Genet. Res., Camb. (1978), **31**, *pp.* 239–254
With 4 text-figures
Printed in Great Britain

The effect of finite population size on models of linked overdominant loci

By P. J. AVERY

Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN Scotland

(*Received* 9 *November* 1977)

SUMMARY

Models of two linked overdominant loci in moderately large, but finite, populations are examined by looking at the variance–covariance matrix of the two gene frequencies and the linkage disequilibrium around stable deterministic equilibrium points. In particular, the effect of genetic drift is examined in cases where, in infinite populations, two stable equilibria with non-zero linkage disequilibrium, D, are maintained. Theoretical formulae are produced and checked by computer simulation. In general, the results show that unless the population size is very large indeed, genetic drift causes the values of D to vary considerably about the equilibrium values and that for many models, where stable equilibria exist at non-zero D values, a wide range of values of D have a high probability. Thus it is very difficult to draw conclusions about the selection regime by measuring linkage disequilibrium in a finite population.

1. INTRODUCTION

Much theoretical work has been done on selected linked loci in order to see how linkage disequilibrium could be maintained between pairs of loci. However, in nearly all studies a deterministic approach has been taken, i.e. the populations have been assumed to be infinitely large, and then different models have been examined to see if polymorphic equilibria exist and what conditions are necessary for these equilibria to be stable. Bodmer & Felsenstein (1967) and, more recently, Karlin (1975) have given a very comprehensive review of the work done using a deterministic approach. The assumption of infinite population size, however, is a very restrictive one and it is necessary to test how large populations need to be before the deterministic results become useful. The analysis of selected linked loci in finite populations is, on the whole, very complicated and thus relatively little work has been done in this area. Levin (1969), Hill (1969) and Hill & Robertson (1968) examined the rate of fixation of linked overdominant loci. Sved (1968) and Hill & Robertson (1968) have considered the unfixed population for two loci, and Franklin & Lewontin (1970) and Yamazaki (1977) have considered this for particular multi-locus models, but most of the work is computer simulation. The only major theoretical work on the examination of unfixed populations for selected linked loci is by Felsenstein (1974). In the absence of mutation or migra-

tion, finite population size will cause loci to eventually be fixed. However, if selection is quite strong and population size moderately large, fixation will be put off virtually indefinitely. In his paper, Felsenstein has considered this case with models which have a single stable equilibrium in linkage equilibrium. After an initial transient stage, the finite population size causes a stable probability distribution of values of the gamete frequencies around the equilibrium values and this distribution was examined by Felsenstein (1974). In this paper we will clarify Felsenstein's approach and extend it to the solution of models which have stable deterministic equilibria with linkage disequilibrium. By examining various selection models we hope to examine the validity of deterministic equilibria analysis to the explanations of experimental determinations of linkage disequilibrium in finite populations, particularly laboratory populations.

2. MODEL AND GENERAL METHOD

The model and method which will be used is the same as used in Felsenstein (1974). A summary will be given in this section but the original paper should be consulted for a more complete explanation. We consider a diploid population of constant size N with discrete generations, mating being at random (i.e. selfing is allowed). Two loci, each with two alleles (i.e. A and a, and B and b) will be considered. Let x_i ($i = 1, 2, 3, 4$) be the frequencies of gametes AB, Ab, aB, ab respectively, let p be the gene frequency of A, let q be the gene frequency of B, and let D be the linkage disequilibrium between the two loci (i.e. $D = x_1x_4 - x_2x_3$). Let w_{ij} be the fitness of the genotype whose two gametes have frequencies x_i and x_j in the population. Then the marginal fitnesses, w_i, and the average fitness $\bar{w}$ are given by

$$w_i = \sum_{j=1}^{4} x_j w_{ij} \quad (i = 1, 2, 3, 4) \tag{1}$$

and

$$\bar{w} = \sum_{j=1}^{4} x_j w_j. \tag{2}$$

These two equations are exact if the x_i's are measured in an infinite population. However, if they are measured from the chromosomes of a finite population at a particular generation they will only be approximately true due to departures from Hardy–Weinberg proportions in the pairing of gametes to form individuals. We will be concerned with fairly large populations and thus these departures will be small and can be ignored.

We shall consider the effect of finite population size by looking at the variance–covariance matrix of p, q and D around a stable equilibrium position (i.e. $\hat{p}$, $\hat{q}$ and $\hat{D}$). If a model has more than one stable equilibria we shall assume that they are far enough apart that they can be considered as two separate distributions around each equilibrium point.

Given values of p, q and D at a particular time t, we can approximate their values at the next generation by splitting them into two components. The first part is that obtained by ignoring finiteness and using the deterministic equations

due to selection and recombination and the second is a random sampling term. We further approximate the deterministic part by assuming that p, q and D at time t are near $\hat{p}$, $\hat{q}$ and $\hat{D}$, the equilibrium values, so that the deviations from the equilibrium values at the next generation are linear functions of the deviations at the previous generation. Putting this in matrix form, we obtain

$$\mathbf{d}_{t+1} = \begin{pmatrix} p_{t+1}-\hat{p} \\ q_{t+1}-\hat{q} \\ D_{t+1}-\hat{D} \end{pmatrix} = \mathbf{A}\mathbf{d}_t + \mathbf{e}, \tag{3}$$

where **A** is a matrix of the first terms of the multivariate Taylor Series expansions for p_{t+1}, q_{t+1} and D_{t+1} as functions of p_t, q_t and D_t evaluated at $\hat{p}$, $\hat{q}$ and $\hat{D}$ and the subscript denotes generation number. **e** is assumed to follow a multivariate normal distribution with mean zero and variance–covariance matrix, **Q**. **Q** is assumed to be independent of the current position, $\mathbf{d}_t$. This independence and the multivariate normal nature of **e** are only approximately true. If **C** is the variance–covariance matrix of **d** for the stable distribution, then as proved by Felsenstein (1974),

$$\mathbf{C} = \mathbf{A}\mathbf{C}\mathbf{A}' + \mathbf{Q}, \tag{4}$$

where a prime denotes the transpose of a matrix. This is a standard result in time series analysis for a multivariate first-order autoregressive process.

In the following sections we shall consider the symmetric viability model of Bodmer & Felsenstein (1967), i.e.

	BB	*Bb*	*bb*
AA	$w_{11} = 1-\delta$	$w_{12} = w_{21} = 1-\beta$	$w_{22} = 1-\alpha$
Aa	$w_{13} = w_{31} = 1-\gamma$	$w_{14} = w_{41} = w_{23} = w_{32} = 1$	$w_{24} = w_{42} = 1-\gamma$
aa	$w_{33} = 1-\alpha$	$w_{34} = w_{43} = 1-\beta$	$w_{44} = 1-\delta$

A more general model could be assumed and followed, if the deterministic equilibria and their stability were known. However, in general, this is not the case and thus in order to try to see the overall trends we shall examine the above fitness model.

3. LINEARIZATION MATRIX

Bodmer & Felsenstein (1967) showed that for the symmetric viability model, an equilibrium of the form $x_1 = x_4 = x$, $x_2 = x_3 = \frac{1}{2}-x$ always occurred, where $0 \leqslant x \leqslant 0{\cdot}5$ and x is a function of the recombination fraction, c, between the two loci and the fitness parameters. Thus we shall consider the behaviour of the system around a stable equilibrium point given by $\hat{p} = \hat{q} = \frac{1}{2}$ and $\hat{D} = -\frac{1}{4}+x$. From deterministic theory (e.g. Lewontin & Kojima, 1960) we know that

$$x_{i,t+1} = (x_{i,t}w_{i,t} - ck_i w_{14} D_t)/\overline{w}_t \quad (i = 1, 2, 3, 4), \tag{5}$$

where $k_1 = k_4 = 1$, $k_2 = k_3 = -1$ and $w_{14} = w_{23}$ (i.e. no difference between coupling and repulsion phase),

$$p_{t+1} = (x_{1,t}w_{1,t} + x_{2,t}w_{2,t})/\overline{w}_t, \tag{6}$$

$$q_{t+1} = (x_{1,t}w_{1,t} + x_{3,t}w_{3,t})/\overline{w}_t, \tag{7}$$

$$D_{t+1} = (x_{1,t}x_{4,t}w_{1,t}w_{4,t} - x_{2,t}x_{3,t}w_{2,t}w_{3,t} - cw_{14}D_t\overline{w}_t)/\overline{w}_t^2. \tag{8}$$

The equilibrium solutions for a particular model can be found from equations

(5)–(8) by putting $x_{i,t+1} = x_{i,t}$, $p_{t+1} = p_t$, $q_{t+1} = q_t$ and $D_{t+1} = D_t$. All the equilibrium points for the models which we shall consider are given by Bodmer & Felsenstein (1967). Using equation (1) and dropping the subscript, t, used above, we obtain

$$\frac{\partial w_i}{\partial p} = w_{i2} - w_{i4} + q\epsilon_i, \tag{9}$$

$$\frac{\partial w_i}{\partial q} = w_{i3} - w_{i4} + p\epsilon_i, \tag{10}$$

$$\frac{\partial w_i}{\partial D} = \epsilon_i, \tag{11}$$

for $i = 1, 2, 3, 4$, where $\epsilon_1 = \epsilon_4 = (2S - \delta)$, $\epsilon_2 = \epsilon_3 = -(2S - \alpha)$ and $S = \frac{1}{2}(\beta + \gamma)$. From equation (6) we find, after correcting a few typographical errors of Felsenstein (1974), that

$$\frac{\partial p_{t+1}}{\partial p} = \frac{1}{\overline{w}}\left[qw_1 + (1-q)w_2 + x_1\frac{\partial w_1}{\partial p} + x_2\frac{\partial w_2}{\partial p} - p_{t+1}\frac{\partial \overline{w}}{\partial p}\right], \tag{12}$$

$$\frac{\partial p_{t+1}}{\partial q} = \frac{1}{\overline{w}}\left[pw_1 - pw_2 + x_1\frac{\partial w_1}{\partial q} + x_2\frac{\partial w_2}{\partial q} - p_{t+1}\frac{\partial \overline{w}}{\partial q}\right], \tag{13}$$

$$\frac{\partial p_{t+1}}{\partial D} = \frac{1}{\overline{w}}\left[w_1 - w_2 + x_1\epsilon_1 + x_2\epsilon_2 - p_{t+1}\frac{\partial \overline{w}}{\partial D}\right]. \tag{14}$$

As stated above, we are evaluating the first-order derivatives with respect to p, q and D at the equilibrium values. If we substitute the equilibrium values in equations (9)–(11) and use the fact that

$$\frac{\partial \overline{w}}{\partial p} = \sum_{i=1}^{4}\frac{\partial x_i}{\partial p}w_i + \sum_{i=1}^{4}x_i\frac{\partial w_i}{\partial p}, \tag{15}$$

we obtain that, at equilibrium,

$$\frac{\partial \overline{w}}{\partial p} = 0. \tag{16}$$

Similarly we can show that

$$\frac{\partial \overline{w}}{\partial q} = 0, \tag{17}$$

and

$$\frac{\partial \overline{w}}{\partial D} = 4x\epsilon_1 + 2(1-2x)\epsilon_2. \tag{18}$$

In general, therefore,

$$\frac{\partial \overline{w}}{\partial D} \neq 0$$

and thus the equilibrium point is not a maximum or minimum for $\overline{w}$ in the three-dimensional space of p, q and D. In the special case of all double homozygotes having equal fitness (i.e. $\alpha = \delta$), an equilibrium exists at $D = 0$ and at this point $\partial\overline{w}/\partial D = 0$, giving a maximum or minimum point for $\overline{w}$.

On substitution of equation (16) and the equilibrium values of equation (9), equation (12) evaluated at the equilibrium point, is

$$\frac{\partial p_{t+1}}{\partial p} = \frac{1}{\overline{w}}(1 - S - \tfrac{1}{2}\beta + \epsilon_1 x - \epsilon_2(\tfrac{1}{2} - x)), \tag{19}$$

https://doi.org/10.1017/S0016672300018024 Published online by Cambridge University Press

where $\bar{w} = 1-S+2\epsilon_1 x^2-2\epsilon_2(\frac{1}{2}-x)^2$. Similarly, we can get all the elements of the linearization matrix, **A**, as, for example

$$\frac{\partial(p_{t+1}-\hat{p})}{\partial(p-\hat{p})} = \frac{\partial p_{t+1}}{\partial p}$$

Thus we have

$$\mathbf{A} = \frac{1}{\bar{w}}\begin{bmatrix} 1-S-\frac{1}{2}\beta+\epsilon_1 x-\epsilon_2(\frac{1}{2}-x) & -2\gamma\hat{D}+\epsilon_1 x+\epsilon_2(\frac{1}{2}-x) & 0 \\ -2\beta\hat{D}+\epsilon_1 x+\epsilon_2(\frac{1}{2}-x) & 1-S-\frac{1}{2}\gamma+\epsilon_1 x-\epsilon_2(\frac{1}{2}-x) & 0 \\ 0 & 0 & \bar{w}\dfrac{\partial D_{t+1}}{\partial D} \end{bmatrix} \tag{20}$$

where

$$\frac{\partial D_{t+1}}{\partial D} = \frac{1}{\bar{w}^2}[(1-S)^2+2(1-S)(\epsilon_1 x(1-x)-\epsilon_2(\tfrac{1}{4}-x^2))-2\epsilon_1\epsilon_2 x(\tfrac{1}{2}-x) - c(\bar{w}-4\hat{D}(\epsilon_1 x+\epsilon_2(\tfrac{1}{2}-x)))].$$

4. SAMPLING MATRIX

Because selection is involved, we must be careful about the ordering of the events selection, recombination and sampling in the evaluation of the sampling matrix. Felsenstein (1974) used a model of differential viability in that he assumed an infinitely large gametic pool which paired at random to form individuals, which survived differentially. N individuals were then chosen and recombination occurred when these N individuals gave gametes to form the new infinite gametic pool. Thus p, q and D are measured in the infinite gametic pool obtained from a set of individuals in Felsenstein's model. As pointed out by Avery & Hill (1978), while p and q remain unchanged in the production of a gametic pool, D decreases. This model we have called SNR ordering. An alternative model which we have used because it lends itself more easily to simulation and experimental verification, is that of differential fertility. Here $2N$ gametes combine randomly to form individuals. The individuals give gametes according to their fertilities to an infinite gametic pool, recombination taking place. $2N$ gametes are then sampled to form the new generation. Thus here p, q and D are measured amongst the chromosomes of a particular generation. This we shall call SRN ordering. Because in SRN ordering, gametic frequencies are measured amongst a finite number of chromosomes, the deterministic equations for changes in gametic frequencies are not exactly correct due to departures from Hardy–Weinberg proportions when gametes pair. However as N is fairly large, these departures are of negligible importance. In SNR ordering, the deterministic equations are exact.

(i) *SRN ordering*

After selection and recombination, we have gametic frequencies, x_i' say ($\equiv x_{i,t+1}$) given by equation (5). Sampling of $2N$ gametes from the gametic pool with frequencies, x_i', is a case of multinomial sampling. Thus if x_i'' are the final gametic frequencies,

$$\text{cov}\,(x_i'', x_j'') = -\frac{x_i' x_j'}{2N} + \delta_{ij}\frac{x_i'}{2N} = \text{cov}\,(\delta x_i, \partial x_j), \tag{21}$$

https://doi.org/10.1017/S0016672300018024 Published online by Cambridge University Press

where $\partial_{ij} = 1$ $(i = j)$ and zero otherwise, and δx_i is the change in x_i due to random sampling. Using equations (21) and (5), and a similar assumption to one made by Felsenstein (1974), i.e.

$$\delta D = x'_1\delta x_4 + x'_4\delta x_1 - x'_2\delta x_3 - x'_3\delta x_2, \tag{22}$$

$\mathbf{Q} = \text{var}(\delta\mathbf{d})$

$$= \begin{bmatrix} \frac{1}{8N} & \frac{1}{2N\bar{w}}[(1-S-c)\hat{D} + \frac{1}{2}(\epsilon_1 x^2 + \epsilon_2(\frac{1}{2}-x)^2)] & 0 \\ \frac{1}{2N\bar{w}}[(1-S-c)\hat{D} + \frac{1}{2}(\epsilon_1 x^2 + \epsilon_2(\frac{1}{2}-x)^2)] & \frac{1}{8N} & 0 \\ 0 & 0 & \frac{1}{2N}\left[\frac{1}{16} - \frac{1}{\bar{w}^2}(\frac{1}{2}(xw_1 - (\frac{1}{2}-x)w_2) - c\hat{D})^2\right] \end{bmatrix} \tag{23}$$

where $w_1 = 1-S+\epsilon_1 x$ and $w_2 = 1-S-\epsilon_2(\frac{1}{2}-x)$.

(ii) *SNR ordering*

Here care must be taken because N individuals are selected rather than $2N$ gametes. Let P_{ij} be the frequency after selection of the genotype whose gametes had frequencies x_i and x_j in the unselected population, and let P'_{ij} be the frequency of the same genotype after selection and sampling. Then,

$$P_{ij} = x_i x_j w_{ij}/\bar{w} \quad (i, j = 1, 2, 3, 4). \tag{24}$$

Let x'_i, x''_i, x'''_i $(i = 1, 2, 3, 4)$ be respectively the gametic frequencies after selection, after selection and sampling, and after selection, sampling and recombination. Then, as shown by Felsenstein (1974) using the theory of multinomial sampling, but this time with sample size N rather than $2N$ as with SRN ordering,

$$\text{cov}(P'_{ij}, P'_{km}) = -P_{ij}P_{km}/N \quad (i, j \neq k, m), \tag{25}$$

$$\text{var}(P'_{ij}) = P_{ij}(1-P_{ij})/N, \tag{26}$$

$$\text{cov}(P'_{14}, x''_1) = -\frac{1}{N}P_{14}x'_1 + \frac{1}{2N}P_{14}, \tag{27}$$

$$\text{cov}(P'_{23}, x''_1) = -\frac{1}{N}P_{23}x'_1, \tag{28}$$

$$\text{cov}(x''_i, x''_j) = \frac{1}{2N}P_{ij} - \frac{1}{N}x'_i x'_j + \delta_{ij}\frac{x'_i}{2N}, \tag{29}$$

$$x'''_i = x''_i - \frac{ck_i}{2}(P'_{14}+P'_{41}) + \frac{ck_i}{2}(P'_{23}+P'_{32}), \tag{30}$$

where, for example, $x'_1 = P_{11} + \frac{1}{2}(P_{12}+P_{21}) + \frac{1}{2}(P_{13}+P_{31}) + \frac{1}{2}(P_{14}+P_{41})$.

https://doi.org/10.1017/S0016672300018024 Published online by Cambridge University Press

Using these above results it is then straightforward to show that

$$\operatorname{cov}(\delta x_i, \delta x_j) = \frac{1}{2N} P_{ij} - \frac{1}{N} x_i' x_j' + \frac{1}{2N} x_i' \delta_{ij} - \frac{c}{N} [a_1 P_{14} + a_2 P_{23} + (P_{23} - P_{14})(k_i x_j' + k_j x_i')] + k_i k_j \frac{c^2}{N} [\tfrac{1}{2}(P_{14} + P_{23}) - (P_{14} - P_{23})^2] \tag{31}$$

where $a_1 = \frac{1}{4}(k_i + k_j + 2k_i k_j)$ and $a_2 = \frac{1}{4}(2k_i k_j - k_i - k_j)$.

Using (31) and (22), we can now derive the elements of **Q** for SNR ordering, i.e.

$$\mathbf{Q} = \frac{1}{8N\bar{w}} \begin{bmatrix} \bar{w} + \frac{1}{4}(1 - 16\hat{D}^2)(\gamma - \beta) - 2(\delta x^2 + \alpha(\tfrac{1}{2} - x)^2) & 4(\hat{D}(1-2S) + \epsilon_1 x^2 + \epsilon_2(\tfrac{1}{2} - x)^2) & 0 \\ 4(\hat{D}(1-2S) + \epsilon_1 x^2 + \epsilon_2(\tfrac{1}{2} - x)^2) & \bar{w} + \frac{1}{4}(1 - 16\hat{D}^2)(\beta - \gamma) - 2(\delta x^2 + \alpha(\tfrac{1}{2} - x)^2) & 0 \\ 0 & 0 & 8N\bar{w} \operatorname{var}(\delta D) \end{bmatrix} \tag{32}$$

where

$$\operatorname{var}(\delta D) = \frac{1}{2N\bar{w}} [(\hat{D} + \bar{D})^2 (1 - S) + 2\epsilon_1 (x')^2 x^2 - 2\epsilon_2 (\tfrac{1}{2} - x)^2 (\tfrac{1}{2} - x')^2] + \frac{1}{N} \left[\frac{1}{32} - \frac{5\bar{D}^2}{2}\right] + \frac{c^2}{N\bar{w}} \left[\tfrac{1}{16} + \hat{D}^2 \left(1 - \frac{1}{\bar{w}}\right)\right] - \frac{c}{N\bar{w}} [\tfrac{1}{16} + \hat{D}(\hat{D} - 2\bar{D})],$$

$$x' = x w_1 / \bar{w} \quad \text{and} \quad \bar{D} = -\tfrac{1}{4} + x'.$$

If we again take the simple model of all double homozygotes being of equal fitness and consider a stable equilibrium at $\hat{D} = 0$ (i.e. no linkage disequilibrium) as was done by Felsenstein (1974), we find that

$$\operatorname{var}(\delta D) = \frac{1}{32N} \left[1 - \frac{2c}{\bar{w}} (1 - c) + \frac{(2S - \alpha)}{4\bar{w}}\right]. \tag{33}$$

This is in contradiction to that given by Felsenstein (1974). His expression ignores the terms in c and c^2.

5. EVALUATION OF FORMULAE

Thus **A** and **Q** are block diagonal, and **C**, the variance–covariance matrix of the departures from equilibrium, has a similar form. Because of this

$$\operatorname{var}_E(D) = \frac{\operatorname{var}(\delta D)}{1 - a_{33}^2}, \tag{34}$$

$$\operatorname{cov}_E(p, D) = \operatorname{cov}_E(q, D) = 0, \tag{35}$$

and

$$\begin{pmatrix} \operatorname{var}_E(p) \\ \operatorname{cov}_E(p, q) \\ \operatorname{var}_E(q) \end{pmatrix} = \begin{pmatrix} 1 - a_{11}^2 & -2a_{11}a_{12} & -a_{12}^2 \\ -a_{11}a_{21} & 1 - a_{12}a_{21} - a_{11}a_{22} & -a_{12}a_{22} \\ -a_{21}^2 & -2a_{21}a_{22} & 1 - a_{22}^2 \end{pmatrix}^{-1} \begin{pmatrix} \operatorname{var}(\delta p) \\ \operatorname{cov}(\delta p, \delta q) \\ \operatorname{var}(\delta q) \end{pmatrix}. \tag{36}$$

The suffix, E, is used to denote that these are variances about a single stable equilibrium point. a_{ij} denotes the (i, j)th element of **A**. In the simple case of all

https://doi.org/10.1017/S0016672300018024 Published online by Cambridge University Press

double homozygotes having equal fitness and all single homozygotes having equal fitness, $\text{var}_E(p) = \text{var}_E(q)$ and simpler equations are possible.

We thus can easily evaluate the variance–covariance matrix around a stable equilibrium if we know N, c and x. In order to exemplify the use of the equations obtained above we shall consider the completely symmetric model (cf. Lewontin & Kojima, 1960) where all double homozygotes have fitness $1-\alpha$ and all single heterozygotes have fitness $1-S$. Trends are more easily observed in this case and the values, stability and zone of attraction of the equilibrium points are well known. For this symmetric model we either have an equilibrium with no linkage disequilibrium (i.e. $\hat{D} = 0$ and $x = 0{\cdot}25$) which is stable if $c > \frac{1}{4}(2S-\alpha)$, or we have two equilibria at $D = \pm\frac{1}{4}\sqrt{(1-4c/(2S-\alpha))}$ which are stable if $c < \frac{1}{4}(2S-\alpha)$. The behaviour for the former case is straightforward as $E(D^2) = \text{var}_E(D)$ is given by the above expressions. However, when there are two stable equilibria there is a bimodal distribution, the two modes being at the two equilibrium points. If all populations are initially at $D = \frac{1}{4}\sqrt{(1-4c/(2S-\alpha))}$, most populations will remain close to this value of D. By chance, however, the disequilibrium in a population may become negative and then D will tend to move towards $D = -\frac{1}{4}\sqrt{(1-4c/(2S-\alpha))}$. When there is a bimodal distribution the variance of D will depend on the initial value of D in the populations. However, $E(D^2) \simeq \hat{D}^2 + \text{var}_E(D)$ regardless of which equilibrium point a replicate is initially nearest to as long as the assumption holds that the two distributions around each equilibrium are reasonably separate. From the above information we can derive approximations for σ_D^2 (Ohta & Kimura (1969)) where $\sigma_D^2 = E(D^2)/E(p(1-p)q(1-q))$. We have been considering equilibria at $\hat{p} = \hat{q} = \frac{1}{2}$ for reasonably large N. Under these assumptions, $E(p(1-p)q(1-q)) \simeq \hat{p}(1-\hat{p})\hat{q}(1-\hat{q}) = \frac{1}{16}$. Thus $\sigma_D^2 \simeq 16E(D^2)$. We shall mainly look at σ_D^2 in our consideration of the effects of different selection schemes. When c is close to $(2S-\alpha)/4$, i.e. the deterministic system is going from one stable equilibrium to two stable equilibria, the assumptions of independence of the two distributions and normality of the sampling distributions no longer hold and thus the approximation breaks down. However, the range of c values for which the approximation is poor is relatively small and decreases with increase in N. This is shown in Figs. 1–3, where σ_D^2 is plotted against c. The dotted vertical lines mark the value of c when the system changes from having one stable equilibrium to two stable equilibria. Figs. 1 and 2 are for a symmetric multiplicative model. In this case, all single homozygotes have fitness $1-S$, all double homozygotes have fitness $(1-S)^2$ and the limit of stability is at $c = S^2/4$. Fig. 3 is for an epistatic model, i.e. all genotypes except the double heterozygotes have fitness $1-S$, and the stability limit is at $c = S/4$. σ_D^2 is plotted against c for only a limited range of values of c in Figs. 1–3. In the range of c values not given, the approximation works very well. Also on the figure we have marked simulation results to show how the approximations deviate from them. The evaluation of these simulation results will be discussed in the following section. As c is increased from zero for fixed N, S and α we find that $\text{var}_E(D)$ increases to a maximum and then declines (using simulation results in the range of instability

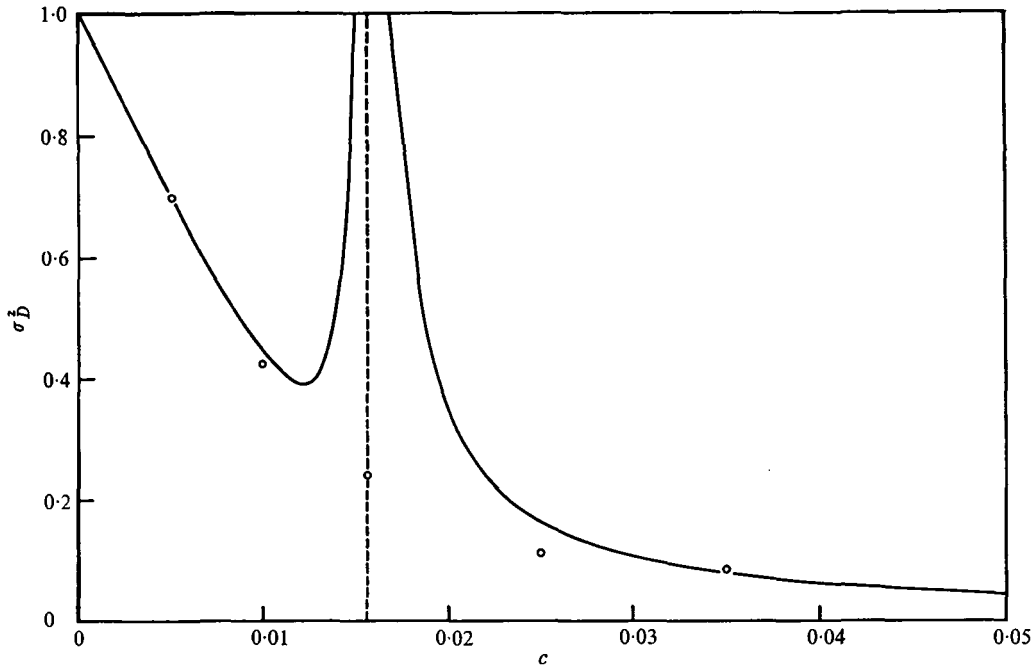

Fig. 1. Predicted values of σ_D^2 are plotted against c for a multiplicative model with $N = 128$, $S = 0{\cdot}25$. Circles are simulation results. $\alpha = 2S - S^2$.

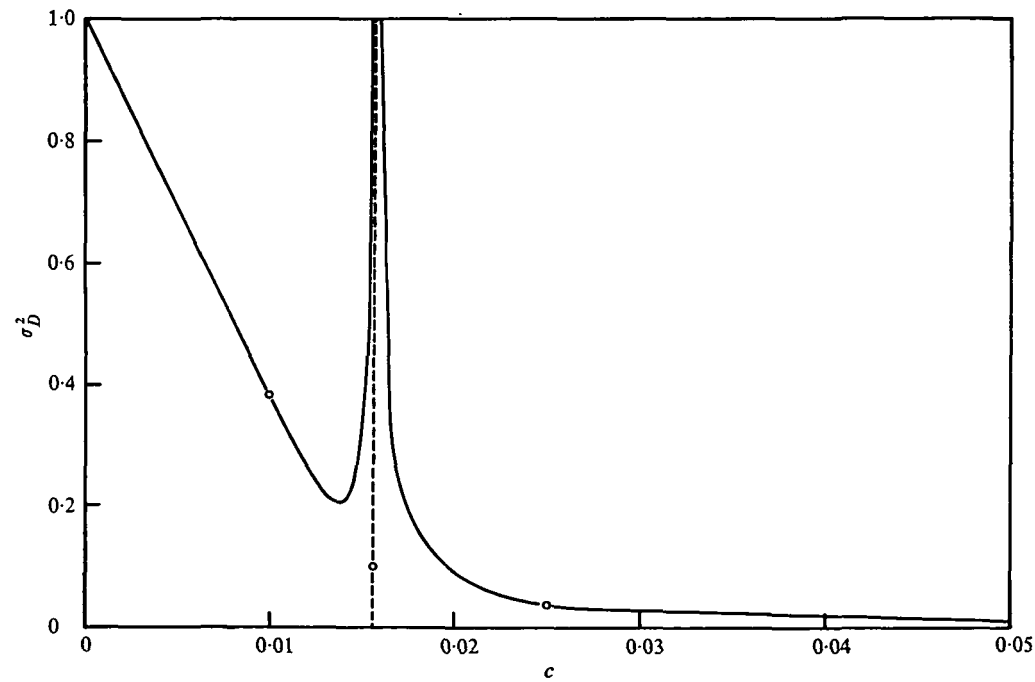

Fig. 2. Predicted values of σ_D^2 are plotted against c for a multiplicative model with $N = 512$, $S = 0{\cdot}25$. Circles are simulation results. $\alpha = 2S - S^2$.

https://doi.org/10.1017/S0016672300018024 Published online by Cambridge University Press

of the approximations), $\text{var}_E(p) = \text{var}_E(q)$ increases to a maximum at $\frac{1}{4}(2S-\alpha)$ and then remains constant while cov (p, q) decreases to zero at $\frac{1}{4}(2S-\alpha)$ and then remains zero.

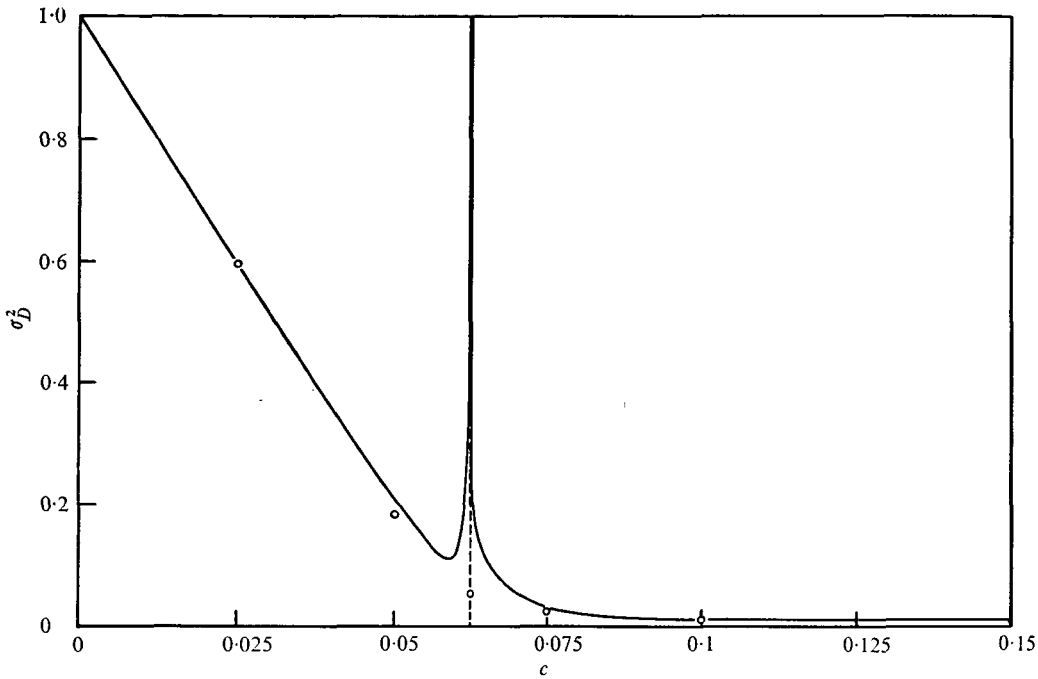

Fig. 3. Predicted values of σ_D^2 are plotted against c for an epistatic model with $N = 512$, $S = 0{\cdot}25$. Circles are simulation results. $\alpha = S$.

6. SIMULATION CHECKS

So far in this paper, we have derived approximations for $\text{var}_E(p)$, $\text{cov}_E(p, q)$, $\text{var}_E(q)$, $\text{var}_E(D)$ and σ_D^2 about stable equilibrium points. In order to validate these approximations, two simulation programs were produced. One was merely a direct simulation of the process using SRN ordering where $2N$ gametes are produced each generation. The second program was an adaptation of an approach suggested by Pederson (1973). Gamete frequencies were changed by selection and recombination (i.e. equation (5)) and then the new values were formed by use of multi-normal sampling. The multi-normal distribution is an approximation to the actual multinomial sampling which takes place in the direct simulation. In the simulation, however, unlike the approximation, the variances and covariances of the multi-normal distributions change with changes in gametic frequencies. For full details, Pederson (1973) should be consulted. As recommended by Pederson (1973), binomial and Poisson distributions were used when the number to be sampled was small or when the probability of a particular gamete became small. The two simulations agreed very well over all values of N and c for σ_D^2, but the second method was far quicker and cheaper on computer time, particularly

for large values of N. Replication was taken high enough to give convergence in σ_D^2 to 2–3 significant figures.

In Tables 1 and 2 predicted and simulation values for σ_D^2 for SRN ordering (i.e. sampling gametes) are given for an additive and a multiplicative model.

Table 1. *Predicted and simulation values for σ_D^2 for an additive selection model ($\alpha = 2S$), and values of the standard deviation of r ($= \sqrt{\sigma_D^2}$), obtained from the predicted SRN values ($S = 0{\cdot}25$)*

		σ_D^2			
N	c	Predicted (SNR)	Predicted (SRN)	Simulation (SRN)	S.D. (r)
32	0·01	0·574	0·590	0·508	0·768
	0·015625	0·363	0·379	0·391	0·616
	0·025	0·223	0·238	0·242	0·488
64	0·01	0·287	0·295	0·299	0·543
	0·025	0·111	0·119	0·126	0·345
	0·05	0·053	0·061	0·063	0·251
	0·1	0·024	0·031	0·034	0·176
	0·2	0·010	0·017	0·017	0·130
	0·5	0·003	0·009	0·009	0·095
128	0·01	0·144	0·148	0·138	0·385
	0·025	0·056	0·060	0·058	0·245
512	0·01	0·036	0·037	0·037	0·192
	0·025	0·014	0·015	0·015	0·122

Table 2. *Predicted and simulation values of σ_D^2 and the deterministic part of σ_D^2 (i.e. $16\hat{D}^2$) for a multiplicative model ($\alpha = 2S - S^2$): $S = 0{\cdot}25$*

			σ_D^2	
N	c	$16\hat{D}^2$	Predicted (SRN)	Simulation (SRN)
32	0·01	0·36	0·705	0·621
	0·025	0	0·642	0·358
64	0·01	0·36	0·533	0·486
	0·025	0	0·321	0·203
	0·05	0	0·089	0·086
	0·1	0	0·038	0·036
	0·2	0	0·018	0·020
	0·5	0	0·009	0·010
128	0·01	0·36	0·446	0·425
	0·025	0	0·161	0·115
512	0·01	0·36	0·382	0·382
	0·025	0	0·040	0·040

For the additive model, all double homozygotes have fitness $1-2S$ as compared with $1-S$ for single homozygotes and there is only one stable polymorphic equilibrium at $\hat{D} = 0$. The approximations in this case break down if c gets very close to zero. As can be seen from Table 1, the fit is reasonably good for all

c and N for the additive model. For the multiplicative model the error is greatest for small N and when c is near the boundary of the stability range. Also in Table 1 we have tabulated the predicted values of σ_D^2 under SNR ordering. For the most part the ordering makes no difference at all. When c is large the difference is quite marked but by then the values of σ_D^2 have got very small.

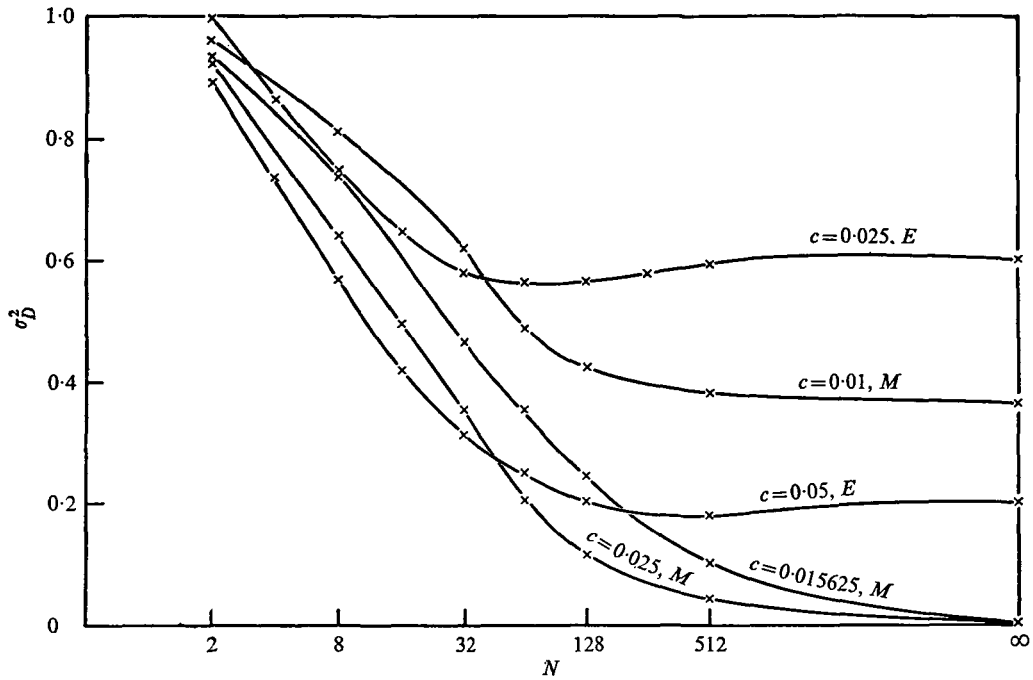

Fig. 4. Simulation results for σ_D^2 are plotted against N for fixed c and given model (E = epistatic, M = multiplicative). $S = 0{\cdot}25$.

In Fig. 4 the results of simulation runs for various values of N are given for three multiplicative models and two epistatic models. The three multiplicative models have only slightly different c values ($c = 0{\cdot}015625 = S^2/4$ is the boundary of stability of $\hat{D} = 0$) but are quite different for all N even though two of them have $\hat{D} = 0$ when $N \to \infty$. The two epistatic cases show the rather surprising results that σ_D^2 falls below its asymptotic value as N increases and then increases again. This appears to be due to a decrease in the mean of $|D|$ caused by the non-normality of the distribution and cannot be predicted theoretically. Thus in the epistatic model when two stable equilibria exist we always overpredict as the approximation gives a monotonically decreasing curve asymptoting at the deterministic value as shown in Table 3.

7. DISCUSSION

From the previous sections we can see that unless selection is very intense and N exceptionally large, the deterministic solutions are not very useful. The problem of having pairs of stable disequilibria means that two replicate lines can have

very different amounts of linkage disequilibrium even if the selection scheme is the same in both. This is because a replicate line, which is originally near one equilibrium point, may, by chance, become closer to the other equilibrium point and thus tend to move nearer to it. Lines moving from around one equilibrium to near the other is more common when the equilibrium values of D are not large. If there is strong epistatic selection and tight linkage, the chance of wandering from one equilibrium to the other will be very small. This is shown by Table 3,

Table 3. *Predicted and simulation values of* σ_D^2 *for an epistatic model* ($\alpha = S$): $S = 0{\cdot}25$

	$c = 0{\cdot}025$		$c = 0{\cdot}05$	
N	Predicted (SRN)	Simulation (SRN)	Predicted (SRN)	Simulation (SRN)
8	0·748	0·751	> 1	0·566
16	0·674	0·646	0·619	0·417
32	0·637	0·578	0·409	0·313
64	0·619	0·563	0·305	0·258
128	0·609	0·566	0·252	0·203
512	0·602	0·594	0·213	0·178
$\to \infty$	0·600	0·600	0·200	0·200

where the contribution to σ_D^2 of variation about equilibria is small. The values for $N \to \infty$ give the deterministic contribution to σ_D^2. When selection is multiplicative, the variance about an equilibrium point contributes quite a large part of σ_D^2, as can be seen from Table 2 where $16\hat{D}^2$, the deterministic contribution to σ_D^2, is tabulated. As the equilibrium values of D, i.e. $\hat{D}$, approach zero, the two distributions begin to overlap greatly and the normal approximation, used for the predicted values, breaks down.

In producing the above theoretical results, a symmetric viability model was used. Theoretically, any model can be analysed in this way as long as the stable equilibria are known. Karlin & Carmelli (1975) have shown that when viabilities are general, numbers of polymorphic equilibria and their stability are difficult to predict. Also when there are two stable equilibria, they will not in general be symmetric about $D = 0$ and thus the zones of attraction will not easily be defined and can only be obtained by computer simulation in the three-dimensional system of p, q and D. Thus generalization of formulae would present considerable difficulties.

In the preceding section, rather unwieldly expressions have been produced. By making certain assumptions, some more tractable expressions can be obtained. Let us take the simple case of all double homozygotes having equal fitness, $1-\alpha$, and assume that α, S and c are such that there is one stable equilibrium with linkage equilibrium (i.e. $\hat{p} = \hat{q} = \frac{1}{2}$ and $\hat{D} = 0$). Then for SRN sampling,

$$\sigma_D^2 = 1/[2N(1-(1+(\tfrac{1}{4}(2S-\alpha)-c)/\bar{w})^2)], \tag{37}$$

and for SNR sampling,

$$\sigma_D^2 = \frac{1-2c(1-c)/\overline{w}+\frac{1}{4}(2S-\alpha)/\overline{w}}{2N(1-(1+(\frac{1}{4}(2S-\alpha)-c)/\overline{w})^2)}, \tag{38}$$

where $\overline{w} = 1-\frac{1}{2}S-\frac{1}{4}\alpha$ in both cases. The difference between the two formulations can now be seen to increase with c. The reason for the difference, as mentioned previously, is that in equation (37), linkage disequilibrium is measured amongst the gametes of individuals of a generation and in equation (38), the linkage disequilibrium is measured in the conceptual gametic pool formed from the individuals of a generation. While gene frequencies do not change in the formation of a gametic pool, the moments of the linkage disequilibrium, D, do.

Table 4. *Predicted values of σ_D^2 for different models and values of S and c*

	$N = 64$				
	Additive		Multiplicative		Neutral
c	$S = 0{\cdot}25$	$S = 0{\cdot}1$	$S = 0{\cdot}25$	$S = 0{\cdot}1$	$S = 0$
0·01	0·295	0·354	0·533†	0·472	0·393
0·025	0·119	0·143	0·203‡	0·159	0·158
0·05	0·061	0·072	0·089	0·076	0·080
0·1	0·031	0·037	0·038	0·038	0·041
0·2	0·017	0·020	0·018	0·020	0·022
0·5	0·009	0·010	0·009	0·010	0·010

† $D \neq 0$.
‡ Simulated value used as prediction breaks down at these parameter values.

When there is no selection or selective parameters are small, then equation (37) simplifies to $\sigma_D^2 = 1/(2Nc(2-c))$ $(= 2/(3N)$ when $c = \frac{1}{2})$ which is a well-quoted result for σ_D^2 in a finite population (e.g. Bulmer, 1976), while equation (38) gives $\sigma_D^2 = ((1-c)^2+c^2)/(2Nc(2-c))$ $(= 1/(3N)$ when $c = \frac{1}{2})$, which is the same as that given in Avery & Hill (1978) for the value of σ_D^2 in a conceptually infinite offspring generation of a finite population, which is equivalent to the infinite gametic pool of SNR ordering. In Avery & Hill (1978) it was also shown that if σ_D^2 was measured in a finite population of size N and departure from Hardy–Weinberg proportions was taken into account, $\sigma_D^2 = (1+c^2)/(2Nc(2-c))$ $(= 5/(6N)$ when $c = \frac{1}{2})$. The departure from Hardy–Weinberg proportions is only of very minor importance unless c is large and thus was not considered in this analysis. Its inclusion would considerably complicate the expressions used.

Calculation of either higher moments of D or its exact distribution does not at present seem tractable. However, an impression of the magnitude of the dispersion around equilibrium values caused by finiteness can be obtained by looking at the standard deviation of r $(= D/\sqrt{(p(1-p)\,q(1-q))})$, the correlation of gene frequencies. As p and q vary little from $\hat{p}$ and $\hat{q}$, the variance of r about an equilibrium point is approximately equal to $\mathrm{var}_E(D)/(\hat{p}(1-\hat{p})\hat{q}(1-\hat{q}))$, which equals 16 $\mathrm{var}_E(D)$ when $\hat{p} = \hat{q} = \frac{1}{2}$ as we have taken in our examples.

r will have a mean value of approximately $\hat{D}/\sqrt{(\hat{p}(1-\hat{p})\hat{q}(1-\hat{q}))}$ $(= 4\hat{D}$ when $\hat{p} = \hat{q} = \frac{1}{2})$ and, being a correlation, lies in the range $[-1, 1]$. Values of the standard deviation of r as calculated from the predicted formulae are given for the additive model in Table 1. If any value of r was equally probable (i.e. r had a uniform distribution on $[-1, 1]$) then its standard deviation would be $1/\sqrt{3}$, i.e. 0·577. Thus it can be seen that, even with large N, the approximate normal distribution must have a wide dispersion unless linkage is very loose. In Tables 1–3, large selection pressures (i.e. $S = 0{\cdot}25$) have been used. By doing so, fixation due to finiteness is delayed almost indefinitely. However, selection in itself has little effect on σ_D^2 and the standard deviation of r when $\hat{D} = 0$. To exemplify this, Table 4 gives values of σ_D^2 for various values of c and S for an additive and a multiplicative model. The values for $S = 0$ have less direct meaning as fixation occurs relatively quickly and a steady distribution of p, q and D does not exist.

Thus care must be taken before using deterministic results when a population is, or has recently been, finite. Also we must be cautious about declaring that different populations have different selection regimes because their linkage disequilibria are different, or in declaring that different population have the same selection regime because their linkage disequilibria are equal.

The author is indebted to Dr W. G. Hill for his comments upon this work as well as on the paper itself and to Dr J. Felsenstein for his comments on the first draft of the paper. The author is grateful to the Science Research Council for financial support.

REFERENCES

Avery, P. J. & Hill, W. G. (1978). Variance in quantitative traits due to linked dominant genes and variance in heterozygosity in small populations. (Submitted to *Genetics*.)

Bodmer, W. F. & Felsenstein, J. (1967). Linkage and selection: theoretical analysis of the deterministic two locus random mating model. *Genetics* **57**, 237–265.

Bulmer, M. G. (1976). The effect of selection on genetic variability: a simulation study. *Genetical Research* **28**, 101–117.

Felsenstein, J. (1974). Uncorrelated genetic drift of gene frequencies and linkage disequilibrium in some models of linked overdominant polymorphisms. *Genetical Research* **24**, 281–294.

Franklin, I. & Lewontin, R. C. (1970). Is the gene the unit of selection? *Genetics* **65**, 707–734.

Hill, W. G. (1969). Maintenance of segregation at linked genes in finite populations. Proceedings of the XIIth International Congress of Genetics, vol. III. *Japanese Journal of Genetics* **44**, Supplement 1, 144–151.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**, 226–231.

Karlin, S. (1975). General two-locus selection models: some objectives, results and interpretations. *Theoretical Population Biology* **7**, 364–398.

Karlin, S. & Carmelli, D. (1975). Numerical studies on two loci selection models with general viabilities. *Theoretical Population Biology* **7**, 399–421.

Levin, B. R. (1969). Simulation of genetic systems. In *Computer Applications in Genetics* (ed. N. Morton), pp. 28–46. Honolulu: University of Hawaii Press.

Lewontin, R. C. & Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 458–472.

Ohta, T. & Kimura, M. (1969). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63**, 229–238.

PEDERSON, D. G. (1973). An approximate method of sampling a multinomial population. *Biometrics* **29**, 814–821.

SVED, J. (1968). The stability of linked systems of loci with a small population size. *Genetics* **59**, 543–563.

YAMAZAKI, T. (1977). The effects of overdominance on linkage in a multilocus system. *Genetics* **86**, 227–236.