A theoretical analysis of linkage disequilibrium produced by genetic drift in *Drosophila* populations

CATHERINE MONTCHAMP-MOREAU AND MARIANO KATZ

Laboratoire de Génétique des Populations, Universités Paris 6 et Paris 7, Tour 42, 2 place Jussieu, 75005 PARIS*

(Received 4 February 1986 and in revised form 8 July 1986)

Summary

We analyse the progression of linkage disequilibrium produced by random genetic drift in populations subject to cyclical fluctuations in size. Our model is applied to natural populations of *Drosophila* which show an annual demographic cycle of bottleneck (finite size) and demographic burst (size supposed to be infinite). In these populations, linkage disequilibrium stabilizes in such a way that, at equilibrium, the expected square of the correlation of gene frequencies $E(r^2)$ shows a stable cycle from year to year. If two loci are tightly linked, $E(r^2)$ barely varies during the annual cycle. Its values remain close to the value expected in a population of the same but constant effective size. If two loci are loosely linked, fluctuations in $E(r^2)$ are large. The maximum value, reached at the end of the bottleneck, is 10 to 100 times greater than the value obtained at the end of the burst. Our results show that the interpretation of observed linkage disequilibrium, by means of statistical tests, requires an accurate knowledge of population demography.

1. Introduction

Many authors have measured linkage disequilibrium D betwen allozymes in natural populations of *Drosophila* (see Hedrick, 1983, for a review). Significant values of D are generally rare and can be due to sampling error. However, Langley (1977) demonstrates a faint negative correlation between the values of the observed gametic χ^2 for two loci and their effective recombination frequency in natural populations of *Drosophila melanogaster*. The same observation was made by Montchamp-Moreau (1985) in natural populations of *Drosophila simulans*. Such results can be due either to random drift or to selection. In order to analyse the behaviour of linkage disequilibrium in small populations with constant size ($8 \leqslant Ne \leqslant 25$), Hill & Robertson (1968) performed Monte-Carlo simulations.

In this paper, we present a theoretical analysis and the results of simulations concerning the linkage disequilibrium which can be produced by random drift in natural populations of *Drosophila*. We based our study on the demographic data available for such populations (Mukai *et al.* 1971; Begon, 1977; McInnis, 1982). Taking into account annual fluctuations in their size, we show that stable linkage disequilibrium can be produced by random drift in these populations.

* UA CNRS 693.
Requests for reprints should be made to C. Montchamp-Moreau.

2. Kinetics of linkage disequilibrium with constant Ne

Weir & Hill (1980) gave an approximate expression for the variance of the gametic correlation coefficient r, produced by random drift in dioecious populations with random mating:

$$E(r^2) \simeq \frac{1+c^2}{2N_e\, c(2-c)}. \tag{1}$$

This approximation is satisfactory when N_e is large and c is not too small. Hill & Robertson (1968) showed that $E(r^2)$ is not very sensitive to the initial allelic frequencies at each of the loci concerned.

Sved & Feldman (1973) established, for monoecious populations with random selfing, the recurrence relationship describing the increasing in $E(r^2)$ from one generation to the next. From this relationship, they obtained for $E(r^2)$ at generation t (with $E(r^2)_0 = 0$):

$$E(r^2)_t = \frac{1-\left[\left(1-\frac{1}{2N_e}\right)(1-c)^2\right]^t}{1+[(2N_e-1)(2c-c^2)]}. \tag{2}$$

We undertook Monte-Carlo simulations in order to compare values of $E(r^2)$ obtained with the values predicted from (1) and (2).

(i) *The model*

We consider a dioecious population with random mating, sex ratio = 1, no mutation, no crossing-over in males (subsequently, c will represent the effective recombination frequency, i.e. half the recombination frequency in the females).

The initial linkage disequilibrium (D_0) is zero; there are two alleles at each locus, and both initial allelic frequencies are 0·5.

We made simulations for the following situations:

$N_e = 1000, 5000,$
$c = 0{\cdot}01; 0{\cdot}05; 0{\cdot}10; 0{\cdot}25.$

In each case, between 500 and 1000 runs were performed, for at least 30 generations.

(ii) *Results and discussion*

The simulated evolution of $E(r^2)$ during the 30 first generations (for $N_e = 1000$) is represented in Fig. 1. The curve obtained from Sved and Feldman's recurrence relationship (2) fits our simulated values quite closely. Similar results were obtained for $N_e = 5000$. Then this formula, strictly established for a monoecious population with random selfing, can also be used to approximate the evolution of $E(r^2)$ in dioecious populations if N_e is large.

Let us consider now the limiting value of $E(r^2)$ for an infinite number of generations ($t \to \infty$); we have two expressions:
- Weir and Hill's formula (1),
- Sved and Feldman's formula (2), which becomes:

$$E(r^2) = \frac{1}{1+[(2N_e-1)(2c-c^2)]}. \tag{3}$$

The values obtained from (1) and (3) in our situations are given in Table 1. The first values (from 1) are always slightly greater than the second values (from 3). Hill (1976) showed that for $N_e = 8$, the exact value $E(r^2)_\infty$ is underestimated by (3) and overestimated by (1). Being close to the equilibrium, the values obtained in our simulations are sometimes overestimates, sometimes underestimates of the values predicted by (1) and (3). However, (1) and (3) give very close values (the difference varies from 1% to 6%); and there is a satisfactory correspondance between the values of $E(r^2)$ observed from generation to generation by simulation and the values obtained by Sved and Feldman's relationship (2). It seems reasonable to use this relationship to calculate the number of generations (G 1%) necessary to approach the limiting value within 1%:

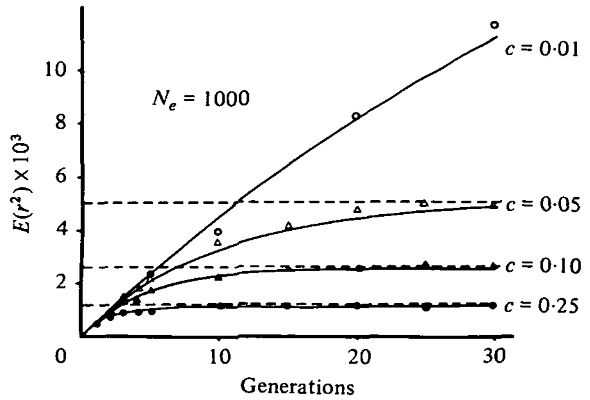

Fig. 1. Evolution of $E(r^2)$ in populations of constant size. $N_e = 1000$ (●, ○, ▲, △: simulated values; --- $E(r^2)$ from eq. (1); —— $E(r^2)$ from eq. (2)).

$$G\,1\% = \frac{2\ln 10}{\ln\left(\frac{1}{1-c}\right)^2\left(\frac{1}{1-(1/2N_e)}\right)}. \tag{4}$$

When N_e is large, (4) is approximated by

$$G\,1\% = \frac{2\ln 10}{\ln\left(\frac{1}{1-c}\right)^2+\frac{1}{2N_e}}$$

which shows that G 1% depends mainly on c. The values of G 1% when $N_e = 1000$ and $N_e = 5000$ are given in Table 1. They are, indeed, independent of N_e. G 1% varies from 8 generations ($c = 0{\cdot}25$) to 228 generations ($c = 0{\cdot}01$). The time necessary before the limiting value of $E(r^2)$ for two loci moderately linked ($c = 0{\cdot}10$) is reached is rather long (22 generations).

The relevance of the relationships established in the case of a constant population size needs to be examined, as the size of a natural population of *Drosophila* fluctuates greatly in the course of a year. We next analyse the evolution of $E(r^2)$ when the population size fluctuates.

3. Linkage disequilibrium evolution with fluctuating Ne

Begon (1977) estimated the number of generations during the winter bottleneck of an English population of *D. subobscura* to be between 1 and 2 (with $N_e = 900$), and the number of generations during the demographic burst to be about 4 ($N_e = 10000$ to 15000). For North American populations of *D. simulans* and *D. melanogaster*, the effective size was estimated to be between 500 and 1000 (for each species) from November to June, and about 10000 from August to October (McInnis *et al.* 1982). From these data, we have analysed the behaviour of $E(r^2)$ when the population shows cyclical fluctuations.

(i) *The model*

T_1 = number of generations with a finite size N_e,
T_2 = number of generations with an infinite size,
$T_1 + T_2$ = number of generations in one cycle.

Initial linkage disequilibrium $D_0 = 0$.

For T_1 generations, $E(r^2)$ increases according to (2), while for T_2 generations, $E(r^2)$ decays according to the following relationship:

$$r_{t+1} = (1-c)\,r_t$$

Table 1. *Square of the correlation of gene frequencies expected at equilibrium in dioecious populations with random mating (from (1)) and in monoecious populations with random selfing (from (3)), for various values of N_e and c. $G_{1\%}$ = number of generations before reaching equilibrium close to 1%.*

N_e	c	$E(r^2)$ from (1) $\frac{1+c^2}{2N_e c(2-c)}$ $(\times 10^3)$	$E(r^2)$ from (3) $\frac{1}{1+(2N_e-1)(2c-c^2)}$ $(\times 10^3)$	$G_{1\%}$
1000	0·01	25·13	24·52	224
	0·05	5·14	5·10	45
	0·10	2·66	2·63	22
	0·25	1·21	1·14	8
5000	0·01	5·03	5·00	228
	0·05	1·03	1·02	45
	0·10	0·53	0·53	22
	0·25	0·24	0·23	8

so

$$E(r^2)_{t+1} = (1-c)^2 E(r^2)_t. \tag{5}$$

Equilibrium is established when the increase Δ_1 of $E(r^2)$ for T_1 generations is exactly counterbalanced by the decrease Δ_2 of $E(r^2)$ for the following T_2 generations.

We have:

$$\Delta_1 = E(r^2)_t - E(r^2)_{t-T_1} \quad \text{(according to (2))},$$
$$\Delta_2 = E(r^2)_t - E(r^2)_{t+T_2} \quad \text{(according to (5))}$$

for $\Delta_1 = \Delta_2$, we obtain the maximum value of $E(r^2)$ for an equilibrium cycle:

$$E(r^2)_+ = \frac{1-\left[\left(1-\frac{1}{2N_e}\right)(1-c)^2\right]^{T_1}}{\left[1-\left(1-\frac{1}{2N_e}\right)^{T_1}(1-c)^{2T_1+2T_2}\right] \times \left[1+(2N_e-1)(2c-c^2)\right]} \tag{6}$$

(see appendix), and the minimum value:

$$E(r^2)_- = (1-c)^{2T_2} E(r^2)_+. \tag{6'}$$

When N_e is large (6) is approximated by:

$$E(r^2)_+ = \frac{1-\left[(1-c)^{2T_1} e^{-T_1/2N_e}\right]}{2N_e\left[1-e^{-T_1/2N_e}(1-c)^{2T_1+2T_2}\right] \times \left[1-e^{-1/2N_e}(1-c)^2\right]}.$$

This shows the inter-relations of generation number and population size.

If $E(r^2)_t < E(r^2)_+$ then $\Delta_1 > \Delta_2$.
If $E(r^2)_t > E(r^2)_+$ then $\Delta_1 < \Delta_2$.

This cycle therefore represents a stable equilibrium.

(ii) *Results*

We have applied this model to the following cases:

$T_1 = 1$ or 2 generations with $N_e = 500$,
$T_2 = 4$ or 8 generations with $N_e = \infty$.

Thus, the annual cycle includes between 5 and 10 generations.

Fig. 2*a*, *b* represent two extreme situations for the recombination frequencies ($c = 0{\cdot}01$ and $c = 0{\cdot}25$, respectively) between the two loci considered. In both cases $T_1 = 2$, $T_2 = 8$ and $N_e = 500$.

The effective size $\overline{N_e}$ for a whole annual cycle is given by the expression:

$$\frac{1}{\overline{N_e}} = \left[T_1 \cdot \frac{1}{N_e} + T_2 \cdot \frac{1}{\infty}\right] \frac{1}{T_1+T_2}$$

so

$$\overline{N_e} = \frac{N_e(T_1+T_2)}{T_1} = 2500.$$

When the recombination frequency is small ($c = 0{\cdot}01$), $E(r^2)$ increases very slowly, oscillating at the same time. After about 225 generations, the cycle of $E(r^2)$ is almost stable. Its value fluctuates between an upper limit $E(r^2)_+$ previously defined in the model (6), and a lower limit $E(r^2)_-$ (6′). The two corresponding effective sizes, calculated from (1) are respectively $N_{e_+} = 2308$ and $N_{e_-} = 2715$. These two values are very close to $\overline{N_e} = 2500$. We also represent in Fig. 2*a* the evolution of $E(r^2)$ obtained from (3) and the value $E(r^2)_{\overline{N_e}}$ expected at equilibrium from (1) for a population of constant size equal to $\overline{N_e}$ (= 2500).

When the recombination frequency is high ($c = 0{\cdot}25$) the stable cycle for $E(r^2)$ is reached after a few cycles (Fig. 2*b*). Starting at the first cycle, $E(r^2)$ fluctuates between two values very close to $E(r^2)_+$ and $E(r^2)_-$. The ratio $E(r^2)_+/E(r^2)_-$ is 65 times greater than when $c = 0{\cdot}01$ (78·5 and 1·2, respectively). N_{e_+} and N_{e_-} are equal to 727 and 72792, respectively. These values are very different from $\overline{N_e}$ (= 2500).

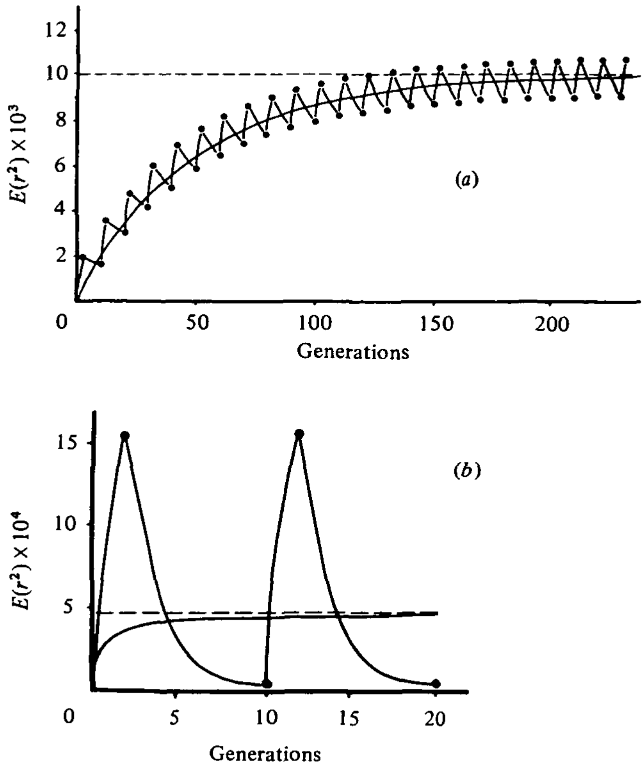

Fig. 2. Evolution of $E(r^2)$ in populations with cyclical size fluctuations (·——·: two generations with $N_e = 500$ then eight generations with $N_e = \infty$; ---$E(r^2)$ from equation (1) when constant $N_e = 2500$; —— $E(r^2)$ from equation (2) when $N_e = 2500$). (*a*) c = 0·01; (*b*) c = 0·25.

The values of the parameters previously described: $E(r^2)_+$, $E(r^2)_-$, N_{e+}, N_{e-} and $E(r^2)_{\overline{N_e}}$ were calculated for different values of c (0·01, 0·05, 0·10, 0·25) when $N_e = 500$ (Table 2). We have also indicated the generation, G 1 % fluctuating, from which the upper value of $E(r^2)$ observed during a cycle approaches $E(r^2)_+$ within 1% and G 1% constant corresponding to a population of constant size $\overline{N_e}$ (from 4). These results show that the number of generations necessary to reach equilibrium is the same, plus or minus the length of a cycle, with either fluctuating or constant population size. It depends only on c. So, in the analysed situations, G 1% fluctuating is independent of T_1 and T_2. In addition, the ratios

$$\frac{N_{e+}}{N_e} \quad \text{and} \quad \frac{N_{e-}}{N_e}$$

depend mainly on c and are only faintly influenced by T_1 and T_2.

4. Discussion

When $E(r^2)$ reaches equilibrium, the characteristics of its fluctuations are different, according to the intensity of the linkage between the loci. This must be taken into account when analysing linkage disequilibria observed in natural populations of *Drosophila*.

When c is small (0·01 – 0·05), the relative variation of $E(r^2)$ is small in the course of a cycle. Thus the effective size calculated at equilibrium from $E(r^2)$, at each point in the annual cycle, is close to the annual harmonic mean ($\overline{N_e}$) of the population sizes.

Consequently, the r values obtained from tightly linked loci can be used to estimate N_e from Hill's formula (1981). If $\overline{N_e}$ has been estimated from data other than linkage disequilibrium (i.e. allelic drift, lethal allelism, ...) this $\overline{N_e}$ value can be used to test drift versus selection as follows. When sampling n gametes in isolated equilibrium populations of constant effective size, we expect, due to the drift effect:

$$E(r) = 0$$

$$V(r) = E(r^2) \simeq \frac{(1-c)^2+c^2}{2N_e c(2-c)} + \frac{1}{n} = \frac{f(c)}{2N_e} + \frac{1}{n}$$

(Weir & Hill 1980).

Table 2. *Square of the correlation of gene frequencies expected by random drift within populations with fluctuating size*

c	T_1	T_2	$E(r^2)_+ \times 10^3$	N_{e+}	$E(r^2)_- \times 10^3$	N_{e-}	$E(r^2)_{\overline{N_e}} \times 10^3$	$\overline{N_e}$	$G_{1\%}$ constant	$G_{1\%}$ fluctuating
0·01	1	4	10·36	2400	9·50	2620	9·95	2500	227	226
	2	4	17·15	1440	15·80	1566	16·48	1500	225	223
	1	8	6·01	4154	5·12	4809	5·55	4500	228	226
	2	8	10·77	2308	9·17	2715	9·95	2500	227	221
0·05	1	4	2·49	2057	1·65	3103	2·05	2500	45	41
	2	4	4·13	1237	2·74	1868	3·41	1500	45	43
	1	8	1·66	3243	0·73	7023	1·14	4500	45	37
	2	8	2·96	1968	1·30	3940	2·05	2500	45	41
0·10	1	4	1·53	1712	0·66	3979	1·05	2500	22	21
	2	4	2·52	1043	1·08	2435	1·75	1500	22	25
	1	8	1·18	2236	0·22	12069	0·58	4500	22	19
	2	8	2·06	1275	0·38	6896	1·05	2500	22	21
0·25	1	4	1·06	1078	0·11	10781	0·46	2500	8	6
	2	4	1·61	709	0·16	7097	0·76	1500	8	9
	1	8	1·01	1136	0·01	114285	0·25	4500	8	1
	2	8	1·57	727	0·02	72792	0·46	2500	8	2

T_1 generations with $N_e = 500$ then T_2 generations with $N_e = \infty$.

https://doi.org/10.1017/S0016672300024952 Published online by Cambridge University Press

Table 3. *Critical values $(\chi^2_\alpha)^{N_e}$ when sampling 600 gametes in a population of effective size N_e.*

		N_e	
c	α	1000	10000
0·01	0·05	60·58	5·54
	0·01	104·60	9·57
	0·001	171·01	15·63
0·10	0·05	8·81	4·34
	0·01	15·21	7·49
	0·001	24·85	12·23
0·25	0·05	5·49	4·00
	0·01	9·47	6·91
	0·001	15·47	11·29

c: recombination frequency, $\chi^2_{\alpha 1 ddl}$ critical values are 3·84, 6·64 and 10·83 for, $\alpha = 0{\cdot}05$, 0·01 and 0·001 respectively.

We deduced the critical values $(\chi^2_\alpha)^{N_e}$ for the Chi square used to test the significance of an observed disequilibrium. These values take into account the disequilibrium produced by random drift in a population of size $\overline{N_e}$:

$$(\chi^2_\alpha)^{N_e} = \chi^2_\alpha\left[\frac{n}{2\overline{N_e}}f(c)+1\right] \tag{7}$$

(χ^2_α = critical value, for the confidence level α, of the 1.d.f. χ^2 distribution).

Table 3 shows some values of $(\chi^2_\alpha)^{N_e}$ for the confidence level $\alpha = 0{\cdot}05$, 0·01 and 0·001 when $n = 600$ and $N_e = 1000$ and 10000. Such sizes encompass the N_{e+} and N_{e-} values estimated for *Drosophila* populations previously cited. When loci are tightly linked, due to random drift there is a large increase in the critical values $(\chi^2_\alpha)^{N_e}$ in all points of the cycle. It must be pointed out that when loci are tightly linked, the stable cycle is reached after a long time. Further, as Prout (1973) showed, migration or transient subdivision of the population can modify and eventually reduce for some time the linkage disequilibrium produced by genetic drift.

When c is large, disequilibrium is established only after a few generations of bottleneck and disappears accordingly during the burst. Correspondingly, for a given cycle, $E(r^2)$ shows large relative variations; it is no longer related to the harmonic mean $\overline{N_e}$. The population size which must be used to test drift versus selection lies between N_{e+} and N_{e-}, and depends on the point of the cycle at which the disequilibrium is observed. A comparison of values presented in Table 3 with those in Table 2 shows that when two loci are far apart, the random drift effect becomes negligible at the end of the demographic burst. For example, for a single generation of bottleneck ($N_e = 500$) and four generations of burst ($N_e = \infty$), with $c = 0{\cdot}25$ we obtain $(\chi^2_{0\cdot05})^{N_{e-}} \simeq 4{\cdot}00$. But with the same parameters we have at the beginning of the burst $(\chi^2_{0\cdot05})^{N_{e+}} \simeq 5{\cdot}49$. Then the annual demographic cycle of the population must be known in order to accurately predict the gametic disequilibrium expected by drift. Unfortunately, these demographic data are not generally available.

Despite these difficulties, it may be possible to compare the observed dynamics of the linkage disequilibrium with the dynamics expected from our model:

– $E(r^2)$ must have the same value from year to year when measured at the same point of the cycle, whether the loci are tightly linked or not;

– $E(r^2)$ must fluctuate during one year, especially when loci are loosely linked. The expected negative correlation between observed linkage disequilibrium and recombination frequency must then be the highest at the end of the demographic burst. Too few data are available to verify this second point; for this reason, the results are unclear (Langley *et al.* 1977).

It is clearly the case that the most precise approach is the use of experimental lab-based populations. By controlling their effective size it is possible to accurately investigate the relative effects of selection and drift in linkage disequilibrium.

Appendix

Equilibrium value of $E(r^2)$ when the population size fluctuates.

– During T_1 generations, the size is finite, so we have:

$$\Delta_1 = E(r^2)_t - E(r^2)_{t-T_1} \tag{A 1}$$

with

$$E(r^2)_t = \frac{1-\left[\left(1-\frac{1}{2N_e}\right)(1-c)^2\right]^t}{[1+(2N_e-1)(2c-c^2)]}$$

writing

$$A = \left(1-\frac{1}{2N_e}\right)(1-c)^2$$

we have

$$E(r^2)_t = \frac{1}{2N_e} + A\,E(r^2)_{t-1}$$

then

$$E(r^2)_t = \frac{1}{2N_e}\frac{(1-A^{T_1})}{(1-A)} + A^{T_1}E(r^2)_{t-T_1}. \tag{A 5}$$

– During T_2 generations, the size is infinite, and we have

$$\Delta_2 = E(r^2)_t - E(r^2)_{t+T_2} \tag{A 3}$$

with

$$E(r^2)_{t+T_2} = (1-c)^{2T_2}E(r^2)_t. \tag{A 4}$$

At equilibrium $\Delta_1 = \Delta_2$ so, from (A 2) and (A 4)

$$E(r^2)_t = E(r^2)_+ = \frac{1-A^{T_1}}{2N_e(1-A)(1-A^{T_1}(1-c)^{2T_2})}.$$

https://doi.org/10.1017/S0016672300024952 Published online by Cambridge University Press

References

Begon, M. (1977). The effective size of a natural *Drosophila subobscura* population. *Heredity* **38**, 13–18.

Hedrick, P. W. (1983). *Genetics of Populations*. Boston: Science Books International.

Hill, W. G. (1976). Non-random association of neutral linked genes in finite populations. In *Population Genetics and Ecology*, (ed. S. Karlin and E. Nevo) pp. 339–376. New York: Academic Press.

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38**, 209–216.

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

Langley, C. H. (1977). Non-random associations between allozymes in natural populations of *Drosophila melanogaster*. In *Measuring Selection in Natural Populations* (ed. F. B. Christiansen and T. M. Fenchel), pp. 265–273, Berlin: Springer-Verlag.

Langley, C. H., Ito, K. & Voelker, R. A. (1977). Linkage disequilibrium in natural populations of *Drosophila melanogaster*, seasonal variation. *Genetics* **86**, 447–454.

McInnis, D. O., Schaffer, H. E. & Mettler, L. E. (1982). Field dispersal and population sizes of native *Drosophila* from North Carolina. *American Naturalist* **119–3**, 319–330.

Montchamp-Moreau, C. (1985). Analyse du déséquilibre gamétique dans des populations naturelles et expérimentales de *Drosophila simulans*. Thèse d'Etat, Université Paris 6, Paris.

Mukai, T., Mettler, L. E. & Chigusa, S. I. (1971). Linkage disequilibrium in a local population of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences U.S.A.* **68**, 1065–1069.

Prout, T. (1973). Appendix to Mitton, J. B. & Koehn, R. C. Population genetics of marine pelecypods. III. Epistasis between functional related isoenzymes in *Mytilus edulis*. *Genetics* **73**, 487–496.

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

Weir, B. S. & Hill, W. G. (1980). Effect of mating structure on variation in linkage disequilibrium. *Genetics* **95**, 477–488.