Theoretical and Applied Genetics 48, 263-268 (1976)

The Effect of Selection on the Standardized Variance of Gene Frequency

F.W. Nicholas and A. Robertson
Institute of Animal Genetics, Edinburgh (Scotland)

Summary. The effect of directional and heterotic selection on the standardized variance of gene frequency ($f = \sigma_q^2/\bar{q}(1-\bar{q})$) has been examined. It has been found that heterotic selection always results in f values lower than those expected due to drift alone. Additive directional selection can result in low f values, but values larger than those expected due to drift will be observed under additive selection with low initial gene frequency, or when the populations have been separated for a very long period of time in which case f expected due to drift is quite high (around 0.7 or greater). The effect of selection on f is unlikely to be detected if the observed value of f is less than 0.1.

Introduction

The relative importance of selection and random drift in determining the observed pattern of evolution is still a major topic of debate in population genetics. For loci at which gene frequency can be determined, one of the lines of study currently being followed is based on an idea apparently first suggested by Cavalli-Sforza (1966), in which the standardized variance of gene frequency of $f = \sigma_q^2/\bar{q}(1-\bar{q})$ is estimated for various loci over several populations. Thus f is estimated from the mean gene frequency at a particular locus over several populations ($\bar{q}$), and the variance of the gene frequency distribution (σ_q^2) over the same populations, at a particular point in time. Since all loci in a given group of populations have been subjected to exactly the same breeding structure, f values obtained from any number of such loci will be homogeneous unless selection has been acting at some of the loci. Lewontin and Krakauer (1973) have recently developed various statistical tests for the homogeneity of f values, though these can be shown to be invalid if the populations within a species have a hierarchical structure (Robertson 1975).

Many of these authors have drawn some conclusions as to the type of selection which is acting. Thus it has been argued that f values lower than those expected due to drift alone could be due to some form of stabilizing selection (e.g. heterozygote superiority), and relatively large f values may be indicative of different strengths of directional selection at the same locus in different populations. But these generalisations are the only knowledge currently available: what is lacking is a proper understanding of the way various models of selection affect the standardized variance of gene frequency.

It is not only natural populations which are being subjected to this type of study. The advent of suitable electrophoretic techniques has recently led to studies of the effect of selection on gene frequency and the variance of gene frequency at polymorphic loci in laboratory populations of mice (Garnett and Falconer 1975). The latter study was concerned solely with the effect of artificial selection for a metric character on gene frequency and variance of gene frequency at various 'electrophoretic' and coat colour loci. A better knowledge of the way in which selection affects the δ standardized variance of gene frequency would assist in the interpretation of such artificial selection experiments.

In attempting to trace the history of human evolution, Cavalli-Sforza (1969) developed an algebraic relationship between f and the time (t) since separation of two populations, for a model of constant but different directional selective values in different populations at the same locus and compared it to the relationship $f = 1-e^{-t/2N}$ expected in the absence of selection. These two relationships were then used to obtain lower and upper limits respectively of t, the time since divergence. But other models of selection would give completely different relationships between f and t and hence completely different estimates of time since divergence. Once again, therefore, a greater understanding of the effect of selection on f would be useful.

For human populations, Cavalli-Sforza and Zei (1967) and Bodmer and Cavalli-Sforza (1968) have obtained the expected value of f for more complex but more realistic models using the Monte-Carlo and migration matrix methods respectively, on a computer. Expected values of f so obtained for situations where sufficient migration and general demographic data are available have been compared with observed f values. But it is difficult to use these methods to determine the effect of selection on f in general terms, as so many parameters of migration and/or demography are required to obtain any specific answer. The cost in computer time is also quite substantial.

As will be shown below, it is impossible to obtain any useful results from algebra alone. An understanding can only be acquired by the use of a transition probability matrix with which it is possible to calculate the expected value of σ_q^2 and q and hence f at any time under various models of selection.

The aim of this study is to obtain a greater insight into the behaviour of the standardized variance of gene frequency under simple models of selection.

The Additive Model

Consider a single locus with alleles A_1 and A_2 and relative fitnesses of the genotypes A_2A_2, A_1A_2 and A_1A_1 1-s/2, 1 and 1+s/2 respectively. If, in a single population, the genotypes are in Hardy-Weinberg equilibrium, then the frequency after selection q_1, is given by

$$q_1 = \bar{q} + sq(1 - q)/2$$

In a finite population, the frequency will also vary because of genetic sampling. We can then only usefully discuss the behaviour of a set of replicate lines with the same effective number of parents N, and selection pressures. If the variance of q over populations is V, then the standardized variance, f, is given by

$f = V/\bar{q}(1 - \bar{q})$ where $\bar{q}$ is the mean frequency.

With no selection, the expected value of f will increase by an amount $(1-f)/2N$ each generation. But selection too will cause a change in f. Putting $q = \bar{q} + \delta q$ and using subscripts for the values after selection, we have

$$\begin{aligned} q_1 &= \bar{q} + s\bar{q}(1-\bar{q})/2 + \delta q(1+s(1-2\bar{q})/2) - s(\delta q)^2/2 \\ &= \bar{q} + s(\bar{q}(1-\bar{q})-V)/2 + \delta q(1+s(1-2\bar{q})/2) - s((\delta q)^2-V)/2 \\ &= \bar{q}_1 + \delta q(1+s(1-2\bar{q})/2) - s((\delta q)^2-V)/2 \,. \end{aligned}$$

Thus the variance after selection, V_1, is given by

$$\begin{aligned} V_1 &= E((\delta q)^2(1+s(1-2\bar{q})) - s(\delta q)^3 + \text{terms in } s^2) \\ &= V(1+s(1-2\bar{q})) - s\mu_3, \text{ where } \mu_3 \text{ is the third moment about the mean.} \end{aligned}$$

Thus the effect of selection on the variance is dependent on the third moment - see Crow and Kimura (1970), p. 239.

The matrix operations

The derivation and subsequent use of a suitable matrix have been described in full, for example, by Hill and Robertson (1968). Only a brief description, therefore, of the matrix operations will be given here.

Consider a population of N diploid individuals mating at random (including selfing). At a particular single locus with two alleles A_1 and A_2, the genotypes A_2A_2, A_1A_2 and A_1A_1 are assumed to have Hardy-Weinberg frequencies of $(1-q)^2$, $2q(1-q)$ and q^2 respectively at conception, where q is the frequency of allele A_1 at conception. The relative fitnesses of these three genotypes are assumed to be S_{22}, 1 and S_{11} respectively.

For a given gene frequency i/2N, the proportion g_i of each genotype in the population of parents at the time of their mating is

$$g_{i22} = \frac{1}{\bar{w}}(1-q)^2 S_{22}$$

$$g_{i12} = \frac{1}{\bar{w}} 2q(1-q)$$

$$g_{i11} = \frac{1}{\bar{w}} q^2 S_{11}$$

where $q = 1/2N$ and $\bar{w}$ is the proportion of zygotes which remain to be included as parents, and is given by

$$\bar{w} = (1-q)^2 S_{22} + 2q(1-q) + q^2 S_{11} \,.$$

The probability of obtaining exactly x A_2A_2, y A_1A_2 and z A_1A_1 genotypes (x+y+z=N) in a popu-

lation of N survivors, given that there were i A_1 alleles in the population of zygotes in the same generation can be expressed as

$$f_i(x,y,z) = \binom{N}{x\,y\,z} g_{i22}^x \, g_{i12}^y \, g_{i11}^z \ ,$$

and can easily be evaluated on a computer for all $i = 0, 1, \ldots, 2N$. It then follows that the probability p_{ij} of obtaining j A_1 alleles in a population of N zygotes at generation $t+1$, given that there were i A_1 alleles in the N zygotes at generation t is

$$p_{ij} = \sum_{2z+y=j} f_i(x,y,z) \qquad i,j = 0, 1, \ldots, 2N \ ,$$

which is an element of the transition probability matrix **P**. The matrix is square of dimension $2N+1$, and within each row $\sum_{j=0}^{2N} p_{ij} = 1$.

The expected value of q and σ_q^2 can then be obtained by post multiplication of **P** by column vectors representing the first and second moments about zero of the distribution of gene frequency. Thus the selection process is commenced by setting up a column vector $\mathbf{u}_0$ with elements $u_i = 1/2N$ and a second vector v with elements $v_i = i/2N \times i/2N$. Then the matrix operations

$$\mathbf{u}_1 = \mathbf{P}\,\mathbf{u}_0$$

and

$$\mathbf{v}_1 = \mathbf{P}\,\mathbf{v}_0$$

result in vectors $\mathbf{u}_1$ and $\mathbf{v}_1$ representing the first and second moments after one generation of selection. The results for subsequent generations are then obtained as

$$\mathbf{u}_2 = \mathbf{P}\,\mathbf{u}_1$$

and

$$\mathbf{u}_t = \mathbf{P}\,\mathbf{u}_{t-1} \qquad (a)$$

$$= \mathbf{P}^t\mathbf{u}_0 \qquad (b)$$

and similarly for **v**. While operations of the form of (b) indicate more clearly the principle of the use of a transition probability matrix, it is operations of the type shown in (a) which are actually carried out, because they involve only the repeated multiplication of the matrix by a vector, rather than the matrix by the matrix as is needed in (b).

At any generation t, the i^{th} element of $\mathbf{u}_t$ represents $E[q_t | q_0 = i/_{2N}]$, and the i^{th} element of $\mathbf{v}_t$ is equivalent to $E[q_t^2 | q_0 = i/_{2N}]$. Thus

$$E\left[\sigma_q^2 \middle| q_0 = \frac{i}{2N}\right] = v_{t(i)} - [u_{t(i)}]^2$$

and

$$E\left[f \middle| q_0 = \frac{i}{2N}\right] = \frac{v_{t(i)} - [u_{t(i)}]^2}{u_{t(i)}\,[1-u_{t(i)}]} \ .$$

Matrix operations of the type shown above have been carried out with a diploid population size of $N = 10$, for a total of $t = 8N$ generations, with various strengths of selection under two simple models, additive and heterotic. The final generation was chosen as 8N simply because it represents a convenient multiple of N, and corresponds to almost all (in this case 98.2%) of the inbreeding process for a locus with neutral alleles. Extrapolation from $t = 8N$ to $t = \infty$ for the parameter f is a relatively easy matter, as E[f] at $t = \infty$ is 1.

An effective population size of $N = 10$ was chosen because it represents a convenient value for matrix operations. It is now commonly realised (see for example, Crow and Kimura 1970) that generalisations to a wide range of population sizes can be made by expressing the results obtained from one value of N as functions of Ns for the additive model, and of $N(s_1 + s_2)$ for the heterotic model, where s is the selection coefficient for additive selection, and s_1 and s_2 are the selection coefficients for heterotic selection. Thus the two models can be represented as

	A_2A_2		A_1A_2		A_1A_1	
Relative fitness)	$1 - \frac{1}{2}s$	:	1	:	$1 + \frac{1}{2}s$	additive model
	$1 - s_1$	:	1	:	$1 - s_2$	heterotic model

It follows that the transition probability matrix P can be set up by taking $S_{22} = 1 - s/2$ and $S_{11} = 1 + s/2$ for the additive model, and $S_{22} = 1 - s_2$ and $S_{11} = 1 - s_1$ for the heterotic model.

The effect of selection on f

An example of the behaviour of f under additive selection in a finite population is given in Fig. 1, in which f is shown as a function of mean gene frequency at

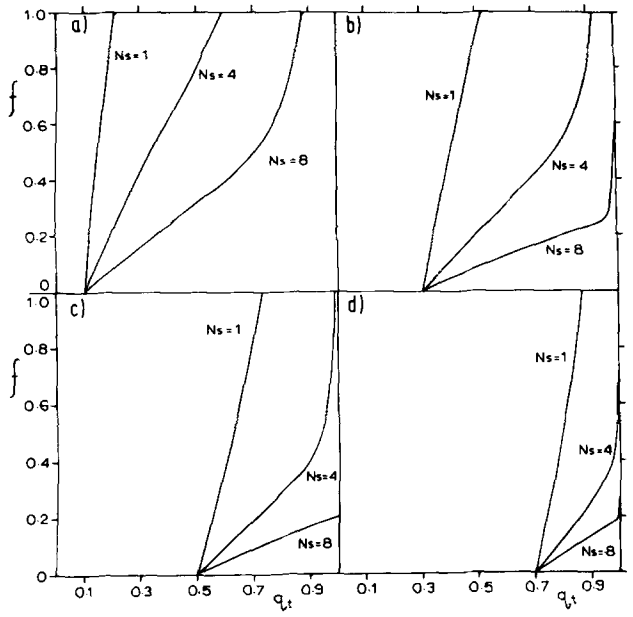

Fig.1. The relationship between standardized variance of gene frequency and mean gene frequency with additive selection with different initial frequencies, drawn from transition probability matrix results

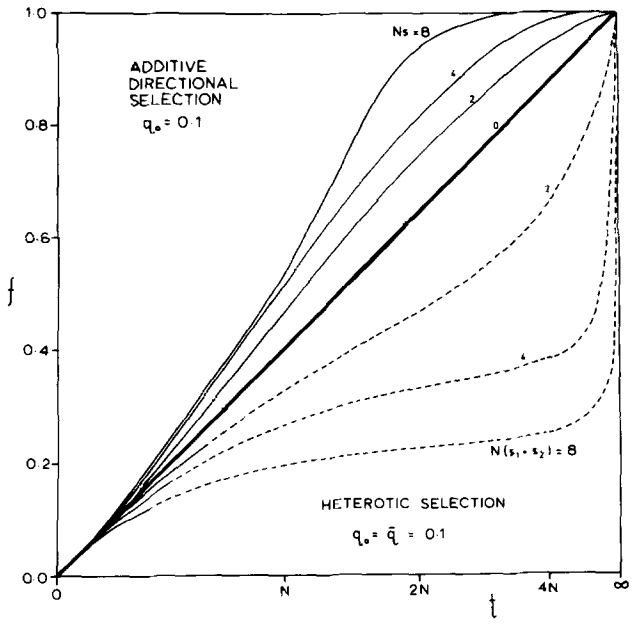

Fig.2. The effect of additive selection and heterotic selection on the standardized variance of gene frequency 0.1. Time is expressed as $1-e^{-t/2N}$, to provide a linear relationship with f with no selection

time t, for the four initial gene frequencies of q_0 = 0.1, 0.3, 0.5 and 0.7. Thus a conceptually infinite population has been subdivided randomly at time t = 0 into several subpopulations each of effective size N. The value of q_0 is the same in all subpopulations, giving $f_0 = 0$. Additive directional selection then occurs with exactly the same coefficient of selection in all subpopulations: the variance of s is zero. The exact matrix results represent the mean value of f which would be observed if the whole process of subdivision followed by selection within subpopulations were repeated a large number of times.

It will be seen that f is linear with q over quite a large range for all cases. The initial slope of the curve can be obtained easily since, when f = 0, we have for the change in the first generation,

$$\Delta f = 1/2N$$

$$\Delta q = s\bar{q}(1-\bar{q})/2$$

so that $df/dq = 1/Ns\bar{q}(1-\bar{q})$.

Another illustration of the way in which f behaves under different strengths of selection for the additive model is given in Fig.2. The results for the heterotic model are also included. The time scale on the x-axis is expressed as $1-e^{-t/2N}$ so as to provide a straight line relationship between f and t in the absence of selection. All the curves in Fig.2 have been obtained for the same initial frequency of allele A_1, namely q = 0.1: results for other initial gene frequencies will be discussed below. In addition, for the heterotic model, it has been assumed that $q_0 = \bar{q}$, where $\bar{q}$ is the large population equilibrium gene frequency, and is given by $s_2/(s_1+s_2)$. This assumption is probably quite a valid description of the situation in real life, because t = 0 in the context of this study represents the time of divergence or separation of one relatively large population into two or more relatively smaller ones. If selection were favouring the heterozygote at a particular locus, then it would not be surprising to find $q = \bar{q}$ in the large population, and hence for any newly formed subpopulation the assumption that $E[q_0] = \bar{q}$ would seem to be quite realistic.

It can be seen from Fig.2 that with low initial frequencies, at any time t additive selection results in f values larger than that expected due to drift alone, and that heterotic selection has the opposite effect. The difference between f under selection and f under drift alone at any time t increases as the values of Ns or $N(s_1+s_2)$ increase. More generally, it has been found that the shape and position of the curves for heterotic selection are very similar for all initial gene frequencies if the initial frequency is the equilibrium value. The effects of heterozygote advantage

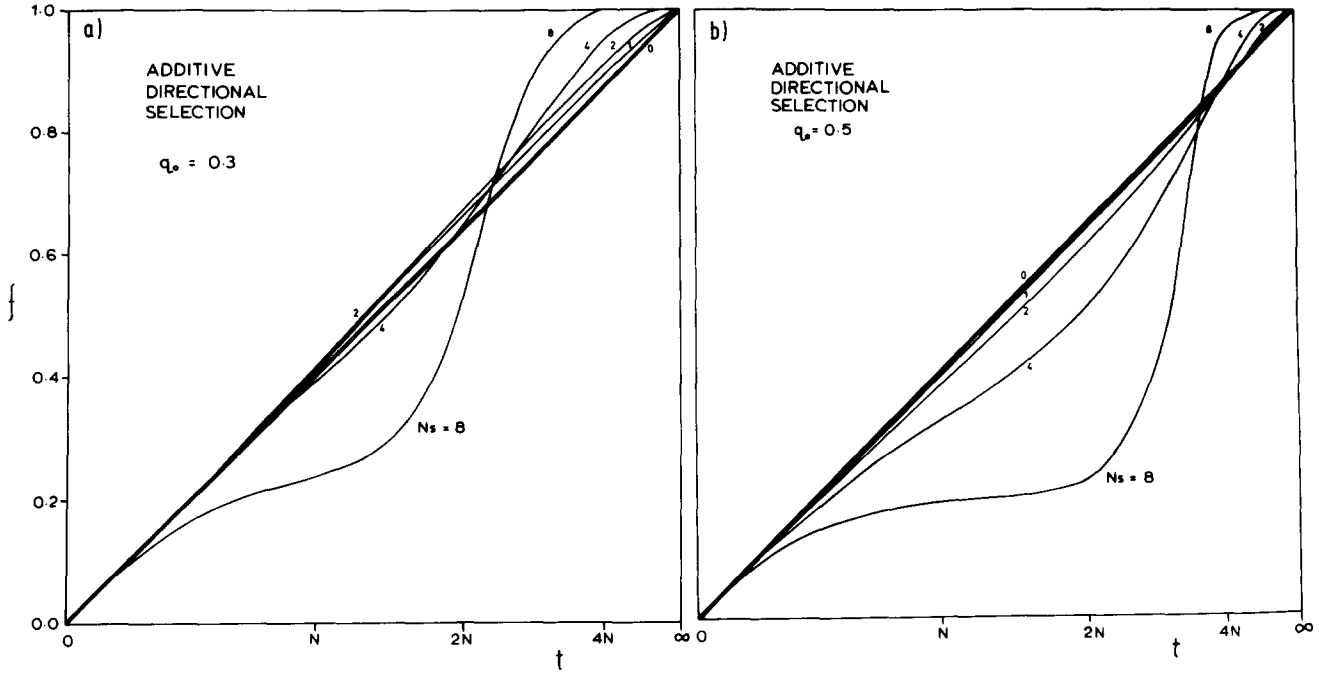

Fig.3. The effect of additive selection on f for initial frequencies of 0.3 and 0.5, with time expressed as $1-e^{-t/2N}$

are thus well in accord with the verbal predictions of Cavalli-Sforza (1966, 1969) and Lewontin and Krakauer (1973).

This apparent independence of f of the equilibrium value of q with heterozygote advantage hides very different behaviour at the different frequencies. When $q_0 = 0.5$, then the mean frequency does not change and the lower f values are due to a true reduction in heterozygosity. But, when $q_0 = 0.1$, heterotic selection does not retard the absolute loss of heterozygosity, it increases it (Robertson 1962). But the mean frequency changes so that the poorer allele is almost always lost from the population. Apparently the decline of the variance is more rapid than that of $\bar{q}$, giving rise to reduced f values.

The effect of additive selection, however, is not so easily generalised. For higher initial gene frequencies, in this case 0.3 and 0.5, Fig.3 shows the effect of various values of Ns on f. It can be seen that f under selection is almost the same as or less than f with drift alone for the majority of the selection process. In fact with $q_0 = 0.5$ the curves for additive selection now resemble the curves for heterotic selection, except for relatively high values of f, of the order of 0.8 or more.

An explanation of the behaviour of f with additive selection can be obtained from Eq.(1), which may be rewritten

$$\Delta V = s[V(1-2\bar{q}) - \mu_3] \qquad (2)$$

whereas for the change in $\bar{q}(1-\bar{q})$, we have

$$\Delta(\bar{q}(1-\bar{q})) = (1-2\bar{q})\Delta q = s\bar{q}(1-\bar{q})(1-2\bar{q})(1-f)/2$$

so that we have

$$f_1 = \frac{V(1+s(1-2\bar{q}) - \mu_3/V)}{\bar{q}(1-\bar{q})(1+s(1-2\bar{q})(1-f)/2)}$$

$$= f(1+s[(1-2\bar{q})(1+f)/2 - \mu_3/V]) \text{ approx.}$$

Further, it can be shown that if s is small, then $V = \bar{q}(1-\bar{q})f$ whereas $\mu_3 = \frac{1}{2}\bar{q}(1-\bar{q})(1-2\bar{q})(3f^2-f^3)$. Thus, in the early stages of selection, the second term in 2) will be smaller than the first and the effect of selection on f will depend on $(1-2\bar{q})$.

Thus the deviation of f from the expected values without selection depends on the initial gene frequency. If this is low, the effect of selection is to increase f above expectation in the early stages and to maintain this throughout, as illustrated in Fig.2 for $q_0 = 0.1$. When $q_0 = 0.3$, both terms in 2) are small when f is small and when Ns = 8, f remains close to expectation as the gene frequency passes through the range close to q = 0.5 where selection has little effect on f.

When $q_0 = 0.5$, on the other hand, selection at the higher values of Ns quickly moves the frequency out of the central range and f values are below theoretical expectation. At very high values of the mean gene frequency, usually higher than 0.95, f suddenly increases until, as fixation is approached, it is greater than theoretical values. A plot of $\log_e(1-f)$ against t (which in the absence of selection would have slope 1/2N) shows that this final acceleration occurs at all values of q_0. It is a consequence of the fact that the limiting rate of loss of heterozygosity increases as Ns increases (see Kimura 1955).

The computer output also included the value of the third moment of gene frequency about the mean. This was usually positive when $\bar{q}$ was less than 0.5 and vice versa though the change in sign did not occur exactly at $\bar{q} = 0.5$. Thus the two terms in 2) are usually opposite in sign. The first was usually the larger though when $q_0 = 0.1$, the two terms were approximately equal for $\bar{q}$ values greater than 0.6.

It can be concluded that heterotic selection always results in f values lower than those expected with drift in the absence of selection. Additive directional selection will produce similarly low values of f unless initial gene frequency is low, or unless observations are made relatively late in the selection process, when f values expected due to drift alone are of the order of 0.7 or greater. In these two situations f with selection is greater than f with drift alone. It is further evident that even with quite large values of Ns or $N(s_1 + s_2)$, the effect of selection on f will never be detected if the observed value of f is less than say 0.1. The effect of selection on f becomes most apparent as f due to drift approaches intermediate values.

Discussion

The results of this study are in broad agreement with verbal predictions already available of the effect of selection on the standardized variance of gene frequency. What has become evident, however, is the way in which two simple models of selection are sufficient to provide expected values of f which cover almost the entire possible range of f values. Furthermore, the possible range of f values at any particular time t can be substantially extended when consideration (not described here) is given to directional selection for a recessive, and for a dominant gene. It must therefore be concluded that while heterogeneous f values certainly can be taken as evidence of selection, any subsequent inference as to the type of selection operating is bound to be of very limited validity in the absence of knowledge of initial gene frequencies.

Literature

Bodmer, W.F.; Cavalli-Sforza, L.L.: A migration matrix model for the study of random genetic drift. Genetics 59, 565- 592 (1968)

Cavalli-Sforza, L.L.: Population structure and human evolution. Proc. Royal Soc. Series B, 194, 362-269 (1966)

Cavalli-Sforza, L.L.: Human diversity. Proc. Twelfth Intern. Congr. Genetics 3, 405-416 (1969)

Cavalli-Sforza, L.L.; Zei, G.: Experiments with an artificial population. Proc. Intern. Congr. on Human Genetics, pp. 473-478 (1967)

Crow, J.F.; Kimura, M.: An introduction to population genetics theory. New York: Harper and Row 1970

Garnett, I.; Falconer, D.S.: Protein variants in strains of mice differing in body size. Genet. Res. 25, 45-58 (1975)

Hill, W.G.; Robertson, A.: The effect of inbreeding at loci with heterozygote advantage. Genetics 60, 615-628 (1968)

Kimura, M.: Stochastic processes and the distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. 20, 33-53 (1955)

Lewontin, R.C.; Krakauer, J.: Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74, 175-195 (1973)

Robertson, A.: Selection for heterozygotes in small populations. Genetics 47, 1291-1300 (1962)

Robertson, A.: Gene frequency distributions as a test of selective neutrality. Genetics 81, 775-785 (1975)

Received June 11, 1976
Communicated by H. Stubbe

F.W. Nicholas
Dept. of Animal Husbandry
University of Sydney
Sydney NSW (Australia)

Dr. A. Robertson
Institute of Animal Genetics
West Mains Road
Edinburgh EH9 3JN (Scotland)