Genet. Res., Camb. (1974), **24**, *pp.* 281–294
Printed in Great Britain

Uncorrelated genetic drift of gene frequencies and linkage disequilibrium in some models of linked overdominant polymorphisms

By JOSEPH FELSENSTEIN
Department of Genetics, University of Washington, Seattle, Washington 98195

(*Received* 15 *April* 1974)

SUMMARY

For large population sizes, gene frequencies p and q at two linked overdominant loci and the linkage disequilibrium parameter D will remain close to their equilibrium values. We can treat selection and recombination as approximately linear forces on p, q and D, and we can treat genetic drift as a multivariate normal perturbation with constant variance-covariance matrix. For the additive–multiplicative family of two-locus models, p, q and D are shown to be (approximately) uncorrelated. Expressions for their variances are obtained. When selection coefficients are small the variances of p and q are those previously given by Robertson for a single locus. For small recombination fractions the variance of D is that obtained for neutral loci by Ohta & Kimura. For larger recombination fractions the result differs from theirs, so that for unlinked loci $r^2 \simeq 2/(3N)$ instead of $1/(2N)$. For the Lewontin–Kojima and Bodmer symmetric viability models, and for a model symmetric at only one of the loci, a more exact argument is possible. In the asymptotic conditional distribution in these cases, various of p, q and D are uncorrelated, depending on the type of symmetry in the model.

1. INTRODUCTION

Much of the work on linked genes in the last few years has centred on the deterministic theory of natural selection of linked polymorphisms. Of the papers which treat genetic drift of linked genes, most assume no selection. The papers of Sved (1968, 1972) and of Ohta & Kimura (1971; see also Kimura & Ohta, 1971) dealt with the effect of a single locus under selection on a neighbouring neutral locus. Sved (1968), Hill (1968, 1969), Levin (1969) and Hill & Robertson (1968) treated the case of two or more linked polymorphic loci in a finite population. Of these, Levin (1969) presented some computer simulations of the rate of fixation of linked overdominant loci. Hill (1968, 1969) used exact matrix methods to find the rate of fixation in very small populations with two linked overdominant loci. For larger populations he used computer simulation. Hill & Robertson (1968) also examined the rate of fixation of linked overdominant loci using computer simulation. In all of these cases, associative overdominance was found to mutually retard the fixation of tightly linked overdominant loci. Of these authors, only Sved (1968) and Hill & Robertson (1968) examined the characteristics of the unfixed populations. Both

found that the measure r^2 of linkage disequilibrium would approach $1/(4Nc)$, where N is the population size and c the recombination fraction.

This paper is concerned with the properties of the doubly unfixed populations when there are linked overdominant polymorphisms in a finite population. Attention is restricted to a few of the many possible models of selection. The variables of interest are the deviations of the gene frequencies from their deterministic equilibria, and the linkage disequilibrium. For the additive–multiplicative family of models, an approximation is constructed by linearizing the processes of selection and recombination. For various symmetric selection models, an exact argument is possible.

Table 1. *Notation for genotype fitnesses*

	BB	*Bb*	*bb*
AA	w_{11}	w_{12}	w_{22}
Aa	w_{13}	$w_{14} = w_{23}$	w_{24}
aa	w_{33}	w_{34}	w_{44}

2. THE MODEL

We consider a diploid population of constant number N with discrete generations. Reproduction is according to a monoecious 'Wright model' with selfing allowed. We consider two loci, each with two alleles (A and a, B and b). The fitnesses of the nine genotypes, assuming no cis–trans effects, are given in Table 1. Technically, we have a different stochastic process, depending on whether these fitnesses represent differences in viability, differences in fertility, or some mixture. The frequencies of the gametes AB, Ab, aB and ab will be denoted by x_1, x_2, x_3 and x_4. The gene frequency of A is p and of B is q (note that q does *not* stand for $1-p$). D is the linkage disequilibrium (or gametic phase unbalance, or gametic phase imbalance, or gametic phase disequilibrium) parameter.

We have the usual relationships between gamete frequencies, gene frequencies, and linkage disequilibrium:

$$\left.\begin{aligned} x_1 &= pq + D, \\ x_2 &= p(1-q) - D, \\ x_3 &= (1-p)q - D, \\ x_4 &= (1-p)(1-q) + D \end{aligned}\right\} \tag{1}$$

and

$$\left.\begin{aligned} p &= x_1 + x_2, \\ q &= x_1 + x_2, \\ D &= x_1 - pq = x_1x_4 - x_2x_3. \end{aligned}\right\} \tag{2}$$

We have the usual marginal gamete fitnesses:

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

and the average fitness:

$$\bar{w} = \sum_{j=1}^{4} x_j w_j = \sum_i \sum_j x_i x_j w_{ij}. \tag{4}$$

https://doi.org/10.1017/S0016672300015287 Published online by Cambridge University Press

We now make use of a quantity which is convenient for handling signs:

$$\left.\begin{aligned} k_i &= 1 \quad \text{if} \quad i = 1, 4 \\ &= -1 \quad \text{if} \quad i = 2, 3. \end{aligned}\right\} \tag{5}$$

The deterministic equations for the change in the x_i by natural selection are:

$$x_i' = [x_i w_i - r k_i w_{14} D]/\bar{w}, \tag{6}$$

where r is the recombination fraction, and the prime indicates the new value of x_i.

Using equations (2) we readily obtain recurrence relations for p, q and D:

$$\left.\begin{aligned} p' &= [x_1 w_1 + x_2 w_2]/\bar{w}, \\ q' &= [x_1 w_1 + x_3 w_3]/\bar{w}, \\ D' &= x_1' x_4' - x_2' x_3' \\ &= [x_1 x_4 w_1 w_4 - x_2 x_3 w_2 w_3 - r w_{14} D\bar{w}]/\bar{w}^2. \end{aligned}\right\} \tag{7}$$

The x_i, w_i and $\bar{w}$ in equations (7) are to be regarded as functions of p, q and D, as given by equations (2) and implied by equations (3) and (4).

Genetic drift is introduced into the model by assuming that the infinite number of survivors of the viability selection are to be reduced to a finite number of adults by multinomial sampling of N individuals. The survivors of viability selection will not in general be in Hardy–Weinberg proportions, in which genotype frequencies are the products of the corresponding gamete frequencies. Therefore sampling N adults is not equivalent to sampling $2N$ haploid genomes. (Of course, the difference will be small if selection coefficients are not large, since then there will be little departure from Hardy–Weinberg proportions after selection.)

All populations which initially have both loci segregating will ultimately drift to fixation at both loci, except for the trivial case in which both homozygotes at a locus are lethal. However, if selection coefficients s or population sizes are large enough that $Ns \gg 1$, fixation will be long delayed. During this interim period we assume that the populations will settle into a stable distribution of completely unfixed populations in the variables p, q and D. It is some characteristics of this distribution which will interest us, in particular the variances and covariances of the variables p, q and D under genetic drift in the presence of overdominance at two loci.

The length of the time until fixation makes the distribution of unfixed populations relevant to equilibrium situations in natural populations. A small rate of mutation to reintroduce the lost alleles into the fixed populations will create an equilibrium distribution of p, q and D. For sufficiently small rates of mutation, the distribution of unfixed populations should be nearly the same as the distribution of unfixed populations without mutation.

3. APPROXIMATE LINEARIZATION OF THE STOCHASTIC PROCESS

For each set of fitnesses there will be various equilibrium values of p, q and D in an infinite population. In a finite population of sufficient size, p, q and D will tend to

remain close to these equilibrium values. Let us define a new set of variables, the d_i, which are the deviations of p, q and D from their deterministic equilibrium values. Thus the row vector

$$\mathbf{d}' = (p - \hat{p}, q - \hat{q}, D - \hat{D}).$$

We can represent the stochastic process as

$$\mathbf{d}^{(t+1)} = f(\mathbf{d}^{(t)}) + \mathbf{e}, \tag{8}$$

where $f(\ )$ represents the deterministic change and $\mathbf{e}$ the stochastic change resulting from sampling of N adults. Now we consider cases in which N is large. Then p, q and D will remain close to their equilibrium values. In that region we can approximate $f(\mathbf{d})$ by a linear transformation of $\mathbf{d}$ (the first terms in a multivariate Taylor series expansion). So

$$\mathbf{d}^{(t+1)} = \mathbf{A}\mathbf{d}^{(t)} + \mathbf{e}. \tag{9}$$

The matrix $\mathbf{A}$ is precisely the Jacobian matrix of the transformation $f(\mathbf{d})$. We approximate the sampling process which provides $\mathbf{e}$ by assuming it follows a multivariate normal distribution with mean 0 and covariance matrix $\mathbf{Q}$. In the actual process the means are zero but the covariance matrix of $\mathbf{e}$ is a function of the current position $\mathbf{d}$. We have approximated it by its value at the equilibrium point.

We are thus approximating the stochastic process by a multivariate random walk with linear return to the equilibrium. This approximation has been used in different contexts by Bodmer (1960) and Smith (1969). Feller (1951) and Karlin & McGregor (1964) found that the continuous-time diffusion process with linear pressure to return to an equilibrium (the Ornstein–Uhlenbeck process well known in physics) is a diffusion approximation to single-locus Wright models when they let $N \to \infty$ while holding mutation constant and making an appropriate change in the scale of gene frequency observations. Norman (1972) also presents conditions sufficient to guarantee the normality of the distribution of gene frequency as $N \to \infty$ and selection and mutation coefficients approach zero. May (1973) has used a multivariate Ornstein–Uhlenbeck approximation in an ecological context.

The random walk is normal and the return process linear, and this guarantees us that the limiting distribution of $\mathbf{d}$ will be multivariate normal, with means zero and some covariance matrix $\mathbf{C}$. To characterize this distribution we need only find $\mathbf{C}$. We have

$$\mathbf{C} = E(\mathbf{d}\mathbf{d}'). \tag{10}$$

Making use of (9) and of the property that $E(\mathbf{d}\mathbf{e}') = E(\mathbf{e}'\mathbf{d}) = \mathbf{0}$, we have

$$\mathbf{C}^{(t+1)} = \mathbf{A}\mathbf{C}^{(t)}\mathbf{A}' + \mathbf{Q} \tag{11}$$

so that at equilibrium, $\mathbf{C}$ satisfies

$$\mathbf{C} = \mathbf{A}\mathbf{C}\mathbf{A}' + \mathbf{Q}. \tag{12}$$

This set of equations is linear in the c_{ij}, and it is possible to solve it on a computer if the matrix $\mathbf{C}$ is small (in the present case, it is only 3×3).

In the cases treated in this paper, we will usually find that matrix **A** is diagonal. In such cases (12) becomes

$$c_{ij} = a_{ii}c_{ij}a_{jj} + q_{ij}, \tag{13}$$

whereby immediately

$$c_{ij} = q_{ij}/(1 - a_{ii}a_{jj}). \tag{14}$$

If **Q** is itself diagonal, so will be **C**, with

$$c_{ii} = q_{ii}/(1 - a_{ii}^2), \tag{15}$$

with all other c_{ij} being zero. In these cases p, q and D are undergoing independent random walks with linear return, and each is independently normally distributed around its deterministic equilibrium value with a variance which can be readily calculated.

4. ADDITIVE–MULTIPLICATIVE FAMILY

We consider the viability model given in Table 2. When $k = 0$ this is the standard additive model, and when $k = 1$ it is the standard multiplicative model. When the loci are individually symmetric ($s = t, u = v$) this model is the symmetric model of

Table 2. *Fitness in the additive–multiplicative model*

	BB	*Bb*	*bb*
AA	$1-s-u+ksu$	$1-s$	$1-s-v+ksv$
Aa	$1-u$	1	$1-v$
aa	$1-t-u+ktu$	$1-t$	$1-t-v+ktv$

Lewontin & Kojima (1960). It is tedious but not difficult to show that for all k there is an equilibrium with

$$p = t/(s+t), \quad q = v/(u+v), \quad D = 0,$$

the position of this equilibrium not depending on r. However, its stability does depend on r, with this equilibrium being unstable if

$$r < k\left(\frac{st}{s+t}\right)\left(\frac{uv}{u+v}\right) \tag{16}$$

(Bodmer & Felsenstein, 1967). There are of course, other conditions which must be satisfied for the equilibrium to be stable, but these will not be derived here.

We can define four epistasis parameters in the usual way (as departures from additivity):

$$e_i = \sum_j k_j w_{ij}. \tag{17}$$

These turn out to be

$$\left.\begin{aligned} e_1 &= ksu, \\ e_2 &= -ksv, \\ e_3 &= -ktu \end{aligned}\right\} \tag{18}$$

and

$$e_4 = ktv.$$

https://doi.org/10.1017/S0016672300015287 Published online by Cambridge University Press

Note that, at the equilibrium mentioned above,

$$x_1e_1+x_2e_2+x_3e_3+x_4e_4=0,$$

$$\left.\begin{aligned} qe_1+(1-q)e_2&=0,\\ qe_3+(1-q)e_4&=0,\\ pe_1+(1-p)e_3&=0\\ \text{and}\qquad pe_2+(1-p)e_4&=0.\end{aligned}\right\} \tag{19}$$

Now we need to know some derivatives. We readily obtain, using (3) and (17),

$$\frac{\partial w_i}{\partial p}=w_{i2}-w_{i4}+qe_i, \tag{20}$$

$$\frac{\partial w_i}{\partial q}=w_{i3}-w_{i4}+pe_i, \tag{21}$$

and

$$\frac{\partial w_i}{\partial D}=e_i. \tag{22}$$

We want to get the derivatives of d_1', d_2' and d_3' with respect to the d_i, so as to compute the matrix **A**. To do this, we compute the derivatives of p', q' and D' with respect to p, q and D in the vicinity of the equilibrium. We have from (2), using (20)–(22),

$$\frac{\partial p'}{\partial p}=\frac{1}{\overline{w}}\left[qw_1+(1-q)w_2+q\frac{\partial w_1}{\partial p}+(1-q)\frac{\partial w_2}{\partial p}\right]-p'\frac{1}{\overline{w}}\frac{\partial \overline{w}}{\partial p}, \tag{23}$$

$$\frac{\partial p'}{\partial q}=\frac{1}{\overline{w}}\left[pw_1-pw_2+p\frac{\partial w_1}{\partial q}-p\frac{\partial w_2}{\partial q}\right]-p'\frac{1}{\overline{w}}\frac{\partial \overline{w}}{\partial q}, \tag{24}$$

$$\frac{\partial p'}{\partial D}=\frac{1}{\overline{w}}[w_1-w_2+x_2e_1+x_2e_2]-p'\frac{1}{\overline{w}}\frac{\partial \overline{w}}{\partial D}. \tag{25}$$

There are similar equations for q' and D. These derivatives are to be evaluated at the equilibrium point where $p'=p$, $q'=q$ and $D'=D$. The equilibrium in which we are interested has the same value of p, q and D whatever the value of r. In particular, it is an equilibrium when $r=0$. This means that it is a point of maximum or minimum $\overline{w}$, so that since it is on the interior of the space of possible values of p, q and D,

$$\frac{\partial \overline{w}}{\partial p}=\frac{\partial \overline{w}}{\partial q}=\frac{\partial \overline{w}}{\partial D}=0. \tag{26}$$

Using this and substituting into (23)–(25) first the expressions in (20)–(22), then (18) and fitnesses in Table 2, we finally find that

$$\frac{\partial p'}{\partial p}=1-\left[\frac{st}{(s+t)}-k\left(\frac{st}{s+t}\right)\left(\frac{uv}{u+v}\right)\right]\Big/\overline{w}, \tag{27}$$

$$\frac{\partial p'}{\partial q}=0, \tag{28}$$

$$\frac{\partial D'}{\partial q}=0. \tag{29}$$

https://doi.org/10.1017/S0016672300015287 Published online by Cambridge University Press

There are similar derivations for q' and D', resulting in

$$\frac{\partial q'}{\partial p} = \frac{\partial D'}{\partial p} = \frac{\partial D'}{\partial q} = 0, \tag{30}$$

$$\frac{\partial q'}{\partial q} = 1 - \left[\frac{uv}{(u+v)} - k\left(\frac{st}{s+t}\right)\left(\frac{uv}{u+v}\right)\right]\Big/\overline{w} \tag{31}$$

and

$$\frac{\partial D'}{\partial D} = 1 + \left[k\left(\frac{st}{s+t}\right)\left(\frac{uv}{u+v}\right) - r\right]\Big/\overline{w}. \tag{32}$$

Since, for example,

$$= \frac{\partial(q'-q)}{\partial(D-\hat{D})} = \frac{\partial q'}{\partial D},$$

equations (27) through (32) give us the values of the a_{ij} and the matrix **A** does indeed turn out to be diagonal, so that we can apply equations (14) and (15). We now need only calculate the q_{ij}.

As mentioned above, this is not quite as simple as it seems, since we are sampling N adult survivors of selection rather than $2N$ gametes.

Let P_{ij} be the frequency of genotype ij (i.e. consisting of gametes i and j) after selection. Then

$$P_{ij} = x_i x_j w_{ij}/\overline{w}. \tag{33}$$

Since the sampling of survivors of selection consists of N multinomial trials, the frequency of genotype ij being P_{ij}, the genotype frequencies P'_{ij} after sampling will have

$$\operatorname{cov}(P'_{ij}, P'_{km}) = -P_{ij}P_{km}/N \quad (ij \neq km) \tag{34}$$

and

$$\operatorname{var}(P'_{ij}) = P_{ij}(1-P_{ij})/N. \tag{35}$$

If x'_i are the gamete frequencies after selection and x''_i after selection and sampling, it is not difficult to show from (34) and (35) that, for example,

$$\operatorname{cov}(P'_{14}, x''_1) = -\frac{1}{N}P_{14}x'_1 + \frac{1}{2N}P_{14}, \tag{36}$$

$$\operatorname{cov}(P'_{23}, x''_1) = -\frac{1}{N}P_{23}x'_1, \tag{37}$$

$$\operatorname{cov}(x''_i, x''_j) = \frac{1}{2N}P_{ij} - \frac{1}{N}x'_i x'_j \quad (i \neq j) \tag{38}$$

and

$$\operatorname{var}(x''_1) = \frac{1}{2N}P_{ii} - \frac{1}{N}(x'_i)^2 + \frac{1}{2N}x'_i. \tag{39}$$

The frequencies of the gametes after selection, sampling and recombination are

$$x'''_i = x''_i - rk_i\tfrac{1}{2}(P'_{14}+P'_{41}) + rk_i\tfrac{1}{2}(P'_{23}+P'_{32}). \tag{40}$$

https://doi.org/10.1017/S0016672300015287 Published online by Cambridge University Press

We are only interested in covariances at the equilibrium point. Making use of the fact that $D = 0$ at that point, we have $x_i' = x_i$ and $P_{14} = P_{41} = P_{23} = P_{32}$, so that

$$\text{cov}(\delta x_i, \delta x_j) = \frac{1}{2N} P_{ij} - \frac{1}{N} x_i x_j + \frac{1}{2N} x_i \delta_{ij} - k_i k_j r(1-r) \frac{1}{N} P_{14}, \tag{41}$$

where δ_{ij} is the Kronecker delta.

One can now show fairly simply that at the equilibrium point,

$$\text{cov}(\delta p, \delta q) = \text{cov}(\delta x_1 + \delta x_2, \delta x_1 + \delta x_3) = 0. \tag{42}$$

To get covariances with D we approximate

$$\delta D \simeq x_1 \delta x_4 + x_4 \delta x_1 - x_2 \delta x_3 - x_3 \delta x_2. \tag{43}$$

This amounts to ignoring higher order terms in N, which we may safely do since the entire approach of linearizing selection equations is valid only for sufficiently large N. At the equilibrium point,

$$\text{cov}(\delta p, \delta D) = \text{cov}(\delta x_1 + \delta x_2, x_1 \delta x_4 + x_4 \delta x_1 - x_2 \delta x_3 - x_3 \delta x_2) = 0, \tag{44}$$

which can be proved, but only after a lot of tedious algebra, including use of the specific fitnesses given in Table 2, as well as the fact that we are at the equilibrium point and that $D = 0$. Clearly there must be an analogous proof that $\text{cov}(\delta q, \delta D) = 0$.

We now have shown that both **A** and **Q** are diagonal matrices, so that we can use (15). It is only necessary to compute the q_{ii}, which in this case will be the variances of δp, δq and δD.

These quantities can be found by further use of (41), of the viabilities in Table 2, and of the fact that we are computing these variances at an equilibrium at which $D = 0$. The proof is straightforward but too tedious to give here. We finally obtain

$$q_{11} = \text{var}(\delta p) = \frac{\hat{p}(1-\hat{p})}{N} \left[1 - \frac{\overline{w}_{Aa}}{2\overline{w}} \right], \tag{45}$$

where

$$\overline{w}_{Aa} = 1 - uv/(u+v),$$

$$q_{22} = \text{var}(\delta q) = \frac{\hat{q}(1-\hat{q})}{N} \left[1 - \frac{\overline{w}_{Bb}}{2\overline{w}} \right] \tag{46}$$

and

$$q_{33} = \text{var}(\delta D) = \frac{1}{2N} \hat{p}\hat{q}(1-\hat{p})(1-\hat{q}) \left[1 + k \left(\frac{st}{s+t} \right) \left(\frac{uv}{u+v} \right) \Big/ \overline{w} \right]. \tag{47}$$

We substitute (45)–(47) and (27), (31) and (32) into (15), and get the c_{ii}:

$$c_{11} = \text{var}(p) = \frac{\hat{p}(1-\hat{p})\,[1 - \overline{w}_{Aa}/(2\overline{w})]}{N[1-(1-Y_1)^2]}, \tag{48}$$

$$c_{22} = \text{var}(q) = \frac{\hat{q}(1-\hat{q})\,[1 - \overline{w}_{Bb}/(2\overline{w})]}{N[1-(1-Y_2)^2]} \tag{49}$$

https://doi.org/10.1017/S0016672300015287 Published online by Cambridge University Press

and

$$c_{33} = \text{var}\,(D) = \frac{\hat{p}(1-\hat{p})\,\hat{q}(1-\hat{q})\,[1+Y_3/\overline{w}]}{N\{1-[1-(r-Y_3)/\overline{w}]^2\}}, \tag{50}$$

where

$$Y_1 = st/(s+t) - Y_3,$$

$$Y_2 = uv/(u+v) - Y_3,$$

$$Y_3 = k\left(\frac{st}{s+t}\right)\left(\frac{uv}{u+v}\right),$$

$$\overline{w} = 1 - st/(s+t) - uv/(u+v) + Y_3,$$

$$\overline{w}_{Aa} = 1 - uv/(u+v)$$

and

$$\overline{w}_{Bb} = 1 - st/(s+t).$$

Of course (48)–(50) could be expressed entirely in terms of s, t, u, v, k and r, but there seems no point in doing so.

We have seen that when the stochastic process is linearized around the equilibrium $p = t/(s+t)$, $q = v/(u+v)$, $D = 0$, then p, q and D drift approximately independently of one another under the additive–multiplicative model. Their covariances are zero and their standing variances may be computed from (48)–(50). It is interesting to note that the formula (50) for the variance of D becomes infinite as r is reduced to the threshold level given by (16) for instability of the equilibrium.

Interestingly enough, the formulas (48) and (49) for var (p) and var (q) do not depend on the recombination fraction r. Thus there is no sign of associative overdominance, which would cause a closely linked neighbour under selection to increase the stability of a polymorphism. However, this does not contradict the work of Ohta & Kimura (1969*b*, 1971; see also Kimura & Ohta, 1971) since the lack of associative overdominance here is purely a result of the approximations used. Associative overdominance, which would surely be present in a more accurate approximation, is an effect of order $1/N^2$, and disappears when we drop terms of order higher than $1/N$. Its absence here thus reflects only the fact that it is of little importance for large N, compared to the selective effects of the loci themselves.

If we let s, t, u and v be small, we can obtain approximations to (48)–(50),

$$\text{var}\,(p) \simeq 1/[2N(s+t)], \tag{51}$$

$$\text{var}\,(q) \simeq 1/[2N(u+v)] \tag{52}$$

and

$$\text{var}\,(D) \simeq \hat{p}\hat{q}(1-\hat{p})\,(1-\hat{q})/\{2N[1-(1-r)^2]\}. \tag{53}$$

The first two expressions are identical to the results of Robertson's (1970) linearization of the stochastic process of a single overdominant locus. Thus each locus is varying as if the other were not present. Another comparison of interest involves the quantity R^2 (not to be confused with the square of the recombination fraction)

$$R^2 = E\left(\frac{D^2}{p(1-p)\,q(1-q)}\right). \tag{54}$$

https://doi.org/10.1017/S0016672300015287 Published online by Cambridge University Press

In this case, the denominator is effectively constant (the linearization discards its variation) so that

$$R^2 = \frac{E(D^2)}{\hat{p}\hat{q}(1-\hat{p})(1-\hat{q})} = \frac{\text{var}(D)}{\hat{p}\hat{q}(1-\hat{p})(1-\hat{q})}, \tag{55}$$

and finally

$$R^2 = 1/\{2N[1-(1-r)^2]\}. \tag{56}$$

R^2 is the quantity called r^2 by Hill & Robertson (1968) and in this approximation is equal to the σ^2 defined by Ohta & Kimura (1969*b*).

When the recombination fraction r is small, we obtain

$$R^2 \simeq 1/(4Nr) \tag{57}$$

in agreement with Hill & Robertson (1968) and Ohta & Kimura (1969*a*), Kimura & Ohta (1971). But for a larger value of r we must use (56) instead of (57). For $r = 0{\cdot}5$,

$$R^2 \simeq 2/(3N) \tag{58}$$

in agreement with Sved & Feldman (1973), and in contrast to Ohta & Kimura's value of $1/(2N)$. It might be expected that (57) would not be accurate for large values of r, since Ohta & Kimura derived it from a diffusion argument which is asymptotically correct as $N \to \infty$ and $r \to 0$ such that Nr remains constant. The present argument should be valid asymptotically as $N \to \infty$, without restriction on r.

It seems a reasonable conjecture that the independence of the drift of gene frequencies and of linkage disequilibrium parameters should hold for additive–multiplicative models with any number of loci and any number of alleles. Nothing in this paper guarantees that this conjecture will be valid. A new elegant and powerful matrix notation for multiple-locus additive and multiplicative models discovered by C. Z. Roux (1974) and by S. Karlin & U. Lieberman (personal communication) might make such a proof possible.

5. SYMMETRIC MODELS

When we consider special symmetric viability models, a stronger argument can be made that some of these quantities are uncorrelated. The argument is exact, without any approximations, and relies on the symmetry. It proves lack of correlation but does not calculate variances of the quantities. This argument should be readily generalizable to any number of loci and any number of alleles (providing that appropriate symmetry is retained).

We make only one assumption: that for all initial states of a population (provided they have both loci segregating initially) the asymptotic conditional distribution of doubly unfixed populations is the same. If this is true, which seems likely for all cases in which the double heterozygote is not lethal, we can compute the covariances $\text{cov}(p, q)$, $\text{cov}(p, D)$ and $\text{cov}(q, D)$ in the asymptotic conditional distribution, knowing only the fitnesses w_{ij}, the population size N, and the recombination fraction r. From this point on, let us subscript the quantities p, q and D with the name of the

allele. Thus p_A is the frequency of A, and D_{AB} is the linkage disequilibrium when AB is defined as coupling.

Suppose that interchanging the identities of alleles A and a, $A \leftrightarrow a$, leaves the fitness matrix unchanged. Then the fitnesses are symmetrical with respect to the A locus. So

$$\text{cov}(p_A, q_B) = \text{cov}(p_a, q_B) \quad \text{and} \quad \text{cov}(D_{AB}, q_B) = \text{cov}(D_{aB}, q_B).$$

But by their basic definitions

$$p_A = 1 - p_a \quad \text{and} \quad D_{AB} = -D_{aB},$$

so that we must also have

$$\text{cov}(p_A, q_B) = -\text{cov}(p_a, q_B) \quad \text{and} \quad \text{cov}(D_{AB}, q_B) = -\text{cov}(D_{aB}, q_B).$$

It follows that all of these covariances must be zero. We have proved the following rule:

I. If the fitness of genotypes are unaffected by exchanging $A \leftrightarrow a$, then, in the asymptotic conditional distribution of double unfixed populations,

$$\text{cov}(p_A, q_B) = 0 \quad \text{and} \quad \text{cov}(D_{AB}, q_B) = 0.$$

Exactly the same argument applied to symmetry at the B locus shows that:

II. If the fitnesses of the genotype are unaffected by exchanging $B \leftrightarrow b$, then in the asymptotic conditional distribution of doubly unfixed populations,

$$\text{cov}(p_A, q_B) = 0 \quad \text{and} \quad \text{cov}(D_{AB}, p_A) = 0.$$

When we have symmetry with respect to the double interchange $A \leftrightarrow a$ and $B \leftrightarrow b$, then since $D_{AB} = D_{ab}$,

$$\text{cov}(p_a, D_{AB}) = \text{cov}(p_a, D_{AB}) \quad \text{and} \quad \text{cov}(q_B, D_{AB}) = \text{cov}(q_b, D_{AB}).$$

It follows that:

III. If the fitnesses of the genotypes are unaffected by exchanging $A \leftrightarrow a$ and $B \leftrightarrow b$ simultaneously, then in the asymptotic conditional distribution of doubly unfixed populations,

$$\text{cov}(p_A, D_{AB}) = 0 \quad \text{and} \quad \text{cov}(q_B, D_{AB}) = 0.$$

Let us now apply rules I, II and III to the various symmetric viability models.

(i) *Lewontin & Kojima's symmetric model*

The model introduced by Lewontin & Kojima (1960) is shown in Table 3. The array of fitnesses remains unchanged when we exchange $A \leftrightarrow a$ or $B \leftrightarrow b$ or both. We can therefore conclude that

$$\text{cov}(p, q) = \text{cov}(p, D) = \text{cov}(q, D) = 0,$$

so that p, q and D are all uncorrelated. Lewontin & Kojima's model is really a special case of the additive–multiplicative model given in Table 2. The lack of correlation of p, q and D which holds approximately for the additive–multiplicative model is exact for the Lewontin–Kojima case.

(ii) *Bodmer's general symmetric model*

This model, shown in Table 4, was introduced in a less general form by Bodmer & Parsons (1962) and in full in Bodmer & Felsenstein (1967). The Lewontin–Kojima model as well as the earlier model of Kimura (1956) are special cases of this model. Inequality of α and δ implies that the fitnesses will not remain unchanged under either of the exchanges $A \leftrightarrow a$ or $B \leftrightarrow b$. Thus, rules I and II do not apply.

Table 3. *Lewontin & Kojima's symmetric viability model*

	BB	*Bb*	*bb*
AA	$1-s-u+e$	$1-s$	$1-s-u+e$
Aa	$1-u$	1	$1-u$
aa	$1-s-u+e$	$1-s$	$1-s-u+e$

Table 4. *Bodmer's general symmetric viability model*

	BB	*Bb*	*bb*
AA	$1-\alpha$	$1-\beta$	$1-\delta$
Aa	$1-\gamma$	1	$1-\gamma$
aa	$1-\delta$	$1-\beta$	$1-\alpha$

Table 5. *The half-symmetric model*

	BB	*Bb*	*bb*
AA	$1-s-u+e_1$	$1-s$	$1-s-v+e_2$
Aa	$1-u$	1	$1-v$
aa	$1-s-u+e_1$	$1-s$	$1-s-v+e_2$

However, the fitnesses do remain unchanged if we exchange both $A \leftrightarrow a$ and $B \leftrightarrow b$ simultaneously. So rule III applies and

$$\text{cov}(p, D) = \text{cov}(q, D) = 0.$$

Therefore, under Bodmer's general symmetric model p and q may covary, but D will be uncorrelated with either p or q.

(iii) *The 'Half-Symmetric' model*

One obvious sort of symmetric model is shown in Table 5. It is not a special case of any of the models given so far. It is symmetric in one locus but not in the other, so I refer to it as half-symmetric. Lewontin & Kojima's symmetric model is a special case of the half-symmetric model. Since we can exchange $A \leftrightarrow a$ and leave the fitness matrix unchanged, by rule I,

$$\text{cov}(p, q) = \text{cov}(q, D) = 0,$$

so q is not correlated with p or D, although they may be correlated with each other.

The half-symmetric model does not seem to have been introduced before. It has an equilibrium with

$$\hat{p} = \tfrac{1}{2}, \quad \hat{q} = (v - \tfrac{1}{2}s - \tfrac{1}{2}e_2)/(u + v - s - \tfrac{1}{2}e_1 - \tfrac{1}{2}e_2) \quad \text{and} \quad \hat{D} = 0.$$

https://doi.org/10.1017/S0016672300015287 Published online by Cambridge University Press

This equilibrium is unstable if $r < \frac{1}{2}\hat{q}(1-\hat{q})(e_1+e_2)$, by the instability criterion given by Bodmer & Felsenstein (1967). Of course, other equilibria might also exist.

If the asymptotic conditional distribution of doubly unfixed populations is independent of the initial position of the population (provided only that it is not fixed already at either locus) then on any small perturbation in the distribution it must return to the same distribution ultimately. For a finite N there are only a finite number of points in this distribution, hence it can be characterized by a finite number of parameters. The method of small parameters introduced into population genetics by Karlin & McGregor (1972) would seem to guarantee that a small change in the fitness matrix will lead to only a small change in the asymptotic conditional distribution, in particular, to only a small change in its second moments cov (p, q), cov (p, D) and cov (q, D). So models whose fitnesses are near those of any of these symmetric models will nearly share their properties. Thus if cov $(p, q) = 0$ in a symmetric model, a nearly symmetric model is likely to have cov $(p, q) \simeq 0$.

It will be interesting to see how much more we can prove using either the linearization or the symmetry arguments used here. In particular, it should not be difficult to investigate linked frequency-dependent polymorphisms. In the meantime, it should be pointed out that it is not difficult to calculate approximate variances and covariances of p, q and D numerically by solving the linear equations in the c_{ij} which are implicit in equation (12) above.

I wish to thank Drs M. Kimura, T. Maruyama, T. Ohta and T. Yamazaki of the National Institute of Genetics, Japan, for their hospitality and the use of their computing facilities during part of the work on this project. I also wish to thank John Gillespie for helpful comments and references. This work was supported by Task Agreement Number 5 of U.S. Atomic Energy Commission Contract AT(45-1)2225 with the University of Washington.

REFERENCES

Bodmer, W. F. (1960). Discrete stochastic processes in population genetics. *Journal of the Royal Statistical Society* B **22**, 218–244.

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

Bodmer, W. F. & Parsons, P. A. (1962). Linkage and recombination in evolution. *Advances in Genetics* **11**, 1–100.

Feller, W. (1951). Diffusion processes in genetics. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 227–246.

Hill, W. G. (1968). Population dynamics of linked genes in finite populations. *Proceedings of the XIIth International Congress of Genetics*, vol. II, pp. 146–147.

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), pp. 144–151.

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

Karlin, S. & McGregor, J. (1964). On some stochastic models in genetics. In *Stochastic Models in Medicine and Biology* (ed. J. Gurland), pp. 245–279. Madison: University of Wisconsin Press.

Karlin, S. & McGregor, J. (1972). Polymorphisms for genetic and ecological systems with weak coupling. *Theoretical Population Biology* **3**, 210–238.

Kimura, M. (1956). A model of a genetic system which leads to closer linkage under natural selection. *Evolution* **10**, 278–287.

KIMURA, M. & OHTA, T. (1971). Theoretical aspects of population genetics. *Monographs in Population Biology*, no. 4. Princeton: Princeton University Press.

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

LEWONTIN, R. C. & KOJIMA, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 116–129.

MAY, R. (1973). Stability in randomly fluctuating versus deterministic environments. *American Naturalist* **107**, 621–650.

NORMAN, M. F. (1972). Markov processes and learning models. *Mathematics in Science and Engineering*, vol. 84. New York and London: Academic Press.

OHTA, T. & KIMURA, M. (1969*a*). Linkage disequilibrium due to random genetic drift. *Genetical Research* **13**, 47–55.

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

OHTA, T. & KIMURA, M (1971). Behaviour of neutral mutants influenced by associated overdominant loci in finite populations. *Genetics* **69**, 247–260.

ROBERTSON, A. (1970). The reduction of fitness from genetic drift at heterotic loci in small populations. *Genetical Research* **15**, 257–259.

ROUX, C. Z. (1974). Hardy-Weinberg equilibria in random mating populations. *Theoretical Population Biology* **5**, 393–416.

SMITH, C. A. B. (1969). Local fluctuations in gene frequencies. *Annals of Human Genetics* **32**, 251–260.

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

SVED, J. A. (1972). Heterosis at the level of the chromosome and at the level of the gene. *Theoretical Population Biology* **3**, 491–506.

SVED, J. A. & FELDMAN, M. W. (1973). Correlation and probability methods for one or two loci. *Theoretical Population Biology* **4**, 129–132.