 1016). We derive this expected change next.

Let p_i be the frequency of an allele of the neutral locus in gamete i (p_i can be 0 or 1). For the moment, we consider one single selected locus j with additive effects of alleles. Locus j is at a genetic distance of xM from the neutral locus. Let us consider a copy of the neutral allele present in a given individual, and let f_j be the selective value contributed by the selected allele present in the same gamete as the neutral allele and f_j'' the selective value contributed by the homologous selected allele in the other gamete. Under random mating, the expected change in gene frequency (*i.e.*, covariance) of the copy of the neutral allele in the first generation (S_1) can be partitioned into the change due to the selected allele in the same gamete as the neutral gene (S_1') and the change due to the homologous selected allele in the other gamete in the same individual (S_1''),

$$S_1 = \text{cov}_1(p_b, f_j' + f_j'') = \text{cov}_1(p_b, f_j') + \text{cov}_1(p_b, f_j'') = S_1' + S_1''.$$

A fraction of the random associations generated in the first generation will remain in the following generations even if the population is expanded to an infinite size after the first generation. The expected value of the remaining covariances in the second generation depends on two factors: the change in expressed genetic variance of the selected locus and the recombination rate between the selected and the neutral loci. The first factor

affects both partial covariances (S_1' and S_1'') in an identical way: Every generation, the genetic variation of the selected locus is assumed to be reduced by selection and drift by a constant proportion $(1 - Z)$. Thus, both covariances are reduced to a proportion Z . On the contrary, the decline because of recombination is different for both partial covariances. The association between the neutral allele and the selective value of the same gamete is maintained with a probability $1 - r = (1 + e^{-2x})/2$ (*i.e.*, if they do not recombine). Therefore the expected partial covariance that remains in generation 2 is

$$S_2' = \text{cov}_2(p_b, f_j') = \text{cov}_1(p_b, f_j') \left(\frac{1 + e^{-2x}}{2} \right) Z = S_1'(1 - r)Z.$$

The effect of recombination on the other partial covariance (between the neutral allele and the selected gene in the other gamete) is opposite to the previous one. Recombination incorporates the selected allele of the homologous gamete into the gamete carrying the neutral gene with a probability r . Therefore, the remaining covariance in generation 2 is

$$S_2'' = \text{cov}_2(p_b, f_j'') = S_1''rZ.$$

In the following generations, both covariances are reduced in the same proportion $(1 - r)Z$ per generation, the selected allele remaining in the same gamete as the neutral gene as the condition for the maintenance of the association, *i.e.*,

$$\begin{aligned} S_3' &= S_2'(1 - r)Z = S_1'(1 - r)^2Z^2, \\ S_3'' &= S_2''(1 - r)Z = S_1''r(1 - r)Z^2, \\ S_4' &= S_3'(1 - r)Z = S_1'(1 - r)^3Z^3, \\ S_4'' &= S_3''(1 - r)Z = S_1''r(1 - r)^2Z^3, \end{aligned}$$

and so on. The sum of all these covariances from generation 1 to infinity is the total change in gene frequency over generations due to the association newly created between the neutral allele and the selected locus j in the initial generation. New associations are created between the neutral gene and the selected locus in successive generations until an asymptotic stage is reached. From Santiago and Caballero (1995) we note that this asymptotic stage is obtained as the sum of the expected changes in gene frequency over generations, given a change of one unit in the first generation,

$$Q_j' = \frac{1}{S_1'} \sum_{i=1}^{\infty} S_i' = \frac{2}{2 - Z - Ze^{-2x}}, \quad (1a)$$

$$Q_j'' = \frac{1}{S_1''} \sum_{i=1}^{\infty} S_i'' = \frac{2(1 - Ze^{-2x})}{2 - Z - Ze^{-2x}}. \quad (1b)$$

[Note that in the derivation of Equation 17 of Santiago and Caballero (1995) the term r has a different meaning than in this article, being the correlation of gene frequencies between mates. In the present derivation

this term is zero because random mating and large population sizes are assumed.]

Robertson (1961) and Santiago and Caballero (1995) showed that the effective population size can be predicted from the variance of the cumulative selective values associated with the neutral gene, as $N_e = N/(1 + \text{Var. of cumulative selective values})$. The variance of the cumulative selective values due to locus j is $Q_j^2 c_j^2$, and this can be again partitioned into the variance due to the selected allele originally located in the gamete with the neutral gene, and the variance due to the selected allele in the other gamete,

$$Q_j^2 c_j^2 = Q_j'^2 \frac{c_j^2}{2} + Q_j''^2 \frac{c_j^2}{2} = \frac{Q_j'^2 + Q_j''^2}{2} c_j^2.$$

If j were the only locus with effect on fitness in the genome, the effective population size would be

$$N_{e(\text{due to } j)} = \frac{N}{1 + c_j^2((Q_j'^2 + Q_j''^2)/2)}. \quad (2)$$

From Equations 1a and 1b we note that Q_j' and Q_j'' take the value $2/(2 - Z) \approx 2$ when the neutral gene and the selected locus are located in different chromosomes (*i.e.*, $x = \infty$), the population size is large, and selection does not change the genetic variance very quickly (*i.e.*, $Z \approx 1$). Therefore, with no linkage, $Q'^2 = Q''^2 \approx 4$, and Equation 2 yields

$$N_e = N/(1 + 4C^2) \quad (3)$$

(Robertson 1961; Santiago and Caballero 1995). Barton (1995) and Nordborg *et al.* (1996, Appendix iii) arrived at the conclusion that with no linkage $N_e = N/(1 + 2C^2)$ instead of Equation 3, but this is not correct as we discuss.

Now consider the n selected loci with different contribution to the variance for fitness. With multiplicative effects among them, the total variance of the cumulative selective values for all the selective loci in the genome is

$$\prod_{j=1}^n \left(1 + c_j^2 \frac{Q_j'^2 + Q_j''^2}{2} \right) - 1.$$

Therefore, the asymptotic value of N_e is

$$N_e = \frac{N}{1 + \prod_{j=1}^n (1 + c_j^2((Q_j'^2 + Q_j''^2)/2))} - 1 \\ \approx N \exp\left(-\sum_{j=1}^n c_j^2 \frac{Q_j'^2 + Q_j''^2}{2}\right). \quad (4a)$$

If c_j^2 and Q_j^2 values are uncorrelated (*i.e.*, independence between Z and c^2 values), Equation 4a reduces to

$$N_e \approx N \exp\left(-C^2 \frac{Q'^2 + Q''^2}{2}\right), \quad (4b)$$

where $Q'^2 = \sum_{j=1}^n Q_j'^2 n$ and $Q''^2 = \sum_{j=1}^n Q_j''^2 n$.

For large populations Q_j'' is nearly 2 when linkage is not very tight (Equation 1b), and it asymptotically tends

to 1 as x tends to 0. Thus, the average Q''^2 ranges from 1 (complete linkage) to 4 (no linkage). Q'^2 approximates $1/r^2$ (Equation 1a) in large populations under weak selection (*i.e.*, $Z \approx 1$), so it may take values much larger than 4 for tight linkage ($r \leq 1/2$). Thus, Q''^2 can usually be neglected relative to Q'^2 , and Equation 4b can be reduced to

$$N_e \approx N \exp\left(-C^2 \frac{Q'^2}{2}\right), \quad (5)$$

without losing much precision. The term Q'^2 , which refers to the effect of selected genes in the same gamete as the neutral gene, has two components. One is due to selected loci in chromosomes other than that of the neutral locus (with probability $[\nu - 1]/\nu$). The other component is due to loci in the chromosome carrying the neutral gene (with probability $1/\nu$). As the neutral gene is assumed to be located in the middle of the chromosome, the second component can be obtained by integration over one-half of the chromosome length. Thus, using Equation 1a and the above probabilities,

$$Q'^2 = \left(\frac{2}{2-Z}\right)^2 \frac{\nu-1}{\nu} + \frac{2}{\nu l} \int_0^{l/2} \left(\frac{2}{2-Z-Ze^{-2x}}\right)^2 dx \\ = \frac{\nu-1}{\nu} \left(\frac{2}{2-Z}\right)^2 \\ + \frac{4}{\nu l} \left(\frac{l}{(2-Z)^2} - \frac{1}{(2-Z)(2-Z-Ze^{-l})} \right) \\ + \frac{\ln(2-Z-Ze^{-l})}{(2-Z)^2} + \frac{1}{(2-Z)(2-2Z)} \\ - \frac{\ln(2-2Z)}{(2-Z)^2}.$$

Numerical analysis (data not shown) indicates that the relevant parameter is the product $L = \nu l$, that is, the genetic size of the genome. Variations in the distribution of the sizes of the chromosomes do not make much difference if the size of the whole genome is constant. Thus, the first term in the above equation can be dropped by setting $\nu = 1$ and substituting l by L :

$$Q'^2 = \frac{4}{L} \left(\frac{L}{(2-Z)^2} - \frac{1}{(2-Z)(2-Z-Ze^{-L})} \right) \\ + \frac{\ln(2-Z-Ze^{-L})}{(2-Z)^2} + \frac{1}{(2-Z)(2-Z)} \\ - \frac{\ln(2-2Z)}{(2-Z)^2}. \quad (6)$$

The following approximations to the above expression can be made:

if $L \rightarrow \infty$ (no linkage), then $Q'^2 = Q''^2 \approx 4$,

and Equation 4b should be used,

if $L \rightarrow 0$ (complete linkage), then $Q'^2 \rightarrow 1/(1-Z)^2$ and Q''^2 can be neglected,

if $L \gg 0$ (say $L > 0.2$), then $Q'^2 \geq 2/(1-Z)L$ and Q''^2 can be neglected. (7)

Then, a general expression for N_e with linkage can be obtained by substituting Q'^2 from Equation 6 into Equation 5. For $L > 0.2$ or so, using the approximation (7), Equation 5 can be simplified to

$$N_e \approx N \exp\left(-\frac{C^2}{(1-Z)L}\right). \quad (8)$$

If selection is weak and linkage is not very tight, *i.e.*, the exponent is smaller than 1 or so, Equation 8 can be expressed in a way more familiar to the classical equations for the effective population size, $N_e \approx N/[1 + C^2/(1-Z)L]$.

Application to particular genetic systems: The above equations for predicting the effective population size are a function of the proportional reduction of the genetic variation $(1-Z)$ at selected loci. Two processes are involved in the dynamics of the genetic variation of loci: selection and drift. It is generally assumed that drift eliminates variation at a constant rate $1/2N_e$. The change in genetic variance due to selection depends on the genetic system. For some models, this change is constant. Particularly, phenotypic selection on an infinitesimal model (Bulmer 1980; Santiago 1998) and the background selection model (Charlesworth *et al.* 1993) erode variation at a rate that is independent of the changes in gene frequency at particular loci.

At equilibrium, the mutational input per generation (V_m) equals the loss of variation by drift and selection. Therefore, the proportion of the expressed genetic variance C^2 that is lost by drift and selection per generation is V_m/C^2 . The remaining fraction of the expressed variation, which is expected to be maintained after one generation of selection, is

$$Z = 1 - \frac{V_m}{C^2}. \quad (9)$$

This term can be substituted in the previous equations to obtain the appropriate Q'^2 and N_e values. For example, the predictive Equation 8 becomes $N_e \approx N \exp[-(C^2)^2/(L V_m)]$.

Infinitesimal model: All the previous equations apply under the infinitesimal model. Favorable and deleterious mutation models reduce to the infinitesimal model if effects are very small. If the population is small, selection is not very strong, and linkage is not very tight, the equilibrium variance can be approximated by $C^2 = 2N_e V_m$ (Lynch and Hill 1986; see Santiago 1998 for a general equation to predict C^2 and V_m under linkage), and $Z = 1 - (V_m/C^2) \approx 1 - 1/(2N_e)$.

Deleterious mutations model: This is equivalent to the background selection model of Charlesworth *et al.* (1993). Predictions of heterozygosity for this model have been developed by Hudson and Kaplan (1995) and more generally by Nordborg *et al.* (1996). In what

follows we show expressions for N_e that generalize those predictions, including the effect of finite populations. Expressions obtained in the previous section are fully applicable, but we consider now a selection coefficient t against heterozygotes. In this case, the assumption previously made of additive gene action within a locus can be removed, because for the background selection model, the effect on the heterozygote and not on the mutant homozygote is critical. If the frequency of a deleterious allele in a particular generation i is q_i , the genetic variance contributed by this locus is proportional to $q_i(1-q_i) \approx q_i$, as the deleterious allele frequency will be generally small, and the expected gene frequency in the next generation is $q_{i+1} \approx q_i - q_i(1-q_i)t \approx q_i(1-t)$. Therefore, the proportional change in the genetic variance of the selected locus due to selection is $q_{i+1}(1-q_{i+1})/q_i(1-q_i) \approx q_{i+1}/q_i = (1-t)$. This is the factor by which genetic variance is changed by selection for this model. This result can also be obtained directly from Equation 9, noting that under mutation-selection balance $C^2 = Ut$ (Crow and Kimura 1970) and $V_m = Ut^2$, where U is the total genomic (diploid) mutation rate for detrimental genes. Combining both drift and selection, the reduction of the association between neutral and selected genes in one generation is approximately

$$Z = (1-t)\left(1 - \frac{1}{2N_e}\right) \approx 1 - t - \frac{1}{2N_e}. \quad (10)$$

Equation 10 can also be obtained from the formula by Keightley and Hill (1988) and Bürger *et al.* (1988), *i.e.*, $C^2 = 2N_e V_m/(1 + 2N_e t)$. Substituting into Equation 9 we get Equation 10. Thus, the appropriate value of Q'_j can be obtained from Equation 1a,

$$Q'_j \approx \frac{1}{r + (1-r)/2N_e + t}. \quad (11)$$

The value of Q'^2 is given by Equation 6; if the genome size is not extremely small ($L > 0.2$), using (7) and (10) we get $Q'^2 \approx 2/[L(t + 1/2N_e)]$, and Equation 8 becomes

$$N_e = N \exp\left(-\frac{C^2}{L(t + 1/2N_e)}\right). \quad (12)$$

When the effect of drift is negligible (*i.e.*, $t \gg 1/2N_e$), then $C^2 = Ut$ and $N_e = N \exp(-U/L) \approx N/[1 + (U/L)]$, which agrees with the approximation of N. H. Barton (unpublished results; see p. 671 of Caballero 1994; Barton 1995). This equation can also be derived from Equation 4 of Nordborg *et al.* (1996), after substituting our Equation 11, which is in fact the average value of the cumulative effect of a large number of selective loci evenly spaced on the genome.

Without recombination ($L = 0$) and $t \gg 1/2N_e$, Equation 6 reduces to $Q'^2 = 1/t^2$. Substituting this and $C^2 = Ut$ into Equation 5, $N_e \approx N \exp[-U/(2t)]$. This equation is identical to the expression of Kimura and Maru-

yama (1966) and Haigh (1978) for the size of the chromosome class with the lowest number of deleterious mutants. As all the chromosomes in the population will be derived from one member of the best class, the size of this chromosome class is indeed the effective population size given in number of chromosomes. Charlesworth *et al.* (1993) found the equivalent algebraic solution for the heterozygosity, $\pi = \pi_0 \exp(-U/2t)$, where π_0 represents the expected heterozygosity if there is no selection.

Favorable mutations model: Assume that the selective values of the three genotypes at any selective locus j are $1, 1 + t$, and $1 + 2t$, respectively. The predictive equations previously shown do not hold if favorable mutations are not effectively neutral (*i.e.*, $t > 1/2N_e$, after Kimura 1983) as changes in variance are dependent on the dynamics of the gene frequencies. However, the general principles of the theory are also applicable for large effects if there is a continuous flux of advantageous mutations. Assume that the frequency of the favorable allele in the generation in which the neutral allele of reference appears by mutation is q_1 . The expected gene frequency of the selected allele in the next generation is $q_2 \approx q_1 + q_1(1 - q_1)t$ and the proportional change in genetic variance due to selection is $q_2(1 - q_2)/q_1(1 - q_1)$. Drift also reduces the genetic variance by $1 - 1/2N_e$, therefore, $Z_1 = (1 - 1/2N_e)q_2(1 - q_2)/q_1(1 - q_1)$. Here we consider only the effect of the gamete associated with the neutral gene (*i.e.*, we neglect Q_j''), obtaining Q_j' with the same argument leading to Equation 1a. If the frequency of recombination between the neutral gene and the selected locus j is r , and the covariance between them in the first generation is S'_1 , the expected changes in gene frequency in the following generations are

$$\begin{aligned} S'_2 &= S'_1(1 - r)Z_1, \\ S'_3 &= S'_1(1 - r)^2 \left(1 - \frac{1}{2N_e}\right)^2 \frac{q_3(1 - q_3)}{q_1(1 - q_1)} \\ &= S'_1(1 - r)^2 Z_2, \\ S'_4 &= S'_1(1 - r)^3 \left(1 - \frac{1}{2N_e}\right)^3 \frac{q_4(1 - q_4)}{q_1(1 - q_1)} \\ &= S'_1(1 - r)^3 Z_3, \end{aligned}$$

and so on. Z_i is the proportional change in genetic variance from the initial generation to generation i due to selection and drift. Therefore, the value of Q_j' given an initial frequency q_1 for the selected locus when the neutral gene appears is (see Equation 1a)

$$Q_j'(q_1) = \frac{1}{S'_1} \sum_{i=0}^{\infty} S'_i = \sum_{i=0}^{\infty} Z_i \left(\frac{1 + e^{-2xt}}{2} \right)^i,$$

where i represents the successive generations and q_i are the sequential frequencies of the selective allele in the consecutive generations, which are obtained as $q_i \approx q_{i-1} + q_{i-1}(1 - q_{i-1})t$. Now, the neutral mutation may appear

when the selected allele has any gene frequency (q_1) in the range 0 to 1. Thus, the appropriate value of Q_j' is the mean weighted value of the $Q_j'(q_1)$ values corresponding to all the possible initial frequencies q_1 in the range 0 to 1, the weights being the product of the proportional contribution of the possible initial frequencies to the observed genetic variance and the probability of being at all the possible values of the initial frequency. In a deterministic mutation-selection model, the probability of having a particular frequency, q , is proportional to $1/[q(1 - q)]$ (Crow and Kimura 1970), and the contribution of the gene to the variance is proportional to $q(1 - q)$. Therefore, the product of these terms is independent of the gene frequency for large $N_e t$. Hence, Q_j' can be approximated as the average value of a number m of $Q_j'(q_1)$ values corresponding to initial frequencies q_1 evenly spaced through the spectrum of gene frequencies from 0 to 1,

$$Q_j' = \frac{\sum_{q_1=1/(m+1)}^{m/(m+1)} Q_j'(q_1)}{m}. \tag{13}$$

The appropriate $Q_j'^2$ value of the neutral locus is the average value of the $Q_j'^2$ values corresponding to all the selective loci j in the genome. This average value has to be substituted into Equation 5. Therefore, although it seems difficult to reach a simple algebraic equation to predict N_e for the model of favorable mutations, the principles previously shown can be applied to find numerical approximations.

ASYMPTOTIC EFFECTIVE SIZE, HETEROZYGOSITY, AND POLYMORPHISM

The parameter N_e that we have derived is the asymptotic effective population size. If a neutral allele appears in the population at a given generation by mutation, the drift process will be initially weak on it, but random associations with selected genes will accumulate over generations making drift increase until an asymptotic value is reached. In the first generation, the magnitude of the drift process on the neutral allele can be quantified by the variance in allele frequency in the first generation. Although this refers only to the particular neutral genes that appeared one generation ago, we refer to it as the effective population size in the first generation,

$$\begin{aligned} N_{e,1} &\approx \frac{N}{1 + C^2((Q_1'^2 + Q_1''^2)/2)} \\ &\approx N \exp\left(-C^2 \frac{Q_1'^2 + Q_1''^2}{2}\right). \end{aligned} \tag{14}$$

Q_1' can be computed as the average value for all the Q_j' values of selected loci, being obviously equal to one, $Q_{j1}' = (1/S'_1) \sum_{j=1}^L S'_j = 1$, so that $Q_1' = 1$. The value of Q_1'' is also 1. As stated before, the asymptotic value of Q'' is less than or equal to 2, and after a few generations it will be much smaller than Q' under linkage, so

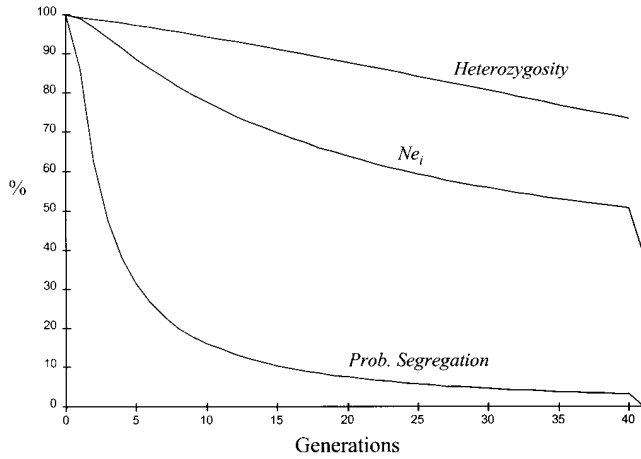


Figure 1.—Reduction of the probability of segregation and the heterozygosity contributed by a locus with a single copy of a neutral gene in the initial generation. The reductions are given as a percentage of the values in the initial generation. The effective population size ($N_{e,i}$) associated with that locus in generation i is also plotted. $N = 100$, $L = 1$, $C^2 = 0.02$, and $t = 0.01$ (deleterious mutations model). The last element in each series corresponds to the asymptotic value (both heterozygosity and probability of segregation equal 0).

we can ignore it in Equation 14 and henceforth. The magnitude of the drift process on the neutral allele from generation 1 to 2 is analogously quantified by the variance in allele frequency from generation 1 to 2, which we refer to as $N_{e,2}$. This is calculated using Q'_2 , the average of all the Q'_{j2} (see the accumulation of terms stated before), obtained as $Q'_{j2} = (1/S'_1)\sum_{i=1}^2 S'_i = 1 + (1-r)Z$, and

$$N_{e,2} \approx N \exp\left(-\frac{C^2 Q_2'^2}{2}\right). \quad (15)$$

From generation 2 to 3, $N_{e,3}$ can be calculated using Q'_3 which is the average of all the Q'_{j3} , obtained as $Q'_{j3} = (1/S'_1)\sum_{i=1}^3 S'_i = 1 + (1-r)Z + (1-r)^2 Z^2$, and so on up to infinite, $Q'_{j,\infty} = Q'_j$, when the asymptotic effective population size, $N_{e,\infty} = N_e$ (equations in the previous sections), is reached.

There is no simple solution for the Q' terms for consecutive generations (except for infinite generations; *i.e.*, Equation 6). Therefore, numerical methods have to be applied to estimate the values of the partial effective sizes in consecutive generations. If genetic variance for fitness is not large, in the first generation $N_{e,1}$ is close to the census size N of the population. In the following generations the effective size drops toward its asymptotic value (see Figure 1). For a new neutral mutation, the decay of genetic variance is $1/2N_{e,1}$ in the first generation, $1/2N_{e,2}$ in the second generation, and so on. A consequence of this cumulative effect of drift on new mutations is that there is not a simple formula to connect asymptotic population size, heterozygosity, and proportion of segregating sites for neutral alleles, as we address next.

Heterozygosity: Under the infinite sites model, the heterozygosity contributed by a new mutation (*i.e.*, with frequency $1/2N$) is $2(1/2N)(1 - 1/2N) \approx 1/N$. Then, with a mutation rate μ per locus and generation, the number of new mutations per generation is $2N\mu$, and the input of heterozygosity per generation is about 2μ . The neutral variability generated by these mutations decreases at an increasing rate, which is a function of the consecutive values of $N_{e,i}$ so the remaining proportion after τ generations is $R_\tau = \prod_{i=1}^\tau (1 - 1/2N_{e,i})$. Therefore, the expected heterozygosity at equilibrium (π) is the sum of the contributions by mutations during all the previous generations,

$$\pi = 2\mu \left(1 + \sum_{\tau=1}^{\infty} R_\tau\right). \quad (16)$$

If there is no selection, or selection acts on a noninherited trait, there is a single value of N_e for the consecutive generations. Thus, $R_\tau = (1 - 1/2N_e)^\tau$, and substituting this into Equation 16, $\pi = 4N_e\mu$, as expected (Crow and Kimura 1970, p. 323). Furthermore, when selection is on an inherited trait and the selective effects are large, the consecutive values of $N_{e,i}$ decay very quickly reaching values close to the asymptotic effective size, N_e , in a few generations. Under this condition, heterozygosity is again well approximated by $\pi = 4N_e\mu$. Otherwise, this equation underestimates heterozygosity because the effective size associated with a mutation is larger than the asymptotic N_e for a long period of time. This is illustrated in Figure 1, which shows the expected heterozygosity for consecutive generations of a neutral allele starting with a single copy in the initial generation. It is observed that the heterozygosity has a rate of reduction lower than that of $N_{e,i}$. However, the degree of disassociation between heterozygosity and the asymptotic N_e is much smaller than that between the proportion of segregating sites and the asymptotic N_e , as we explain next.

Proportion of segregating sites: Under an infinite sites model, the proportion of segregating sites increases by $2N\mu$, the number of new mutations per generation. The equilibrium proportion of segregating sites, s , can be obtained by calculating the probability that mutants appearing in previous generations are still segregating in the current one. Looking backward in time, the remaining fraction of the segregating sites produced τ generations ago is a function of the magnitude of the drift process until the current generation. As we have seen, this magnitude is represented by the partial $N_{e,i}$ values from generation 1 to generation τ and it can be summarized by the harmonic mean $N_{e,H\tau}$ of these τ values, *i.e.*, $1/N_{e,H\tau} = (1/\tau) \sum_{i=1}^\tau (1/N_{e,i})$. Thus, the probability of segregation in the current generation of mutations appeared τ generations ago (P_τ) can be approximated by

$$P_\tau = 1 - \exp\left(-2\frac{N_{e,H\tau}}{N\tau}\right) \quad (17)$$

(Gale 1990, p. 108). This equation gives overestimates of P_τ in the long term, say for $\tau > N_{e,H\tau}$. Therefore, in practice we utilize this equation until the difference for two consecutive generations, $P_\tau - P_{\tau+1}$, is smaller than the expected asymptotic rate of decay $1/2N_e$. After that, the recursive equation $P_{\tau+1} = P_\tau(1 - 1/2N_e)$ is used. The proportion of segregating sites s can be computed as the sum of the remaining contributions from all the previous generations,

$$s = 2 N \mu \sum_{i=0}^{\infty} P_i. \tag{18}$$

The proportion of segregating sites is generally much more dependent on N than on N_e because only a small proportion of new mutations segregate for a long period. For example, if there is no selection, the s value for the whole population is approximately $4N\mu \ln 2N$ (see Ewens 1979). The input of segregating sites per generation is $2N\mu$. At equilibrium, this is also the number of sites that become monomorphic per generation. Therefore, the proportion of polymorphic loci that become monomorphic per generation is $2N\mu/s = 1/(2 \ln 2N)$, which is a relatively large proportion. For example, for a population of $N = 100$, about 10% of the segregating sites are lost by drift every generation. This is also illustrated in Figure 1. The probability of segregation of an initially single-copy neutral allele has most of its reduction in the initial generations. Given this large rate of loss of polymorphic loci per generation, it is clear that the proportion of segregating sites is very dependent on the mutations arising few generations ago and, therefore, the $N_{e,i}$ values of the initial generations have much influence. Because these initial $N_{e,i}$ values are closer to the census size N than to the asymptotic effective size, N_e , the proportion of segregating sites in the whole population is only slightly dependent on N_e . On the contrary, the rate of loss of heterozygosity per generation is relatively small ($1/2N$ with no selection). For example, for a population of size $N = 100$, it is only 0.5% per generation. Therefore, the heterozygosity is more dependent on the asymptotic effective population size, N_e . The above arguments indicate that if the asymptotic N_e is much smaller than the census size N , heterozygosity will be more affected by selection than the proportion of segregating sites because the latter depends strongly on N . This dependence of the asymptotic reduction of s , π , and $N_{e,i}$ on population size predicted under a model of deleterious mutations is shown in Figure 2. The larger the population size the stronger the selection as the mutant effects are assumed to be constant ($t = 0.01$). Reductions of $N_{e,i}$ and π are close and tend to be equal with large N (strong selection), as previously noted by Charlesworth *et al.* (1995) for background selection. The proportion of segregating sites is much less affected by the increase in strength of selection.

Allele frequency spectrum: The application of the classic theory of N_e provides methods to predict the

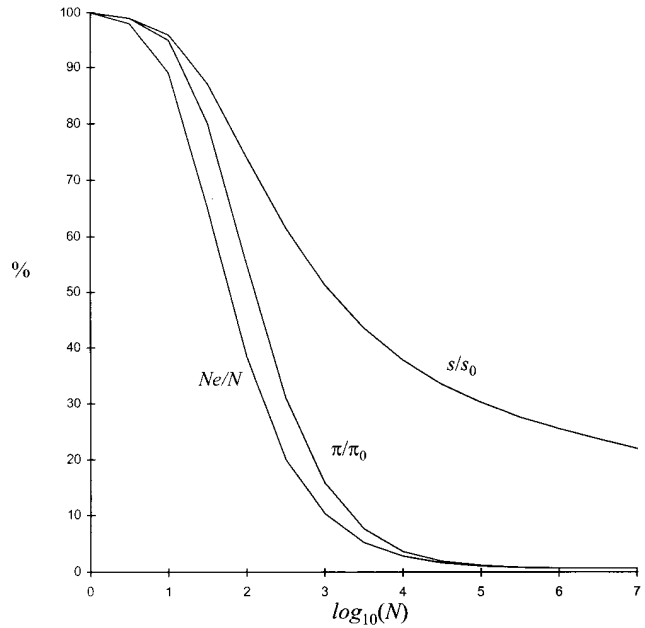


Figure 2.—Example of the dependence of the asymptotic reduction of proportion of segregating sites (s), heterozygosity (π), and effective population size (N_e) on the number of reproductive individuals (N). $C^2 = 0.001$, $t = 0.01$ (deleterious mutations model), and $L = 0$.

spectrum of frequencies of neutral genes a number of generations after their appearance (*e.g.*, Crow and Kimura 1970). According to our results, these methods should consider the evolution of the consecutive values of the effective size for the n generations, but the application would be quite complex. However, the precision is not much affected if the harmonic mean $N_{e,H\tau}$ of the partial $N_{e,i}$ values for the τ generations is used as the constant effective population size for the τ generations. The distribution of neutral gene frequencies in the population can then be computed as a combination of distributions for neutral mutations that appeared in the actual generation, one generation ago, two generations ago, etc., up to infinity. An illustration of this is given in the next section.

EVALUATION OF RESULTS

The above predictions and equations were checked by Monte Carlo simulations. Random mating populations with N diploid individuals were simulated. The selective system was controlled by n loci evenly distributed in linear chromosomes. Further n neutral loci were allocated alternating with the selected loci. The population was initially run for thousands of generations so that the selective system could reach mutation-selection-drift equilibrium. Thereafter, two different sets of runs were carried out according to the objective. In the simulations used to evaluate N_e , alleles from each neutral locus were initially set at frequency 0.5. The population was then simulated for 100–300 generations until the asymp-

TABLE 2
Simulations (and predictions in parentheses) based on multiplicative gene action
with mutants of equal effect, t , on the heterozygote

| $C^2 \times 10^2$ | N | L | N_e/N | π/π_0 | s/s_0 |
|--|-----------|-----|-------------|------------------|------------------|
| Deleterious mutations ($t = 0.05$ when $N = 100$, otherwise $t = 0.02$) | | | | | |
| 4.94 | 100 | 1 | 0.37 (0.41) | 0.42 (0.45) | 0.60 (0.65) |
| 3.51 | 100 | 0 | 0.07 (0.11) | 0.19 (0.19) | 0.45 (0.52) |
| 0.20 ^a | 800 | 0 | — (0.17) | 0.16–0.30 (0.21) | 0.54–0.60 (0.54) |
| 0.20 ^a | 10,000 | 0 | — (0.09) | 0.09 (0.10) | 0.39 (0.41) |
| 0.20 ^a | 1,000,000 | 0 | — (0.08) | 0.08 (0.08) | 0.21 (0.31) |
| Favorable mutations ($t = 0.14$) | | | | | |
| 3.57 | 1,000 | 0 | — (0.007) | 0.023 (0.021) | 0.30 (0.28) |
| 1.68 | 10,000 | 0 | — (0.001) | 0.004 (0.003) | 0.17 (0.24) |

Predictions for the model of deleterious mutations were made using Equations 5, 6, and 10, and those for the model of favorable mutations using Equations 5 and 13, as explained in the text.

^aSimulations taken from Table 1 of Charlesworth *et al.* (1995).

Effective size was clearly reached. Fifty additional generations were run. At least 200 independent replicates of this process were simulated. The variance (Var_i) of the frequency of the neutral genes was computed for each generation i over loci and replicates. The effective population size at a given generation i was computed as $N_{e,i} = 0.5(0.25 - \text{Var}_{i-1}) / (\text{Var}_i - \text{Var}_{i-1})$. The observed asymptotic N_e value was computed as the average of the $N_{e,i}$ values of the 50 additional generations. A different set of simulations was run to evaluate the heterozygosity and the segregation of polymorphic loci. In this case, the neutral genes were introduced as mutants, and the population was run until the equilibrium heterozygosity and polymorphism was reached. The selective value of an individual was calculated as $(1 + t)^k$ for the model of favorable mutations and $(1 - t)^k$ for the model of detrimental mutations, where k is the number of mutants carried by the individual. Every generation the mean fitness of the population was set to 1, and the variance of relative fitnesses of individuals (C^2) was computed.

Table 2 shows some simulations of asymptotic values of N_e , π , and s . Predictions were generally close to simulations. As was explained before, the effective size is progressively reduced over generations until the asymptotic value is reached. A comparison with simulations is made in Figure 3. Predictions of the equilibrium heterozygosity and proportion of segregating sites in Table 2 were made from these values of $N_{e,i}$ in consecutive generations as explained above. As expected, the absolute reduction of N_e is generally greater than the reduction of heterozygosity and polymorphism (cf. Figure 2) because π and s depend not only on the asymptotic N_e but also on nonasymptotic values, particularly s . A tendency of convergence between the ratios π/π_0 and s/s_0 with increasing population size is predicted, as noted by Charlesworth *et al.* (1995) for background selection (see also Figure 2).

Predictions are also accurate when mutations of unequal effects are considered. For example, simulations

from Nordborg *et al.* (1996) with $U = 0.4$, $L = 1$ and mutation effects with mean $t = 0.04$ drawn from a gamma distribution with parameters $\alpha = 0.70$ and $\beta = 0.032$ give an average π/π_0 of 0.67. The prediction from Equation 4a is 0.65, and that from the approximation (4b) (assuming no correlation between C_i^2 and Q_i^2) is 0.67, suggesting that the shape of the distribution of effects is not very important for the effective population size.

Finally, Figure 4 represents an example of the agreement between observed and expected allele frequency spectrums. The expected frequency distribution, under selection for the whole population of mutations originated τ generations ago, was obtained by using transition matrix methods. The partial $N_{e,i}$ values for generations 1 to τ were predicted, and the harmonic mean

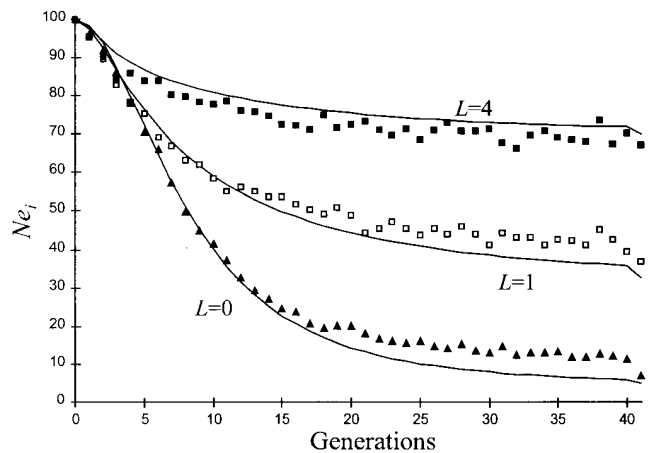


Figure 3.—Three examples of the progressive reduction of $N_{e,i}$ associated with a new neutral mutation in a population of size $N = 100$. Simulated (boxes) and predicted (lines) values for 40 generations. The last element in each series corresponds to the asymptotic N_e value. $C^2 \approx 0.044$ and $t = 0.05$ (deleterious mutations model). Solid boxes, four chromosomes 1 Morgan long each; open boxes, one chromosome 1 Morgan long; and solid triangles, one chromosome with no recombinations.

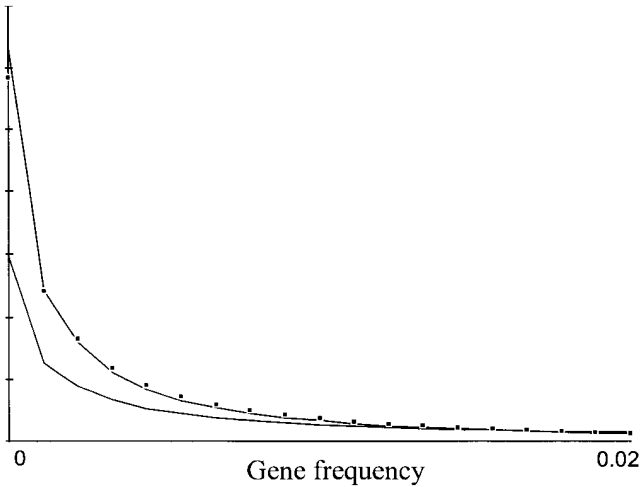


Figure 4.—Example of gene spectrum in a population of 500 individuals under a multiplicative deleterious mutations model and no recombination. $C^2 = 0.00426$, $t = 0.01$. Only gene frequencies between 0 and 0.02 are represented. Simulated values in boxes. Top line: prediction under selection; bottom line, prediction under a pure neutral model (see text for explanations).

$N_{e,H\tau}$ of these was used as the constant effective size of mutations originated τ generations ago. Predictions (top line) were made by accumulating all the expected distributions for neutral mutations originated in all the previous generations and in the current one. Simulations (boxes) were very close to these predictions. The bottom line shows the expected distribution, which would have been predicted under a pure neutral model without selection. This was calculated assuming the constant effective population size, which explains the observed level of heterozygosity in the population.

DISCUSSION

The fundamental concept in our analysis is that the parameter N_e , which summarizes the magnitude of the drift process in a genomic region or in the whole genome, is a function of the rate of reduction of the covariance between the neutral genes and the selected system. This reduction depends on three factors: the genetic size of the genome (*i.e.*, the recombination rate), the change of variance of the selected loci due to selection, and the reduction of variance due to drift. At equilibrium, the total rate of reduction is $V_m/C^2 = 1/2N_e + t$ for models in which this rate is independent of the gene frequencies (*i.e.*, infinitesimal model or deleterious mutations model). When the effects of the selected loci on fitness are large in relation to N_e , say $t \gg 1/2N_e$, the relative influence of genetic drift is small and predictions become independent of N_e . In this case, there is full agreement with equations from Hudson and Kaplan (1995), Barton (1995), and Nordborg *et al.* (1996) for background selection (deleterious muta-

tions). As the effect of the genes decreases, with $t < 1/2N_e$ and getting close to the assumptions of the infinitesimal model, the predictions are more dependent on N_e , and the bigger the population, the smaller the ratio N_e/N (see Table 2).

Our predictions of N_e can be made in terms of compound parameters, such as the variance for fitness, C^2 , and the new input of mutational variance, V_m , but not necessarily on mutation rates and mutational effects of spontaneous mutations, whose magnitudes are in a current debate (*e.g.*, Peck and Eyre-Walker 1997). Houle *et al.* (1996) have reviewed estimates of C^2/V_m for a variety of traits and species, obtaining an average value of 50 for life-history traits. This is an estimate of the average persistence time of detrimental mutations. For viability in *Drosophila* $C^2 = 0.01$, approximately, for the whole genome (Mukai 1988). Because the genome size of *Drosophila melanogaster* is about 1.25 (considering that there is no recombination in males), substituting $V_m/C^2 = 0.02$, $C^2 = 0.01$, and $L = 1.25$ into Equations 8 and 9, we obtain $N_e = 0.67N$, which is a considerable reduction in effective size due to inherited differences in viability alone.

A main requisite for our model to work is the continuous flux of genetic variation for fitness in all the chromosome regions. Mutation introduces new variation at neutral sites while selection reduces the genetic variability. This requirement is far away from the strong selective sweep model assumed by Wiehe and Stephan (1993), for which the hitchhiking of neutral alleles by favorable mutations can be considered as a two-step process. First, a strongly selected gene passes quickly through the population, wiping out linked variation, and second, polymorphism is recovered by mutation in a period where no hitchhiking occurs. Therefore, our equations for favorable mutations are not applicable to the assumptions made by Wiehe and Stephan (1993). For example, with parameters $N = 1000$, $L = 0$, $t = 0.2$, the average simulated heterozygosity (π/π_0) is 0.031, 0.084, and 0.246 when a single selective locus is segregating all the time, one-third of the time, or one-twelfth of the time, respectively. The corresponding predictions obtained with our method are 0.028, 0.033, and 0.041, respectively. Thus, the two latter, including periods of recovery of polymorphism, deviate from the assumptions of our model, and predictions become more and more inaccurate.

Our derivation follows the arguments of Robertson (1961) and Santiago and Caballero (1995). The variance of long-term selective values (Q^2C^2) is partitioned into two components, one due to the chromosome carrying the neutral allele of reference (Q'^2) and the other due to the homologous chromosome (Q''^2). With no linkage, large populations and weak selection, both terms approximate a value of 4 and Equation 3, $N_e = N/(1 + 4C^2)$, is obtained, as deduced by Robertson (1961) and Santiago and Caballero (1995). However,

Barton (1995) and Nordborg *et al.* (1996) arrived at the conclusion that with no linkage $N_e = N/(1 + 2C^2)$. Nordborg *et al.* (1996, Appendix iii) linked this to the previous expression by arguing that in the former the term C^2 is, in fact, $C^2/2$, because it refers to the variance of fitness of families (couples) instead of individuals. This is not correct, however. In the argument of Robertson (1961) and Santiago and Caballero (1995), C^2 is the variance of the relative fitness values of couples because these were fixed (monogamous matings) but this was assumed to be the fitness associated with neutral alleles, and all four alleles (in the couple) had the same associated fitness. The model would be equivalent for monoecious populations (Caballero and Santiago 1995).

The reason for the confusion is clear from the derivation in this article. Barton (1995) and Nordborg *et al.* (1996) considered only the gamete carrying the neutral allele as the determinant for the fitness associated to this allele. This is the same as neglecting Q''^2 , as we did, for example, to obtain Equation 5. Now, for large population size ($Z \approx 1$), $Q' \approx 1/r$, and from Equation 5, $N_e = N/(1 + C^2/2r^2)$, which agrees with Barton's expression. If now $r = 0.5$, the above expression yields $N_e = N/(1 + 2C^2)$. However, neglecting Q''^2 is allowed only for moderate or strong linkage because only for $r \ll 0.5$ is $Q'^2 \gg Q''^2$. The intuitive explanation is that with tight linkage, the fitness associated with the neutral allele depends mostly on that of the gamete carrying it and $Q'^2 \gg Q''^2$. However, for very loose linkage or no linkage, the fitness of the homologous gamete is also important: $Q'^2 \approx Q''^2 \approx 4$, and $N_e = N/(1 + 4C^2)$.

To reduce the complexity of the derivation, we have considered that the recombination rate is constant across the chromosome and the neutral gene is located in the middle of a chromosome. An equivalent derivation can also be developed for a neutral gene at any location. The neutral location does not make a big difference unless the gene is in the final region of the chromosome tip. In this region, the effect of drift is smaller as closely linked selective genes can only (or mainly) be located at one side of the neutral gene, reducing down to a half the random associations with selected genes. As these regions in both tips are very small, their weight on the average N_e value for the whole genome is irrelevant and the result for the central location is a very good approximation to the average N_e . This effect indicates that N_e is mainly determined by the strength of selection acting on the region closely linked to the neutral locus. An equivalent conclusion has been reached by Nordborg *et al.* (1996) under the background selection model.

Regional variations in the frequency of recombination are often observed (see Lichten and Goldman 1995), with a general pattern of reduced recombination in proximal regions (*e.g.*, Nachman and Churchill 1996). Additionally, the distribution of transcriptional

genes throughout the genome does not seem to be uniform (Gardiner 1996), suggesting that the source of genetic variability for fitness is not evenly distributed in the genome. The exact magnitude of these deviations is unknown, but the former equations could also be applied if the density of genetic variability for fitness were more or less proportional to the rate of recombination. Otherwise, the computation of the appropriate value of Q'^2 must consider the particular distribution of selected genes and genetic distances between these genes and the neutral locus.

The reduction of effective size associated with a neutral gene is progressive: The magnitude of the drift process is smaller for new neutral mutations than for old ones, and this process accumulates on neutral genes over generations until an asymptotic value is reached. The consequence is that heterozygosity will always be larger than that expected if all the neutral genes in the population had a constant effective size equal to the asymptotic value (N_e) and, therefore, cannot be formally predicted in the simple way, $4N_e\mu$. The magnitude of the underprediction depends on how quickly the asymptotic N_e value is reached (see Figure 3). For a given genome size or recombination rate, this relies on the rate of reduction of the genetic variance. Under the assumptions of the infinitesimal model, the reduction of the variance in the selected system will be mainly due to drift if selection is weak, the reduction of effective size will be slow, and the difference between the real heterozygosity and that expected from the asymptotic N_e will be the highest. As the effect t of selected genes becomes larger, the rate of reduction increases and the asymptotic N_e is reached earlier. In a model of deleterious mutations of large effect (the background selection model), heterozygosity tends to be almost equal to $4N_e\mu$ as mutations reach their asymptotic value of N_e in a few generations. Nordborg *et al.* (1996) developed a prediction of the reduction of heterozygosity due to linked selected loci under background selection (formally π/π_0), which is identical to our prediction of the asymptotic N_e/N when drift is not considered. This prediction turns inexact as the population size or the effect of selected genes decreases (Nordborg *et al.* 1996), because the associated effective sizes of neutral alleles become further and further away from the asymptotic value. An issue related to those above refers to the reductions in the probability of fixation of advantageous mutations because of their linkage to other selected loci (see Barton 1994). Mutants of very large effect, whose fate is decided in a few generations, are affected by the asymptotic N_e less than mutants of small effect and, therefore, the fixation probability of the former is reduced by a smaller amount (Caballero and Santiago 1998).

The progressive reduction of the effective size associated with mutations can also explain the apparent disconnection between heterozygosity and number of segregating sites under selection, which is the basis of

statistical tests of neutrality (*e.g.*, Fu 1996). Actually, those tests compare the observed spectrum of gene frequencies with its expectation under a pure neutral model. As we have seen, the proportion of segregating sites is little dependent on N_e , as it is mostly due to recent mutations, which have associated effective sizes close to the census size of the population and far away from the asymptotic value. For very large populations, however, the number of generations that effectively contribute to the proportions of segregating sites is larger. Therefore, the drop of effective size during the initial generations affects it more, making heterozygosity and number of segregating sites similarly reduced. This effect has been described by Charlesworth *et al.* (1995).

The spectrum of gene frequencies can be approximated from the evolution of N_e associated to mutations over generations. For mutations originated τ generations before the actual generation, the magnitude of the drift process can be summarized by the harmonic mean ($N_{e,H\tau}$) of the $N_{e,i}$ values from generation $i = 1$ to τ . The remaining proportion of heterozygosity can be predicted by $(1 - 1/2N_{e,H\tau})^\tau$. Analogously, the proportion of segregating sites can be approximated from $N_{e,H\tau}$ using the general theory of the effective population size (Equations 17–18). In other words, the spectrum of gene frequencies for mutations originated τ generations ago is approximately the expected under a neutral model using the appropriate $N_{e,H\tau}$. Deviations from the pure neutral spectrum arise when the contributions of all previous generations are accumulated. Different spectra corresponding to different $N_{e,H\tau}$ values of previous generations (from $\tau = 1$ to ∞) are superimposed, one over the others, building a general spectrum that cannot be explained by a single N_e value under a neutral model (see Figure 4).

When statistical tests are applied to compare predicted and observed spectra of gene frequencies, the finite size of the samples can make the deviations from the neutral model difficult to detect. Observations in natural populations of *Drosophila* denote reduced diversity in regions with low recombination rates (Begun and Aquadro 1992), but most data show no deviations from the neutral spectrum (see Charlesworth *et al.* 1995). Although virtually any model considering directional selection could account for the observed correlation between nucleotide variation and recombination rate, simple selective sweep models with strong selection cannot explain the statistical agreement with the neutral spectrum (Hudson 1994; Braverman *et al.* 1995). Predictions using computer simulations reveal that the statistical agreement is consistent with the background selection model (Charlesworth *et al.* 1995; Hamblin and Aquadro 1996). The general theory that we have described can help to determine the conditions for the background selection model, alone or combined with weak selective sweep models, which could explain the pattern of observed variation.

Finally, some remarks concerning artificial selection can be made. In the general theory of quantitative traits, linkage is usually ignored as farm species generally have several chromosomes, suggesting that the assumption of free recombination is close to reality. Additionally, linkage makes the analytical model more cumbersome: Additive models are complicated by the effect of the generation of negative covariances between genes affecting fitness (Bulmer 1980; Santiago 1998). Although our theory takes into account this effect, which is included the term $Z = 1 - V_m/C^2$, its application to a model in which parents are selected individuals and the genetic values of both “gametes” are negatively correlated is not straight. Further insight into these models is necessary to assess the impact of linkage in artificial selection programs.

We thank B. Charlesworth, W. G. Hill, and N. Barton for helpful comments. This work was supported by grant PB95-0909-C02-02 from Ministerio de Educación y Cultura (Spain) to E.S. and by grant 64102C605 from Universidad de Vigo to A.C.

LITERATURE CITED

- Barton, N. H., 1994 The reduction in fixation probability caused by substitutions at linked loci. *Genet. Res.* **64**: 199–208.
- Barton, N. H., 1995 Linkage and the limits to natural selection. *Genetics* **140**: 821–841.
- Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.
- Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140**: 783–796.
- Bulmer, M. G., 1980 *The Mathematical Theory of Quantitative Genetics*. Clarendon Press, Oxford.
- Bürger, R. E., G. P. Wagner and F. Stettinger, 1988 How much heritable variation can be maintained in finite populations by a mutation selection balance? *Evolution* **43**: 1748–1766.
- Caballero, A., 1994 Developments in the prediction of the effective population size. *Heredity* **73**: 657–679.
- Caballero, A., and E. Santiago, 1995 Response to selection from new mutation and effective size of partially inbred populations. I. Theoretical results. *Genet. Res.* **66**: 213–225.
- Caballero, A., and E. Santiago, 1998 Survival rates of major genes in selection programmes. *Proc. 6th World Congress Genet. Appl. Livestock Production* **26**: 5–12.
- Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.
- Charlesworth, D., B. Charlesworth and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. *Genetics* **141**: 1619–1632.
- Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.
- Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, Berlin.
- Fu, Y.-X., 1996 New statistical tests of neutrality for DNA samples from a population. *Genetics* **143**: 557–570.
- Gale, J. S., 1990 *Theoretical Population Genetics*. Unwin Hyman, London.
- Gardiner, K., 1996 Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet.* **12**: 519–524.
- Haigh, J., 1978 The accumulation of deleterious genes in a population—Muller’s Ratchet. *Theoret. Pop. Biol.* **14**: 251–267.
- Hal dane, J. B. S., 1919 The combination of linkage values and the calculation of the distances between the loci of linked factors. *J. Genet.* **8**: 299–309.

- Hamblin, M. T., and C. F. Aquadro, 1996 High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. *Mol. Biol. Evol.* **13**: 1133–1140.
- Houle, D., B. Morikawa and M. Lynch, 1996 Comparing mutational variabilities. *Genetics*. **143**: 1467–1483.
- Hudson, R. R., 1994 How can the low levels of DNA sequence variation in regions of the *Drosophila* genome with low recombination rates be explained? *Proc. Natl. Acad. Sci. USA* **91**: 6815–6818.
- Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. *Genetics* **141**: 1605–1617.
- Keightley, P. D., and W. G. Hill, 1988 Quantitative genetic variability maintained by mutation-stabilizing selection balance in finite populations. *Genet. Res.* **52**: 33–43.
- Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.
- Kimura, M., and Maruyama, T., 1966 The mutational load with epistatic gene interactions in fitness. *Genetics* **54**: 1337–1351.
- Lichten, M., and A. S. Goldman, 1995 Meiotic recombination hotspots. *Annu. Rev. Genet.* **29**: 423–444.
- Lynch, M., and W. G. Hill, 1986 Phenotypic evolution by neutral mutation. *Evolution* **40**: 915–935.
- Maynard-Smith, J., and J. Haigh, 1974 The hitchhiking effect of a favourable gene. *Genet. Res.* **23**: 23–35.
- Mukai, T., 1988 Genotype-environment interaction in relation to the maintenance of genetic variability in populations of *Drosophila melanogaster*, pp. 21–31 in *Proceedings of the 2nd International Congress on Quantitative Genetics*, edited by B. S. Weir, E. J. Eisen, M. J. Goodman and G. Namkoong. Sinauer, Sunderland, MA.
- Nachman, M. W., and G. A. Churchill, 1996 Heterogeneity in rates of recombination across the mouse genome. *Genetics* **142**: 537–548.
- Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 The effect of recombination on background selection. *Genet. Res.* **67**: 159–174.
- Peck, J. R., and A. Eyre-Walker, 1997 The muddle about mutations. *Nature* **387**: 135–136.
- Robertson, A., 1961 Inbreeding in artificial selection programmes. *Genet. Res.* **2**: 189–194.
- Santiago, E., 1998 Linkage and the maintenance of variation by mutation-selection balance in finite populations: an infinitesimal model. *Genet. Res.* **71**: 161–170.
- Santiago, E., and A. Caballero, 1995 Effective size of populations under selection. *Genetics* **139**: 1013–1030.
- Wiehe, T. H. E., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* **10**: 842–854.
- Woolliams, J. A., N. R. Wray and R. Thompson, 1993 Prediction of long-term contributions and inbreeding in populations undergoing mass selection. *Genet. Res.* **62**: 231–242.
- Wright, S., 1931 Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Communicating editor: B. S. Weir

