Theor Appl Genet (1993) 86:775–780

THEORETICAL AND APPLIED GENETICS

Effects of population size and linkage on optimal selection intensity

F. Hospital[1], C. Chevalet[2]

[1] Institut National de la Recherche Agronomique, Station de Génétique Végétale, Ferme du Moulon, 91190 Gif sur Yvette Cedex, France
[2] Institut National de la Recherche Agronomique, Laboratoire de Génétique Cellulaire, BP 27 – 31326 Castanet Tolosan Cedex, France

Received: 30 November 1992 / Accepted: 4 January 1993

Abstract. Following Robertson (1970a) it is generally considered that for mass selection the selected proportion that maximises ultimate response is 0.5. This prediction has been partly tested by different authors. Here we explicitly address the question using computer simulations of selection in finite populations with linkage. The results show that the response achieved is always lower than the one predicted by standard methods, and that optimum selection intensity may be much lower than predicted unless population size is small.

Key words: Selection response – Linkage – Genetic drift – Computer simulations

Introduction

As soon as predictions are not restricted to the first few generations, the effects of continuous selection on a quantitative character depend in particular on population size and on the number and distribution of the genes controlling the character. However, due to excessive mathematical complexity, these parameters are hardly taken into account in analytical approaches. Analytical solutions rely on hypotheses concerning either the genetic system: one locus (Robertson 1960), two linked loci (Hill and Robertson 1966), many unlinked loci (Bulmer 1971; Verrier et al. 1990), or else allelic effects: Gaussian distribution (Lande 1976; Chevalet 1988), or distribution restricted to the first four moments (Turelli and Barton 1990).

Communicated by E. J. Eisen
Correspondence to: F. Hospital

Most predictions on mid- and long-term response to selection are based on the approximate analyses initiated by Dempster (1955) and Robertson (1960) and subsequently developed by Robertson (1970a). They depend on the ultimate probability of fixation of a favourable allele in a finite population (Malécot 1952; Kimura 1957), on the assumption of Gaussian distributions of genotypes and phenotypes, and on the change with time of genetic variance under random genetic drift:

$$V_A(t) = V_A(0)\left(1 - \frac{1}{2N}\right)^t \tag{1}$$

where $V_A(t)$ is the genetic variance in the tth generation and N the effective population size (Wright 1931). The change in mean breeding value G in one generation is then:

$$\Delta G(t) = G(t) - G(t-1) = \frac{iV_A(t-1)}{\sigma} \tag{2}$$

where i is the normalized selection differential and σ the phenotypic standard deviation. Assuming σ constant, the cumulative response after t generations of selection is:

$$G(t) - G(0) = \sum_{k=1}^{t} \Delta G(k)$$

$$\simeq 2N\frac{iV_A(0)}{\sigma}(1 - e^{-(t/2N)}). \tag{3}$$

If at each generation the N individuals selected are a proportion p of the T individuals measured, we have:

$$N = Tp \text{ and } i(p) = \frac{z(p)}{p}$$

with $z(p)$ being the ordinate of the normal distribution at the point where an upper tail area p is cut off. Hence $Ni(p) = Tz(p)$, and this relationship leads to the prediction of the expected genetic response at intermediate generations t, as a function of the proportion selected p:

$$G(t) - G(0) \simeq 2Tz(p)\frac{V_A(0)}{\sigma}(1 - e^{-(t/2Tp)}). \quad (4)$$

Since $z(p)$ is maximal for $p = 0.5$, these approaches were used to show that the optimal proportion of individuals that should be used for reproduction is 0.5, if the aim is to obtain the highest possible ultimate response. Furthermore, equation 4 in the limit yields the same qualitative result as that derived from the one-locus theory of Malécot (1952) and Kimura (1957), namely that the ultimate cumulative response is $2N$ times the response in the first generation.

In theory, these approximations hold in the limiting case of weak selection, a situation obtained under the assumptions of the infinitesimal model if the trait is due to infinitely many unlinked genes with small individual effects. In this paper, we investigate the robustness of the predictions drawn from equation 4 with respect to departures from its basic assumptions, from both qualitative and quantitative points of view. Because analytical approaches cannot be developed unless similar hypotheses are stated, we systematically used computer stochastic simulations of the process to investigate the effects of population size, linkage, and selection intensity on the response to selection. The analysis is restricted to a purely additive model of gene action.

Model and method

The computer program simulates selection in a finite population. A diploid individual is described by n pairs of loci with a recombination fraction r between them, along the map length L (in centiMorgans) on which they are distributed. Each locus has two alleles with additive effects 0 and 1, so that the genotypic value of an individual is the sum of all ones at all $2n$ loci. The phenotypic value is the sum of the genotypic value and a random normal variable with mean 0 and an appropriate variance to achieve the specified heritability in the first generation. The population is split into two groups (sexes) of equal size. In each group, a proportion p of individuals is selected in each generation. One parent is drawn at random from each selected group, and produces a gamete after cross-overs have been generated according to a Poisson distribution assuming no interference, the two gametes forming one offspring zygote. This procedure is repeated until the required number T of zygotes is obtained. Generations are non-overlapping, and the process starts with an initial population obtained by drawing at random two alleles at each locus for all individuals, with a specified frequency q of favourable alleles (alleles with effects equal to 1).

The parameters are:

n, total number of loci
r, recombination fraction between adjacent loci
L, map length in centiMorgans on which loci are evenly distributed
q, frequency of the favourable allele at each locus in the initial population
h^2, realized heritability in the first generation
T, total number of zygotes in each generation ($T/2$ males and $T/2$ females)
p, proportion selected in each group at each generation
t, time in generations, starting at 0 for the initial population.

For each set of parameters, the program simulates 1000 runs of selection from a random initial population until complete fixation at all loci for nine different values of $p(0.1, 0.2, \ldots, 0.8, 0.9)$. The output at each generation is the average response over the 1000 runs for each p value, or the value of p corresponding to the highest average response.

Results

Figure 1 presents the average responses obtained with 50 loci equally spaced on a chromosome of length 50 cM for different values of p and two population sizes, at the same t/T. In the case of a small population size ($T = 20$), for low values of p the response is high in the early generations but soon reaches a limit, and for high values of p the response per generation is lower but is sustained over a longer period of time. Thus, the responses at the limit are roughly symmetrical around a maximal value obtained for a p of about 0.5. These qualitative results are not much different from the theoretical predictions. For a larger population size ($T = 200$), responses for all p values are higher than in the preceding case, and the same contrast can be seen between fast but limited response to strong selection and low but longer-lasting response to weak selection. The important feature is that the limit reached with weak selection is much higher than that obtained under strong selection, so that the optimal p value

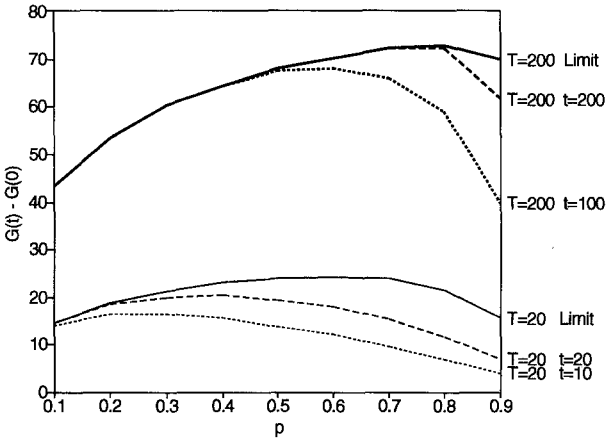

Fig. 1. Response to selection with nine proportions p of selected individuals $n = 50$, $L = 50$, $q = 0.2$, $h^2 = 0.5$. Results for two population sizes, T, are shown at the limit and at two intermediate generations t such that t/T is the same for different T

exceeds 0.5 after about the 60th generation, being equal to 0.8 in the limit.

The discrepancy between these results and the theoretical prediction is seen in Fig. 2, where we have plotted both the optimal p values for each generation and the one predicted according to equation 4, on a common t/T logarithmic abscissa. Whereas the curves for $T = 20$ agree with the theoretical predictions, the optimal proportion for $T = 200$ is always much higher, even at intermediate generations long before the limit; the approximate theory underestimates the optimal p value as soon as population size is not very small. The various conditions considered in Fig. 2 show that this

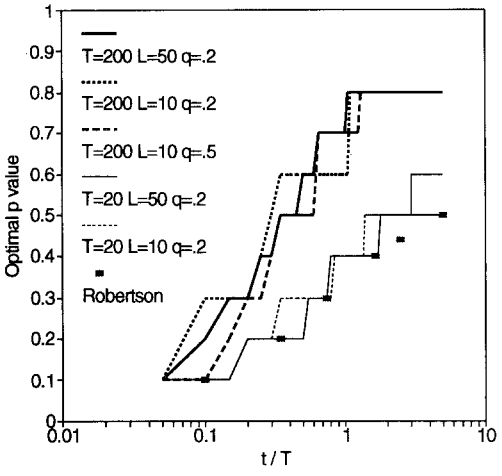

Fig. 2. Optimal proportion p giving the highest response at each t plotted against t/T. Fifty loci spread over two map lengths, L, for two population sizes, T, and two initial frequencies q. $h^2 = 0.5$. Simulation results can be compared to the results derived from equation 4 by Robertson (1970a)

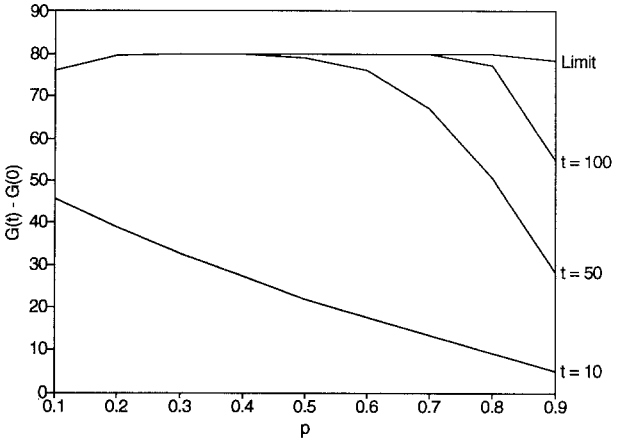

Fig. 3. Response to selection with nine proportions p of selected individuals independent loci. $n = 50$, $T = 200$, $q = 0.2$, $h^2 = 0.5$. Results at the limit and at three intermediate generations, t

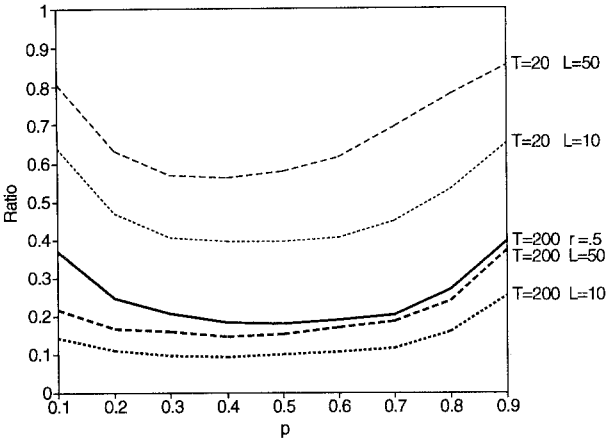

Fig. 4. Ratio of response at the limit to $2N$ times the response in the first generation: $\{[G(\infty) - G(0)]/2N[G(1) - G(0)]\}$. Fifty loci either independent or spread over two map lengths, L, for two population sizes, T. $q = 0.2$, $h^2 = 0.5$

discrepancy appears to depend mainly on population size, and not much on the tightness of linkage or the initial gene frequencies.

The case of unlinked loci is considered in Fig. 3. It appears that with 50 independent loci, the maximum possible response, corresponding to fixation of *all* favourable alleles, is obtained with any proportion selected between 0.2 and 0.8, giving a flat optimum, which is at variance with the sharper optimum predicted by infinitesimal theory (equation 4). If many independent loci are considered ($n = 250$, data not shown) the results become closer to the predictions of the infinitesimal model. Such a situation seems to be outside the range of realistic values for n, while $n = 50$ is about the maximum number of simultaneously independent loci that could be considered on a genome.

A test of the prediction that the ultimate response is expected to be $2N$ times the response in the first generation is shown in Fig. 4. Comparing corresponding curves for $T = 20$ and $T = 200$ clearly shows that the larger the population the farther are our simulation results from theoretical predictions. The divergence is both quantitative and qualitative. From a quantitative point of view, the ratio $\{[G(\infty) - G(0)]/2N[G(1) - G(0)]\}$ is smaller than one for all values of p, showing that the theoretical predictions always overestimate selection response. From a qualitative point of view, two points are worth stressing, considering for example the case $T = 200$, $L = 50$. First, the curve is not symmetrical – for two values of p symmetrical about 0.5, the ratio is always lower for the low p values, indicating that the theoretical overestimation of the response is more important for strong selection than for weak selection. Second, the curve is concave, suggesting that

the theory overestimates the long-term response more the higher this response.

Discussion and conclusion

The discrepancies between theory and simulations can be seen more sharply by considering the respective predictions for the change with time of additive genetic variance; some illustrations are given in Fig. 5. The comparison of our results with the theoretical predictions shows that the latter underestimate the decrease of variance with time, the more so for greater selection intensity. Even in the case of unlinked loci, the theoretical predictions have little connection with the results observed in the simulations.

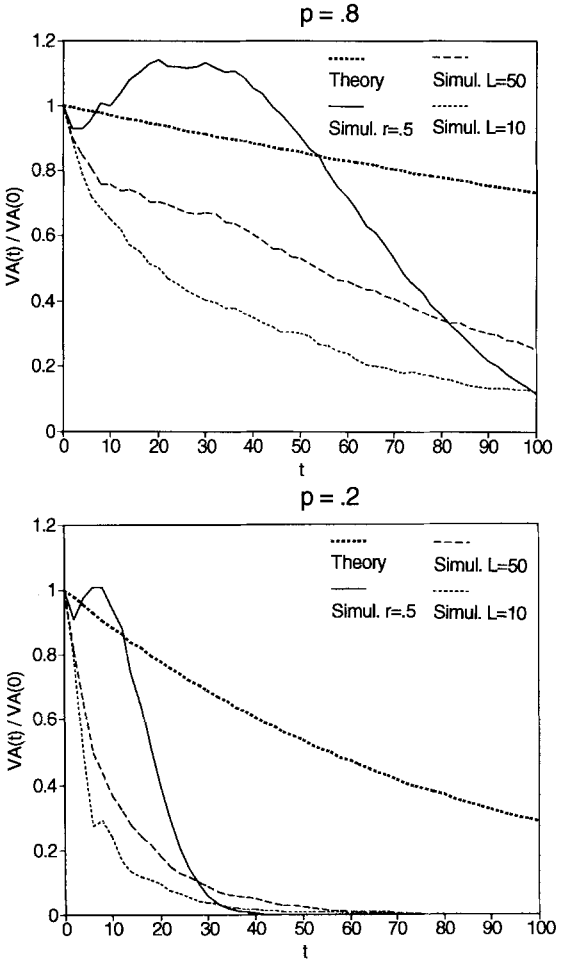

Fig. 5. Decrease with time of additive genetic variance relative to variance in the initial population, for two selected proportions. Simulation results with $n = 50$, three linkage values, $T = 200$, $q = 0.2$ and $h^2 = 0.5$ can be compared to corresponding theoretical predictions from equation 1

When dealing with selection in finite populations, the theory derived from equation 1 only takes account of drift. In this case, the reduction of variance due to selection relies only on the reduction of population size T to its effective value $N = Tp$, whereas selection by itself and its interactions with drift have other effects that may reduce the additive variance available for further selection.

First, directional selection on a multilocus system generates negative covariances between gene effects at different loci. This was first described by Bulmer (1971). The consequences on variance can be summarized by writing:

$$V_A = V_g + C$$

where V_A is the total additive genetic variance available for selection, V_g is the sum of variances of gene effects at each locus, that is the total additive variance in the case of complete linkage equilibrium ("*genic variance*" in Bulmer's terminology), and C is the sum of covariances between gene effects over all pairs of different loci. In the case of directional selection, C is negative, so that the variance available for selection is reduced from V_g to V_A. In a purely additive multilocus system such as the present model, C is strictly equivalent to the sum of linkage disequilibria between all pairs of different loci, and the simulations showed that selection actually generates strong negative pairwise disequilibria (data not shown).

Second, negative linkage disequilibrium means that gametes in repulsion (bearing both favourable and unfavourable alleles at different loci) are more frequent than gametes in coupling in the parent population. The tighter the linkage, the more gametes are likely to stay in repulsion in the offspring generation. Selection on such gametes then induces "*hitch-hiking*" effects of unfavourable alleles in linkage disequilibrium with favourable ones, so that unfavourable alleles may increase in frequency up to fixation. This effect can be described by considering a reduced effective population size in equation 1 (Birky and Walsh 1988). This will possibly reduce both response at intermediate generations and at the end of the process, giving a response at the limit which is lower than the maximum possible one. This effect is illustrated in Table 1, which shows the percentage of unfavourable alleles fixed at the limit, and the time at which this limit is reached. The difference between fixation under weak and strong selection is again much higher for larger population size.

These two effects reduce the available genetic variance and the corresponding response. They are due to selection and are more intense as selection is stronger. The observed response depends finally on the reduction of variance due to both drift and selection.

Table 1. Fixation of unfavourable alleles. Fix 0 = percentage of unfavourable alleles fixed at the limit; t = generation at which this limit is reached, 50 loci, $q = 0.2$, $h^2 = 0.5$

Map	p	$T = 20$		$T = 200$	
		Fix 0	t	Fix 0	t
$r = 0.5$	0.2	47.7	66	0.5	56
	0.8	48.0	252	0.2	198
$L = 50$	0.2	61.2	106	26.6	100
	0.8	58.6	234	7.2	304
$L = 10$	0.2	66.1	53	45.2	113
	0.8	65.3	212	32.4	428

For very small population size, the effects of drift appear to override the effects of selection, so that the optimum p value shown by the simulations is close to the one predicted from equation 4.

For a population which is not too small, the effects of selection are no longer negligible, resulting in a more severe reduction of variance at low values of p. Since these effects of selection are not taken into account in equation 1, the stronger the selection, the more the theory overestimates the response. Hence, the theory underestimates the optimal value of p, the proportion selected.

The effect of variation in gene frequencies can be seen from the evolution of variance with independent loci (Fig. 5). The early *increase* of variance in this case, which is not taken into account by equation 1, yields a seemingly correct quantitative adjustment of simulations with theory. In fact, this masks important qualitative differences between variance behaviour in the two cases.

Studies have been published that check the theoretical predictions of equation 4, using either experimentation (Ruano et al. 1975; Frankham 1977) or computer simulations (Robertson 1970b; Harris 1982). Concerning the quantitative evaluation of selection response, the results were generally lower than the theoretical predictions. Concerning the qualitative definition of an optimum intensity of selection, they either showed disagreement with theory, but on too few combinations of the parameters to reach a general conclusion, or agreement under special conditions which in the light of our results appear to be restricted to cases where the theory holds approximately. Moreover, the slight increase in the optimal p value due to linkage noticed by both Hill and Robertson (1966) and Robertson (1970b) was not considered by these authors as inconsistent with the theoretical predictions. Hence, the theoretical predictions concerning this optimum were never clearly disputed.

The present work allows us to emphasize the respective effects of population size and linkage. First, the theory very strongly overestimates the mid- and long-term responses to continuous selection as soon as the genetic assumptions are not restricted to the infinitesimal model. Second, qualitative differences occur as soon as the population size is not very small: the optimum for the proportion p of selected individuals is shifted towards values higher than 0.5.

The practical consequences of our results may be important for mid- and long-term economical evaluation of selection programs. The usual methods for predicting the response to selection over several generations are derived from an approximation that holds for weak selection. Our results show that, apart from quite unrealistic situations (very small population size, a quantitative trait due to several hundred unlinked loci with equal effects), these predictions are generally unreliable except in the first few generations. We emphasize that linkage, and the joint effects of selection and drift on additive genetic variance, must not be neglected in selection theory.

These joint effects are illustrated in Fig. 6, which shows the predictions of genetic variance decrease under continuous selection according to various models taking into account drift alone (equation 1, Robertson 1970a), selection in an infinite population for the infinitesimal model (Bulmer 1980), both selection and drift for the infinitesimal model (Verrier et al. 1990), and selection and drift in the multinormal approximation for a finite number of linked loci (Chevalet 1988). Comparing these curves to the results of our simulations clearly shows that none of these models is able to properly describe the genetic system. Actually, Turelli and Barton (1990) showed that the resolution of the complete system of equations is not possible without some simplifying hypothesis. However, one can

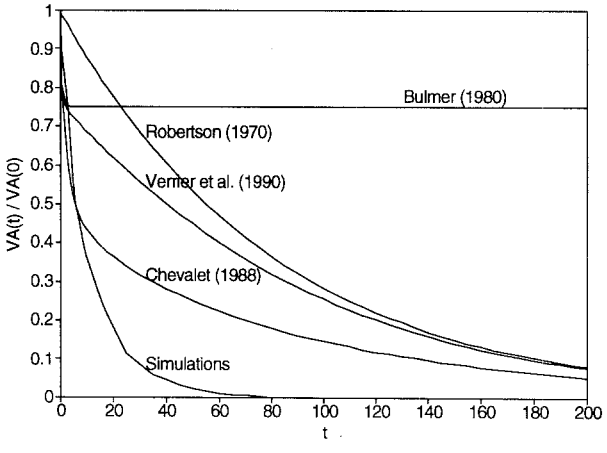

Fig. 6. Decrease of variance observed in the simulations as compared to that predicted by other models as explained in the text. $n = 50$, $L = 50$, $T = 200$, $p = 0.2$ and $h^2 = 0.5$ for models taking any of these parameters into account

still hope to obtain an analytical description of the system, using empirical laws for the behaviour of multilocus systems under selection that can be provided by simulations (Franklin and Lewontin 1970; Robertson 1977; Robertson and Hill 1983).

Deriving better predictions will require a generalization of the model to take account of non-additive gene action, non-uniform distribution of gene effects (some genes with large, some with small effects), and mutation, which was beyond the scope of the present paper. It will also require data on the number of quantitative trait loci and their linkage relationships. New analytical approaches would be valuable for the general understanding of genetic structure in selected populations, and for the analysis of complex population structures (overlapping generations, several groups with different genetic levels, subgroups involving very many individuals). However, computer simulations can already be used directly by including the programs used here in programs which take account of the population structures encountered in actual selection schemes. This would make it possible to develop more realistic predictions of the response to selection in the situations encountered in applied quantitative genetics. Moreover, it will probably be easier to include genetic marker data in simulation programs than in analytical tools, when such new genetic data become available to develop marker-assisted selection schemes.

Acknowledgements. Thanks are due to Daniel Wallach for careful reading of the manuscript.

References

Birky CW Jr, Walsh JB (1988) Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci USA 85: 6414–7518

Bulmer MG (1971) The effect of selection on genetic variability. Am Nat 105:201–211

Bulmer MG (1980) The mathematical theory of quantitative genetics. Clarendon Press, Oxford

Chevalet C (1988) Control of genetic drift in selected populations. In: Weir BS, Eisen EJ, Goodman MM, Namkoong G (eds) Proc 2nd Int Conf Quant Genet, Raleigh NC May 31–June 5 1987. Sunderland, Mass., pp 379–394

Dempster ER (1955) Genetic models in relation to animal breeding. Biometrics 11:534

Frankham R (1977) Optimal selection intensities in artificial selection programmes: an experimental evaluation. Genet Res 30:115–119

Franklin IR, Lewontin RC (1970) Is the gene the unit of selection? Genetics 65:707–734

Harris DL (1982) Long-term response to selection I. Relation to breeding population size, intensity, and accuracy with additive gene action. Genetics 100:511–532

Hill WG, Robertson A (1966) The effects of linkage on limits to artificial selection. Genet Res 8:269–294

Kimura M (1957) Some problems of stochastic processes in genetics. Ann Math Stat 28:882–901

Lande R (1976) The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genet Res 28:221–235

Malécot G (1952) Les processus stochastiques et la méthode des fonctions génératrices ou caractéristiques. Publ Inst Stat Univ Paris 1:1–25

Robertson A (1960) A theory of limits in artificial selection. Proc R Soc Lond B 153:234–249

Robertson A (1970a) Some optimal problems in individual selection. Theor Pop Biol 1:120–127

Robertson A (1970b) A theory of limits in artificial selection with many linked loci. In: Kojima K (ed) Mathematical topics in population genetics. Springer-Verlag, Berlin, pp 246–288

Robertson A (1977) Artificial selection with a large number of linked loci. In: Pollak E, Kempthorne O, Barley TB Jr (eds) Proc 1st Int Conf Quant Genet. Iowa State Univ Press, Ames, pp 307–322

Robertson A, Hill WG (1983) Population and quantitative genetics of many linked loci in finite populations. Proc R Soc Lond B 219:253–264

Ruano RG, Orozco F, López-Fanjul C (1975) The effect of different selection intensities on selection response in egg-laying of *Tribolium castaneum*. Genet Res 25:17–27

Turelli M, Barton NH (1990) Dynamics of polygenic characters under selection. Theor Pop Biol 38:1–57

Verrier E, Colleau JJ, Foulley JL (1990) Predicting cumulated response to directional selection in finite panmictic populations. Theor Appl Genet 79:833–840

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159