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

Enrique Santiago* and Armando Caballero²

*Departamento Biología Funcional, Universidad de Oviedo, 33071 Oviedo, Spain and ²Departamento 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

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

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 descendants 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
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, \( Q = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2 \) (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 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 chromosomes with \( n \) chromosomes \( M \) 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), \( r = \frac{1 - \exp(-2x)}{2} \), 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-

### Table 1

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>Number of reproductive individuals</td>
</tr>
<tr>
<td>( N_a )</td>
<td>Asymptotic effective population size (with subscript ( i ) refers to generation ( i ))</td>
</tr>
<tr>
<td>( N_{a,H} )</td>
<td>Harmonic mean of the effective size between generations 1 and ( \tau )</td>
</tr>
<tr>
<td>( l )</td>
<td>Chromosome length Morgans</td>
</tr>
<tr>
<td>( \nu )</td>
<td>Number of chromosomes in the genome</td>
</tr>
<tr>
<td>( L )</td>
<td>Total length of the genome in Morgans (( = l \nu ))</td>
</tr>
<tr>
<td>( r )</td>
<td>Recombination fraction</td>
</tr>
<tr>
<td>( x )</td>
<td>Genetic distance in Morgans</td>
</tr>
<tr>
<td>( t )</td>
<td>Heterozygous effect of selected gene</td>
</tr>
<tr>
<td>( \mu )</td>
<td>Mutation rate per locus, gamete and generation</td>
</tr>
<tr>
<td>( U )</td>
<td>Total genomic (diploid) mutation rate</td>
</tr>
<tr>
<td>( \pi )</td>
<td>Equilibrium heterozygosity (with subscript 0 refers to no selection)</td>
</tr>
<tr>
<td>( s )</td>
<td>Equilibrium proportion of segregating sites (subscript 0 refers to no selection)</td>
</tr>
<tr>
<td>( Z )</td>
<td>Reduction in genetic variance per generation due to selection and drift</td>
</tr>
<tr>
<td>( Z_i )</td>
<td>Reduction in genetic variance between generations 1 and ( i )</td>
</tr>
<tr>
<td>( C^2 )</td>
<td>Variance of relative fitnesses of individuals</td>
</tr>
<tr>
<td>( C^2 )</td>
<td>Contribution of selected locus to the genetic variance for fitness</td>
</tr>
<tr>
<td>( V_m )</td>
<td>Mutational input of variance per generation for fitness</td>
</tr>
<tr>
<td>( Q' )</td>
<td>Long term selective values of the chromosome initially carrying the neutral allele</td>
</tr>
<tr>
<td>( Q'' )</td>
<td>Long term selective values of the homologous chromosome</td>
</tr>
</tbody>
</table>
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 $C^2 = \Pi_{i=1}^{\infty} (1 + c_i^2) - 1$, where $c_i^2$ is the square of the coefficient of variation contributed by locus $i$, 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 $x$ M 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 $S_j$ the selective value contributed by the homologous selected allele in the other gamete. Under random mating, the expected change in gene frequency ($S_j$) of the copy of the neutral allele in the first generation ($S_j$) can be partitioned into the change due to the selected allele in the same gamete as the neutral gene ($S_j'$) and the change due to the homologous selected allele in the other gamete in the same individual ($S_j''$),

$$S_j = \text{cov}(p, f_j' + f_j'') = \text{cov}(p, f_j') + \text{cov}(p, f_j'') = S_j' + S_j'' .$$

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 ($S_j'$ and $S_j''$) in an identical way: Every generation, the genetic variance of the selected locus is assumed to be reduced by selection and drift by a constant proportion $(1 - r)$. 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

$$S_j' = \text{cov}_e(p, f_j') = \text{cov}_e(p, f_j') (1 + e^{-2x}) / 2 Z .$$

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

$$S_j'' = \text{cov}_e(p, f_j'') = S_j'' r Z .$$

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.,

$$S_j' = S_j'(1 - r)Z = S_j'(1 - r)^2 Z^2 ,$$
$$S_j'' = S_j''(1 - r)Z = S_j'' r(1 - r)Z^2 ;$$
$$S_j' = S_j'(1 - r)Z = S_j'(1 - r)^2 Z^3 ,$$
$$S_j'' = S_j''(1 - r)Z = S_j'' r(1 - r)^2 Z^4 ,$$

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,

$$Q_j' = \frac{1}{S_j' \times 1} \sum_{i=1}^{S_j'} = \frac{2}{2 - Z - Ze^{-2x}} , \quad (1a)$$
$$Q_j'' = \frac{1}{S_j'' \times 1} \sum_{i=1}^{S_j''} = \frac{2(1 - Ze^{-2x})}{2 - Z - Ze^{-2x}} . \quad (1b)$$

[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 pop- ulation 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,

\[
Q_j^2 c_j^2 = Q_j^2 c_j^2 \frac{Q_j^2}{2} + \frac{Q_j^2 + Q_j^2}{2} = \frac{Q_j^2 + Q_j^2}{2} c_j^2.
\]

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

\[
N_e(\text{due to } j) = \frac{N}{1 + c_j^2((Q_j^2 + Q_j^2)/2)}.
\]

From Equations 1a and 1b we note that \( Q_j^2 \) and \( Q_j^2 \) take the value \( 2(2 - Z) = 2 \) 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., \( Z = 1 \)). Therefore, with no linkage, \( Q_j^2 = Q_j^2 = 4 \), and Equation 2 yields

\[
N_e = \frac{N}{1 + 4C} \approx \frac{N}{1 + 4C}.
\]

(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 selected loci in the genome is

\[
\Pi_{j=1}^{n} \left(1 + c_j^2 \frac{Q_j^2 + Q_j^2}{2}\right) - 1.
\]

Therefore, the asymptotic value of \( N_e \) is

\[
N_e = \frac{N}{1 + \Pi_{j=1}^{n}((1 + c_j^2((Q_j^2 + Q_j^2)/2)) - 1
= N \exp\left(-\sum_{j=1}^{n} c_j(Q_j^2 + Q_j^2)/2\right).
\]

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

\[
N_e = N \exp\left(-C^2 Q_j^2 + Q_j^2/2\right),
\]

where \( Q_j^2 = \sum_{j=1}^{n} Q_j^2/n \) and \( Q_j^2 = \sum_{j=1}^{n} Q_j^2/n \).

For large populations \( Q_j^2 \) 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 \( Q_j^2 \) ranges from 1 (complete linkage) to 4 (no linkage). \( Q_j^2 \) approxi- mates \( 1/r^2 \) (Equation 1a) in large populations under weak selection (i.e., \( Z = 1 \)), so it may take values much larger than 4 for tight linkage (\( r \ll 1/2 \)). Thus, \( Q_j^2 \) can usually be neglected relative to \( Q_j^2 \), and Equation 4b can be reduced to

\[
N_e = N \exp\left(-C^2 Q_j^2/2\right),
\]

without losing much precision. The term \( Q_j^2 \), 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 \( 1 - 1/\nu \)). 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,

\[
Q_j^2 = \left(\frac{\nu}{2 - Z}\right)^2 + \frac{2}{\nu} \int_0^{1/\nu} \left(\frac{2}{2 - Z - Ze^{-x}}\right)^2 dx
= \frac{1}{\nu^2} \int_0^{1/\nu} \left(\frac{2}{2 - Z - Ze^{-x}}\right)^2 dx
\]

\[
+ \frac{4}{\nu} \left(\frac{1}{(2 - Z)^2} - \frac{1}{(2 - Z)(2 - 2Z)} - \frac{\ln(2 - 2Z)}{(2 - Z)^2}\right).
\]

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 \( L \):

\[
Q_j^2 = \frac{1}{L} \left(\frac{L}{(2 - Z)}\right)^2 - \frac{1}{(2 - Z)(2 - 2Z)} - \frac{\ln(2 - 2Z)}{(2 - Z)^2}\right).
\]

The following approximations to the above expression can be made:

if \( L \to \infty \) (no linkage), then \( Q_j^2 = Q_j^2 = 4 \),
and Equation 4b should be used,
if \( L \to 0 \) (complete linkage), then \( Q_j^2 \to 1/(1 - Z)^2 \)
and \( Q_j^2 \) can be neglected,
if \( L \gg 0 \) (say \( L > 0.2 \)), then \( Q'^2 \approx 2/ (1 - Z)L \) 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

\[
N_e = N \exp \left( -\frac{C^2}{(1 - Z)L} \right) .
\] (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, \( N_e = N / [1 + C^2 / (1 - Z)L] \).

Application to particular genetic systems: The above equations for predicting the effective population size are a function of the proportional reduction of the genetic variance \((1 - Z)\) 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 mutation rate for detrimental genes. Combining both processes for the effective population size, \( N_e \), expressed in a way more familiar to the classical equations, 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 \( V_m = U t^2 \), 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

\[
Z = (1 - t) \left( 1 - \frac{1}{2N_e} \right) = 1 - t - \frac{1}{2N_e} .
\] (10)

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

\[
Q' = \frac{1}{r + (1 - r) / 2N_e + t} .
\] (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 \( Q'^2 = 2 / [L (t + 1 / 2N_e) \), and Equation 8 becomes

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

When the effect of drift is negligible (i.e., \( t \gg 1 / 2N_e \)), then \( C^2 = Ut \) and \( N_e = N \exp (-U / L) = N / [1 + (U / L)] \), 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, \( N_e = N \exp (-U / (2t)) \). 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 \( \pi_0 \) 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 \( q_2 = q_1 + q_1(1 - q_1) t \) and the proportional change in variance due to selection is \( q_1(1 - q_1)/q_1(1 - q_1) \). Drift also reduces the genetic variance by \( 1 - 1/2N_e \), therefore, \( Z_i = (1 - 1/2N_e) q_1(1 - q_1)/q_1(1 - q_1) \). Here we consider only the effect of the gamete associated with the neutral gene (i.e., we neglect \( Q_j^{z} \)), obtaining \( Q_j^{z} \) with the same argument leading to Equation 1a. If the frequency of recombination between the neutral and the selected locus \( j \) is \( r \) and the covariance between them in the first generation is \( S_j \), the expected changes in gene frequency in the following generations are

\[
S_j = S_j(1 - r)Z_1, \\
S_j = S_j(1 - r)^2 \left( 1 - \frac{1}{2N_e} q_1(1 - q_1) \right), \\
S_j = S_j(1 - r)^3 Z_2, \\
\]

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

\[
Q_j^{z}(q_1) = \frac{1}{S_j} \sum_{i=0}^{\infty} S_j^i = \sum_{i=0}^{\infty} Z \left( 1 + e^{-2^i} \right),
\]

where \( i \) represents the successive generations and \( q \) are the sequential frequencies of the selected allele in the consecutive generations, which are obtained as \( q = q_{-1} + q_{-1}(1 - q_{-1}) t \). 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 \( Q_j^{z} \) is the mean weighted value of the \( Q_j^{z}(q_1) \) 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, \( q_1 \), 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, \( Q_j^{z} \) can be approximated as the average value of a number \( m \) of \( Q_j^{z}(q_1) \) values corresponding to initial frequencies \( q_1 \) evenly spaced through the spectrum of gene frequencies from 0 to 1,

\[
Q_j^{z} = \frac{\sum_{i=0}^{\infty} Z \left( 1 + e^{-2^i} \right)}{m},
\]

The appropriate \( Q_j^{z} \) value of the neutral locus is the average value of the \( Q_j^{z} \) 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, HETEROZYGOITY, 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,

\[
N_{e1} = \frac{N}{1 + C^{2}((Q_j^{z} + Q_j^{z'})/2)},
\]

\[
= N \exp \left( -C^{2}Q_j^{z} + Q_j^{z'} \right),
\]

where \( Q_j^{z} \) can be computed as the average value for all the \( Q_j^{z} \) values of selected loci, being obviously equal to one, \( Q_j^{z} = (1/S_j) \sum_{i=1}^{\infty} S_j^i = 1 \), so that \( Q_j^{z} = 1 \). The value of \( Q_j^{z'} \) is also 1. As stated before, the asymptotic value of \( Q_j^{z} \) is less than or equal to 2, and after a few generations it will be much smaller than \( Q_j^{z} \) under linkage, so
Effective Population Size with Linkage

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 \( N_e \) 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 \( N_{e1} \). This is calculated using \( Q_{e1} \), the average of all the \( Q_{e} \)'s (see the accumulation of terms stated before), obtained as \( Q_{e1} = (1/S) \sum i=1 \infty Z^i = 1 + (1 - r)Z + (1 - r)^2Z^2 \), and so on up to infinite, \( Q_{e\infty} = Q_{e} \), when the asymptotic effective population size, \( N_{e\infty} = N_e \) (equations in the previous sections), is reached.

The 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 \( \tau \) 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 effective size associated with a mutation is larger than which we refer to as \( N_e \) 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 effective size \( N_e \) is much smaller than the asymptotic effective size, \( N_{e\infty} \) 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 \( N_{e\infty} \). 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.

Heterozygosity: Under the infinite sites model, the heterozygosity contributed by a new mutation (i.e., with frequency \( 1/2N_e \)) is \( (1/2N_e)(1 - 1/2N_e) = 1/2N_e \). Then, with a mutation rate \( \mu \) per locus and generation, the number of new mutations per generation is \( 2N \mu \), and the input of heterozygosity per generation is about \( 2N \mu \). The neutral variability generated by these mutations decreases at an increasing rate, which is a function of the consecutive values of \( N_{e\infty} \), so the remaining proportion after \( \tau \) generations is \( R_{\tau} = 11_{\infty}(1 - 1/2N_{e\infty}) \). Therefore, the expected heterozygosity at equilibrium (\( \pi \)) is the sum of the contributions by mutations during all the previous generations,

\[
\pi = 2\mu(1 + \sum_{\tau=1}^{\infty} R_{\tau}). \tag{16}
\]

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 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 \( N_{e\infty} \). 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 \( \tau \) 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 effective size associated with a mutation is larger than which we refer to as \( N_e \) 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 effective size \( N_e \) is much smaller than the asymptotic effective size, \( N_{e\infty} \) 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 \( N_{e\infty} \). 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.
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 \( 4N\mu \). In \( 2N \) (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 \( N_{ei} \) values of the initial generations have much influence. Because these initial \( N_{ei} \) 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 (1/2\( 2N \) 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 \( s, \pi, \) and \( N_{ei} \) 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 \( N_{ei} \) and \( \pi \) 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 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 \( N_{eh} \), of the partial \( N_{ei} \) values for the \( t \) generations is used as the constant effective population size for the \( t \) 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 \( N_{ei} \) alleles from each neutral locus were initially set at frequency 0.5. The population was then simulated for 100–300 generations until the asymp-
Effective Population Size with Linkage

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

<table>
<thead>
<tr>
<th>C² x 10²</th>
<th>N</th>
<th>L</th>
<th>N/N</th>
<th>t/α₀</th>
<th>s/s₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.94</td>
<td>100</td>
<td>1</td>
<td>0.37(0.41)</td>
<td>0.42(0.45)</td>
<td>0.60(0.65)</td>
</tr>
<tr>
<td>3.51</td>
<td>100</td>
<td>0</td>
<td>0.07(0.11)</td>
<td>0.19(0.19)</td>
<td>0.45(0.52)</td>
</tr>
<tr>
<td>0.20</td>
<td>800</td>
<td>0</td>
<td>— (0.17)</td>
<td>0.16-0.30(0.21)</td>
<td>0.54-0.60(0.54)</td>
</tr>
<tr>
<td>0.20</td>
<td>10,000</td>
<td>0</td>
<td>— (0.09)</td>
<td>0.09(0.10)</td>
<td>0.39(0.41)</td>
</tr>
<tr>
<td>0.20</td>
<td>1,000,000</td>
<td>0</td>
<td>— (0.08)</td>
<td>0.08(0.08)</td>
<td>0.21(0.31)</td>
</tr>
</tbody>
</table>

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.

Simulations taken from Table 1 of Charlesworth et al. (1995).

Note: The effective size was clearly reached. Fifty additional generations were run. At least 200 independent replicates of this process were simulated. The variance (Var.) 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 = 0.5(Var._i) / (Var._i + Var._z). The observed asymptotic N_e value was computed as the average of the N_e 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 (C²_t) was computed.

Table 2 shows some simulations of asymptotic values of N_e, t, 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 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 t and s depend not only on the asymptotic N_e but also on nonasymptotic values, particularly s. A tendency of convergence between the ratios N_e / N_α and s/s₀ 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 α = 0.70 and β = 0.032 give an average N_e / N_α of 0.67. The prediction from Equation 4a is 0.65, and that from the approximation (4b) (assuming no correlation between C²_t and Q_t²) 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 spectra. The expected frequency distribution, under selection for the whole population of mutations originated N generations ago, was obtained by using transition matrix methods. The partial N_e values for generations 1 to N were predicted, and the harmonic mean
and predictions become independent of Kaplan and Hudson there is full agreement with equations from N parameter of the gene frequencies (i.e. N selected loci on fitness are large in relation to for explanations). Considerable reduction in effective size due to inherited 500 individuals under a multiplicative deleterious mutations nome size of 2114 E. Santiago and A. Caballero for explanations). At equilibrium, the total rate of reduction is rate), the change of variance of the selected loci due to drift, the recombination recovered 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^-) 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^-), is obtained, as deduced by Robertson (1961) and Santiago and Caballero (1995). However,
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 $Q$ 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, $4N_e u$. 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 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 $4N_e u$, 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 $n \pi_0 / n_0$), 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 reduction 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 \( t \) generations before the actual generation, the magnitude of the drift process can be summarized by the harmonic mean \( (N_{eht})_i \) of the \( N_{eht} \) values from generation \( i = 1 \) to \( t \). The remaining proportion of heterozygosity can be predicted by \( (1 - 1/2N_{eht})^t \). Analogously, the proportion of segregating sites can be approximated from \( N_{eht} \) using the general theory of the effective population size (Equations 17–18). In other words, the spectrum of gene frequencies for mutations originated \( t \) generations ago is approximately the expected under a neutral model using the appropriate \( N_{eht} \). Deviations from the pure neutral spectrum arise when the contributions of all previous generations are accumulated. Different spectra corresponding to different \( N_{eht} \) values of previous generations (from \( t = 1 \) to \( \infty \)) 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 \( \mathcal{Z} = 1 - V_m/C^2 \), 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 641026/05 from Universidad de Vigo to A.C.

**LITERATURE CITED**


Fu, Y.-X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143**: 557–570.


Haldane, 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.


Communicating editor: B. S. Weir