Genet. Res., Camb. (1974), **23**, *pp.* 281–289
Printed in Great Britain

Linkage disequilibrium and genetic variability

BY M. G. BULMER
Department of Biomathematics, University of Oxford

(*Received* 25 *September* 1973)

SUMMARY

It has been shown previously that, even in the absence of linkage, selection can cause an appreciable change in the genetic variance of a metric character due to disequilibrium; this change is temporary and is rapidly reversed when selection ceases. This result is here extended to allow for the effect of linkage, and it is shown that the change in the variance is effectively determined by the harmonic mean of the recombination fractions. The validity of the approximate general formula derived here has been checked by comparison with exact results obtained from models with five or six loci. In order to determine the likely value of the harmonic mean recombination fraction, a simple model was constructed in which it was assumed that loci are distributed at random along the chromosome maps. Results of computer simulations of this model are reported for different chromosome numbers and numbers of loci.

1. INTRODUCTION

It has been shown in a previous paper (Bulmer, 1971) that, if a metric character is determined by an effectively infinite number of unlinked loci, selection cannot cause any permanent change in the genetic variance but will cause a temporary change which is rapidly reversed when selection ceases. This effect must be due to genetic disequilibrium, that is to say to the correlation between pairs of loci which is induced by selection. This result depends on the assumptions that there is no linkage and that the number of loci is effectively infinite. When the number of loci is finite, as it must be in any actual situation, selection can also cause a permanent change in the genetic variance due to a change in the gene frequencies, but it seems likely that the results about the temporary change in the variance due to disequilibrium will remain approximately valid provided that the number of loci is not very small; the magnitude of the permanent change in the genetic variance has recently been discussed by O'Donald (1972). The purpose of this paper is to remove the first restriction and to show the modifications which must be made to the analysis in the presence of linkage.

We shall therefore consider a metric character determined by N loci, where N is large, and we shall denote by V_i the phenotypic variance (measured before selection operates) in the ith generation of selection, so that V_0 is the variance in the absence of selection. It will be assumed that the effect of selection in the ith generation is to change the variance from V_i to $V_i+\Delta V_i$. The effect of selection will usually be to decrease the phenotypic variance so that ΔV_i will be negative.

https://doi.org/10.1017/S0016672300014920 Published online by Cambridge University Press

We define the *disequilibrium contribution* in the ith generation of selection as

$$d_i = V_i - V_0. \tag{1}$$

It has been shown previously that d_i behaves as if it were a component of the additive genetic variance, A_i, so that we may write

$$A_i = A_0 + d_i, \tag{2}$$

where A_0 is the additive genetic variance in the absence of selection. The heritability in the ith generation can be defined as

$$h_i^2 = A_i/V_i. \tag{3}$$

The basic result proved in the previous paper (Bulmer, 1971) is that, in the absence of linkage,

$$d_{i+1} = \tfrac{1}{2}d_i + \tfrac{1}{2}h_i^4 \Delta V_i. \tag{4}$$

The term $\frac{1}{2}h_i^4\Delta V_i$ is due to the fresh disequilibrium introduced by the action of selection in the ith generation; the term $\frac{1}{2}d_i$ is due to the fact that, in the absence of linkage, only half the disequilibrium contribution present in the previous generation is preserved. If ΔV_i approaches a limiting value, ΔV^*, under continued selection, then d_i will tend to a limiting value d^*, which can be evaluated by putting $d_{i+1} = d_i$ in equation (4); hence

$$d^* = h^{*4}\Delta V^*, \tag{5}$$

where h^{*2} is the limiting value of the heritability.

2. THE EFFECT OF LINKAGE

The basic equation (4) has been derived (Bulmer, 1971) by considering the regression between relatives, such as child on parent or grandchild on grandparent, on the assumption that the joint probability distribution of these related individuals is multivariate normal so that the regressions are linear and homoscedastic. It was shown that this assumption is true for parent and child when there is a large number of loci whether or not there is linkage. The expression $\frac{1}{2}h_i^4\Delta V_i$ for the fresh disequilibrium contribution in the offspring generation generated by selection in the previous generation was based on the child–parent regression and is therefore unaffected by the presence of linkage. On the other hand rather more than half of the disequilibrium contribution present in the previous generation will be preserved in the presence of linkage, so that d_i must be multiplied by a factor larger than $\frac{1}{2}$ in the first term on the right-hand side of equation (4).

To compute the effect of this change, suppose that δ_i is the contribution to d_i from a particular pair of loci with recombination fraction r. Then in the next generation a fraction $(1-r)\delta_i$ of the contribution from this pair of loci will be preserved since under random mating δ_i must be determined entirely by gametic phase disequilibrium which decays at a rate $(1-r)$ per generation. We now make the simplifying assumption that the genetic effects of all the loci are the same, so that the fresh contribution to δ_{i+1} as a result of selection in the ith generation must be the same for

https://doi.org/10.1017/S0016672300014920 Published online by Cambridge University Press

all loci, and must therefore be equal to $\frac{1}{2}h_i^4\Delta V_i/\frac{1}{2}N(N-1)$, since there are altogether $\frac{1}{2}N(N-1)$ pairs of loci which contribute to d_i if there are N loci. (It is assumed in the above argument that the fresh contribution to δ_{i+1} does not depend on δ_i. The validity of this assumption will be investigated numerically in the next section.) It follows that

$$\delta_{i+1} = (1-r)\,\delta_i + \frac{\frac{1}{2}h_i^4\Delta V_i}{\frac{1}{2}N(N-1)}\,. \tag{6}$$

If δ_i tends to a limiting value δ^*, then this limiting value can be evaluated by putting $\delta_{i+1} = \delta_i$ in equation (6), so that

$$\delta^* = \frac{\frac{1}{2}h^{*4}\Delta V^*}{\frac{1}{2}N(N-1)} \times \frac{1}{r}\,. \tag{7}$$

Hence

$$d^* = \Sigma\delta^* = \frac{\frac{1}{2}h^{*4}\Delta V^*}{\frac{1}{2}N(N-1)} \times \Sigma\frac{1}{r} = \tfrac{1}{2}h^{*4}\Delta V^*/H, \tag{8}$$

where H is the harmonic mean of the recombination fractions. In the absence of linkage the rate of approach to the limiting value is very rapid, more than half the final value being attained after one generation of selection (Bulmer, 1971). The rate of approach will be slower in the presence of linkage, but nevertheless it is still likely to be quite rapid.

Equation (8) provides an implicit equation for d^* whose solution can usually be found if the selection function is known. Consider, for example, a population subject to 'nor-optimal' selection under which the fitness of an individual with phenotypic value y is $\exp\,[-c(y-\theta)^2]$. If y is normally distributed with mean θ (as it must be at equilibrium) and with variance V, the change in the variance as the result of selection is

$$\Delta V = -2cV^2/(1+2cV). \tag{9}$$

By using (1)–(3), Equation (8) can be written

$$d^*[1+2c(V_0+d^*)]H + c(A_0+d^*)^2 = 0, \tag{10}$$

where A_0 and V_0 are the additive genetic variance and the phenotypic variance in the absence of selection (i.e. in linkage equilibrium). This leads to a quadratic equation for d^*; only one of the roots lies in the permissible range, so that the solution is unique. Similarly, under the quadratic optimum model in which the fitness function is $1-c(y-\theta)^2$, it will be found that

$$\Delta V = -2cV^2/(1-cV), \tag{11}$$

which again leads to a quadratic equation for d^*:

$$d^*[1-c(V_0+d^*)]H + c(A_0+d^*)^2 = 0. \tag{12}$$

Finally, under many forms of artificial selection, $\Delta V = -kV$, where k is a constant. This leads to the quadratic equation:

$$d^*(V_0+d^*)H + \tfrac{1}{2}k(A_0+d^*)^2 = 0. \tag{13}$$

The practical applications of these results will be discussed further in Section 4.

https://doi.org/10.1017/S0016672300014920 Published online by Cambridge University Press

3. COMPARISON WITH EXACT RESULTS

The preceding results depend on two main assumptions, first that the fresh contribution to linkage disequilibrium in any generation is the same for all pairs of loci, so that the final disequilibrium is inversely proportional to the recombination fraction, and secondly that the regression of child on parent is linear and homoscedastic so that the fresh disequilibrium contribution in any generation is altogether $\frac{1}{2}h^4\Delta V$. The first assumption seems plausible provided that the linkage disequilibrium between all pairs of loci is fairly small. The second assumption is true when the number of loci is large since in the limit the joint distribution of offspring and parents is multivariate normal (Bulmer, 1971), but may break down when there are only a few loci, particularly in the presence of dominance which may introduce some curvature into the regression. It is therefore desirable to compare the predictions made here with exact results obtained by computing equilibrium gametic frequencies. Unfortunately it is not possible to calculate exact equilibria with more than five or six loci in a reasonable amount of computer time.

Lewontin (1964) has tabulated equilibrium gametic frequencies and linkage disequilibrium parameters for a five-locus model under quadratic optimum selection. At each locus the three genotypes have effects 6, 4·8 and -6 respectively, and the five loci contribute additively to the phenotypic value, y. The fitness function is $1-(y-24)^2/3000$. The loci are spaced at equal distances along a line with recombination fraction R between adjacent loci. To test whether linkage disequilibrium is inversely proportional to the recombination fraction I have calculated $D'_{ij}\times r_{ij}$, where D'_{ij} is the linkage disequilibrium between loci i and j relative to its maximum possible value (as given by Lewontin) and r_{ij} is the recombination fraction. The results are shown in Table 1. It will be seen that this quantity is nearly constant in each column except when the linkage is very tight ($R < 0{\cdot}01$). If the line corresponded to a chromosome of length 100 centimorgans, the map distance between adjacent loci would be 20 centimorgans, which corresponds to a recombination fraction of 0·165. Thus D'_{ij} is inversely proportional to r_{ij} to a good approximation over the range of values of R likely to be of biological significance.

Table 1 also shows the observed value of d^* ($= V^*-V_0$, where V^* is the variance in linkage disequilibrium and V_0 is the variance in linkage equilibrium with the same gene frequencies) and the value of d^* predicted from equation (12). The predicted value is higher than the observed value for all R, but the agreement is nevertheless remarkably good in view of the small number of loci and the presence of dominance.

As a second example I have considered a model with five loci without dominance, the three genotypes at each locus having effects 6, 0 and -6 respectively; the fitness function was taken as exp $(-y^2/3000)$. The five loci are spaced evenly along a line as in Lewontin's model. The equilibrium gametic frequencies with all gene frequencies equal to $\frac{1}{2}$ were computed and the results are shown in Table 2. (This is an unstable equilibrium, but it serves as well as a stable equilibrium to demonstrate the amount of linkage disequilibrium generated by selection.) The quantity $D'_{ij}\times r_{ij}$ is not as constant as in Lewontin's model, due to the development of two position effects

https://doi.org/10.1017/S0016672300014920 Published online by Cambridge University Press

under tight linkage; D'_{12} is larger than expected (and in particular larger than D'_{23}), while D'_{24} is smaller than expected (and in particular smaller than D'_{13}). Nevertheless the observed value of d^* is close to the value predicted from equation (10) even under very tight linkage; it seems that the two position effects, which are in opposite directions, have to a large extent cancelled each other when the total disequilibrium is considered.

Table 1. *Linkage disequilibrium under a five-locus model* (Lewontin, 1964)

R between adjacent loci	0·0005	0·00125	0·005	0·01	0·02	0·03	0·05	0·1	0·234
$-D'(12)\times r(12)\times 10^3$	0·41	0·82	1·76	2·23	2·58	2·72	2·86	3·03	3·07
$-D'(13)\times r(13)\times 10^3$	0·71	1·26	2·11	2·44	2·65	2·74	2·76	2·96	3·12
$-D'(14)\times r(14)\times 10^3$	0·93	1·45	2·19	2·47	2·66	2·66	2·66	3·04	3·39
$-D'(15)\times r(15)\times 10^3$	1·03	1·44	2·18	2·45	2·59	2·61	2·55	3·03	3·16
$-D'(23)\times r(23)\times 10^3$	0·42	0·86	1·81	2·25	2·59	2·73	2·85	2·97	3·06
$-D'(24)\times r(24)\times 10^3$	0·73	1·30	2·13	2·45	2·65	2·75	2·77	2·93	3·12
$-d^*$ (observed)	7·63	4·44	1·43	0·76	0·40	0·27	0·16	†	0·04
$-d^*$ (predicted)	8·79	5·51	1·91	1·01	0·52	0·35	0·21	†	0·05

† Not calculated owing to an internal inconsistency in the data.

Table 2. *Linkage disequilibrium under a five-locus model without dominance*

R between adjacent loci	0·001	0·0025	0·005	0·01	0·025	0·05	0·1	0·25	0·5
$-D'(12)\times r(12)\times 10^3$	0·45	1·00	1·66	2·47	3·57	4·29	4·82	5·21	5·31
$-D'(13)\times r(13)\times 10^3$	0·23	0·56	1·02	1·73	2·95	3·86	4·57	5·12	5·31
$-D'(14)\times r(14)\times 10^3$	0·20	0·49	0·93	1·63	2·87	3·81	4·54	5·11	5·31
$-D'(15)\times r(15)\times 10^3$	0·28	0·66	1·19	1·77	3·20	4·04	4·67	5·16	5·31
$-D'(23)\times r(23)\times 10^3$	0·24	0·57	1·04	1·75	2·99	3·91	4·61	5·14	5·31
$-D'(24)\times r(24)\times 10^3$	0·13	0·33	0·65	1·32	2·60	3·62	4·44	5·08	5·31
$-d^*$ (observed)	68·0	62·7	55·4	45·0	29·3	19·0	11·4	5·7	3·8
$-d^*$ (predicted)	71·6	62·7	54·1	44·1	30·0	20·4	12·9	6·7	4·6

As a final example I have considered a model with six loci without dominance, the three genotypes at each locus having effects 6, 0 and -6 as before, distributed in three pairs on three chromosomes, with recombination fraction R between the two loci on the same chromosome. The fitness function was taken as $\exp(-y^2/3600)$, which gives the same intensity of selection as in the previous model with five loci. The equilibrium gametic frequencies with all gene frequencies equal to $\frac{1}{2}$ were computed as before; the results are shown in Table 3. The linkage disequilibrium is inversely proportional to the recombination fraction if $R > 0{\cdot}05$, but this relationship breaks down badly under tight linkage. The observed value of d^* is in reasonable agreement with the value predicted from equation (10) when $R > 0{\cdot}01$, but is rather larger than its predicted value under tight linkage. However, if there are only two loci on a chromosome, the recombination fraction between them is unlikely to be as low as 0·01.

It is concluded that equation (8) is likely to provide a reasonable approximation

https://doi.org/10.1017/S0016672300014920 Published online by Cambridge University Press

to the disequilibrium generated by selection for a character determined by about five loci. When the number of loci becomes larger the assumption that the regression of child on parent is linear and homoscedastic will become more accurate since it is an asymptotic result based on the central limit theorem which becomes exact when the number of loci is infinite. The first assumption, that linkage disequilibrium is inversely proportional to the recombination fraction, will be affected by two factors when the number of loci increases; adjacent loci will tend to be closer together but at the same time the selection pressure at each locus will become weaker. The disequilibrium between adjacent loci is thus likely to remain approximately unchanged, and the first assumption is likely to be satisfied as accurately when the number of loci is large as when it is small. It is therefore suggested that equation (8) will provide a satisfactory approximation under most circumstances likely to be of biological significance.

Table 3. *Linkage disequilibrium under a six-locus model with three chromosomes*

R between adjacent loci	0·001	0·0025	0·005	0·01	0·025	0·05	0·1	0·25	0·5
$-D'$ (adj. loci) $\times R \times 10^3$	0·91	1·94	3·01	3·90	4·35	4·40	4·41	4·40	4·40
$-D'$(non-adj. loci) $\times \frac{1}{2} \times 10^3$	0·04	0·24	0·74	1·71	3·09	3·74	4·10	4·32	4·40
$-d^*$ (observed)	97·6	83·8	65·7	43·6	21·4	12·7	8·3	5·6	4·7
$-d^*$ (predicted)	71·8	57·1	44·7	32·7	19·8	13·3	9·2	6·5	5·5

4. APPLICATIONS

Before discussing the disequilibrium generated under natural or artificial selection it is necessary to consider the value of the harmonic mean recombination fraction, H, likely to be found in a natural population. As an approximate model we shall suppose that there are n pairs of chromosomes, each with the same length of 100 centimorgans in map units, that N loci are distributed at random along the chromosome maps, and that the recombination fraction, r, between a pair of loci is $\frac{1}{2}$ if the loci are on different chromosomes, and is given by the mapping function

$$r = \tfrac{1}{2}(1-e^{-2x}) \tag{14}$$

if the loci are on the same chromosome at a map distance of x morgans. (See Bailey, 1961 for an account of the theory of mapping functions.) It is of course realised that chromosomes do not all have the same length, that loci may not be distributed at random along their length, and that the above mapping function may be oversimplified since it assumes that there is no interference. Nevertheless, it seems reasonable to suppose that this model will give results of the right order of magnitude. The only complication which will be considered is absence of crossing over in the male sex in *Drosophila* which can be taken into account by using the mapping function

$$r = \tfrac{1}{4}(1-e^{-2x}) \tag{15}$$

for loci on the same chromosome, and $r = \frac{1}{2}$ as before for loci on different chromosomes. As a rough model for *Drosophila melanogaster* it seems adequate to assume a haploid number of 3 chromosomes, each 100 centimorgans long, since the fourth chromosome is extremely short; in fact, of course, the first chromosome is shorter and the second and third chromosomes are longer than 100 centimorgans, but it seems unlikely that these departures will make an appreciable difference in this context.

Since the loci are assumed to be randomly distributed along the chromosome maps, the harmonic mean recombination fraction will be a random variable, depending on the positions of the loci, with a highly intractable distribution. Recourse was therefore made to simulation. Consider as an example the case with 6 pairs of chromosomes and 12 loci. 12 random numbers between 1·0 and 6·99 were obtained by means of a pseudo-random number generator. The integral part of each number gives the chromosome on which the locus lies, and the decimal part its position along the chromosome, measured in morgans. For each of the 66 possible pairs of loci the recombination fraction was determined from equation (14), and then the harmonic mean, H, of these 66 recombination fractions was evaluated. This procedure was repeated 100 times to give 100 different values of H, which were then arranged in rank order. The median and the lower and upper deciles of this distribution (i.e. the 50th, 10th and 90th observations in rank order) were recorded; the median was 0·33, the lower decile 0·16 and the upper decile 0·42. These quantities are shown in Table 4 for different haploid numbers and numbers of loci. As might be expected the likely range of values of H depends critically on the number of chromosomes.

Table 4. *Median and inter-decile range of the harmonic mean recombination fraction based on* 100 *computer simulations*

Haploid number …	3†	3	6	12	24	48
Number of loci			Median			
6	0·17	0·26	0·39	0·46	0·50	0·50
12	0·11	0·20	0·33	0·41	0·46	0·49
24	0·10	0·17	0·27	0·38	0·43	0·48
48	0·08	0·15	0·26	0·35	0·41	0·46
96	0·07	0·13	0·21	0·32	0·39	0·45
			Inter-decile range			
6	0·06–0·28	0·11–0·40	0·17–0·47	0·25–0·50	0·37–0·50	0·43–0·50
12	0·05–0·19	0·09–0·29	0·16–0·42	0·25–0·48	0·29–0·49	0·44–0·50
24	0·03–0·15	0·07–0·25	0·15–0·36	0·18–0·43	0·29–0·47	0·40–0·49
48	0·04–0·12	0·09–0·20	0·17–0·32	0·22–0·40	0·31–0·45	0·38–0·48
96	0·04–0·10	0·07–0·17	0·15–0·26	0·23–0·37	0·32–0·43	0·39–0·47

† No crossing-over in one sex, e.g. *Drosophila*.

We can now consider the significance of the basic equation

$$d^* = \tfrac{1}{2}h^{*4}\Delta V^*/H. \tag{8 bis}$$

In a natural population which has been subject to stabilizing selection for a considerable time h^{*2} is the observed heritability and ΔV^* the observed difference in the variance before and after selection. For example, if the heritability is $\frac{1}{2}$ and if the variance is found to decline from 100 before selection to 90 after selection, it can be concluded that $d^* = -1\cdot25/H$. If the chromosome number is large, then H is unlikely to be much less than $\frac{1}{2}$, so that d^* would be about $-2\cdot5$; if selection were relaxed the genetic variance would rise from 50 to 52·5. For *Drosophila*, on the other hand, H is more likely to be of the order of 0·1, which makes d^* about $-12\cdot5$; relaxation of selection would cause an increase in the genetic variance from 50 to 62·5.

Table 5. *The final reduction in the genetic variance when the initial heritability is* $\frac{1}{2}$ *and the initial phenotypic variance is* 100

H	k ... 0·01	0·10	0·25	0·50	0·75	0·90	0·99
0·02	5·3	23·3	31·4	36·4	38·7	39·7	40·1
0·04	2·9	16·7	25·4	31·4	34·5	35·7	36·3
0·06	2·0	13·1	21·5	28·0	31·4	32·9	33·6
0·08	1·5	10·8	18·8	25·4	29·0	30·6	31·3
0·10	1·2	9·2	16·7	23·3	27·1	28·7	29·5
0·20	0·6	5·3	10·8	16·7	20·5	22·3	23·2
0·30	0·4	3·7	8·0	13·1	16·7	18·4	19·3
0·40	0·3	2·9	6·4	10·8	14·1	15·7	16·6
0·50	0·2	2·3	5·3	9·2	12·2	13·7	14·6

We consider next an artificial selection experiment with stabilizing selection in which a proportion P of the population nearest the mean is chosen to provide the parents of the next generation. If the character is normally distributed with variance V, then $\Delta V = -kV$, where

$$k = 2f(z)/P. \tag{16}$$

In this equation z is the standard normal deviate corresponding to $\frac{1}{2}(1-P)$ and $f(z)$ is the standard normal density function. For example, if the middle 20% are selected $k = 0\cdot9785$; the values of k for the middle 50% and the middle 80% are 0·8574 and 0·5623 respectively. The resulting quadratic equation for d^* is given by equation (13). Values of $-d^*$, the reduction in the additive genetic variance, are shown in Table 5 for different values of k and H, in the case when $A_0 = 50$ and $V_0 = 100$. For example, if the middle 50% are selected, so that $k = 0\cdot8574$, then d^* is about -13 if $H = 0\cdot5$; in an organism with a high chromosome number the additive genetic variance would be reduced from 50 before selection to about 37 after several generations of selection. For *Drosophila*, on the other hand, H might be 0·1 or less, so that the additive genetic variance might be reduced eventually to about 20.

We consider finally an artificial selection experiment in which a proportion P of individuals with the highest phenotypic values is chosen to provide the parents of the next generation. In this case $\Delta V = -kV$, where

$$k = \frac{f(z)}{P}\left[\frac{f(z)}{P} - z\right], \tag{17}$$

z being the standard normal deviate corresponding to P. For example, if $P = 0{\cdot}2$ so that 20% of the population with the highest phenotypic values are selected in each generation, then $k = 0{\cdot}7818$. The effect of selection when $A_0 = 50$ and $V_0 = 100$ can be found from Table 5. However, this analysis does not take into account the permanent change in the variance which occurs under directional selection due to the change in the gene frequencies. As the number of loci becomes infinitely large the permanent change in the variance will become infinitesimally small, but it may clearly be of considerable importance in most practical situations. It is suggested that equation (13) will still remain approximately valid after several generations of selection if A_0 and V_0 are re-interpreted as being the additive genetic variance and the phenotypic variance which would be observed in linkage equilibrium under the gene frequencies prevailing at the time, in other words the values which would be observed if selection were relaxed.

REFERENCES

BAILEY, N. T. J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage.* Oxford: Clarendon Press.

BULMER, M. G. (1971). The effect of selection on genetic variability. *American Naturalist* **105**, 201–211.

LEWONTIN, R. C. (1964). The interaction of selection and linkage. II. Optimum models. *Genetics* **50**, 757–782.

O'DONALD, P. (1972). Natural selection for a quantitative character over several generations. *Nature* **237**, 113–114.

https://doi.org/10.1017/S0016672300014920 Published online by Cambridge University Press