Heredity (1974), **33** (2), 229-239

ESTIMATION OF LINKAGE DISEQUILIBRIUM IN RANDOMLY MATING POPULATIONS

WILLIAM G. HILL
Institute of Animal Genetics, Edinburgh EH9 3JN

Received 15.xi.73

Summary

The degree of linkage disequilibrium, D, between two loci can be estimated by maximum likelihood from the frequency of diploid genotypes in a sample from a random-mating population. Haploid genotypes can be identified directly in some species from a sample of chromosomes extracted from the population and made homozygous, or by test crossing. The maximum likelihood estimators of D are described, with examples, for both methods, including the cases where both loci are codominant and one or both are dominant.

The efficiencies of the methods are compared when $D = 0$: If both loci are codominant the estimate of D has the same variance,

$$V(\hat{D}) = p(1-p)q(1-q)/N,$$

from a sample of N identified diploids as from N identified haploid types, where p and q are the gene frequencies; therefore the diploid method is more efficient in practice since less labour is required. With dominance at either locus $V(\hat{D})$ is lower for samples of the same size using the haploid method if the dominant alleles are at high frequency.

1. Introduction

Now that it is possible to use starch gel electrophoresis to type the same individual for several different polymorphic loci, some of which may be linked, associations between the frequencies of alleles at two or more loci are being studied. Allard and his group with plants (*e.g.* Allard, Babbel, Clegg and Kahler, 1972), several groups with *Drosophila* (Prakash and Lewontin, 1968, 1971; Kojima, Gillespie and Tobari, 1970; Zouros and Krimbas, 1972; Charlesworth and Charlesworth, 1973; Franklin, 1973) and Webster (1973) in salamander have found such linkage disequilibrium, although in the *Drosophila* cases usually associated with a chromosomal inversion. Mukai, Mettler and Chigusa (1971) however, did not find any associations among linked genes in *D. melanogaster*. Sinnock and Sing (1972*a*, *b*) found some evidence of disequilibrium among loci in man, but these loci were not known to be linked. A group in this laboratory (D. A. Briscoe, J. M. Malpica and A. Robertson) are also doing similar analyses on *Drosophila* populations which will be reported subsequently. In view of the number of these studies being undertaken, whatever their possible contribution to population genetics, it seems worth while to investigate some of the statistical problems of estimation of linkage disequilibrium.

The degree of linkage disequilibrium can be estimated directly from the genotypic frequencies in a sample of individuals taken from the population. The coupling and repulsion heterozygotes can not normally be distinguished, however, and if either locus is dominant (which for electrophoretic variants usually implies the existence of null alleles) other classes are also confounded.

An alternative approach which is only applicable to *Drosophila* is the isolation of single chromosomes from natural populations against crossover-suppressor stocks. These single chromosomes may thus be made homozygous before establishing their allelic content (*e.g.* Kojima *et al.*, 1970; Mukai *et al.*, 1971). An equivalent procedure is to test cross individuals against a marker stock. The technique of chromosome isolation, in particular, involves much more labour per observation, *i.e.* a diploid or a haploid (chromosome) individual identified, and we may ask whether this labour is justified in terms of improved accuracy of estimation of the disequilibrium. This question was raised with me by Dr D. A. Briscoe, and an attempt is made to provide an answer in this paper by predicting the sampling variance of estimates of disequilibrium obtained by the alternative methods.

It is recommended that maximum likelihood (ML) estimation be used in any such analysis of data, for even where numerical solutions are required these can be obtained easily using relevant computer programs. (A program specifically for handling the analysis of designs discussed in this paper is available from the author.) Whilst the main results of this paper are predictions of sampling variances, it has been extended to include methods of estimation, together with examples to help the experimentalist. For the case of two codominant loci an ML procedure has been given by Bennett (1965), but an alternative method is presented here; and the ML solution for two dominant loci has been given by Turner (1968) and Cavalli-Sforza and Bodmer (1971) but is repeated for completeness.

2. Analysis

The population is assumed to be random mating and to be in Hardy-Weinberg equilibrium at each locus. At the first locus there are two alleles, A and a, with frequencies p and $1-p$, and at the second locus two alleles, B and b, with frequencies q and $1-q$. The frequencies of the chromosome types AB, Ab, aB and ab are f_{11}, f_{12}, f_{21} and f_{22} respectively, and the linkage disequilibrium, D, is given by

$$D = f_{11}f_{22} - f_{12}f_{21} = f_{11} - pq.$$

The frequencies are summarised in table 1 (*a*). We shall alternate between use of the (f_{ij}) and (p, q, D) to define the model, according to which gives the more condensed form of results, and utilise the property that the same transformation applied to the ML estimators $(\hat{f}_{ij})$ gives the ML estimators $(\hat{p}, \hat{q}, \hat{D})$, and vice versa (*e.g.* Elandt-Johnson, 1971, p. 298).

We consider three models in which diploid individuals are identified: both A and B codominant (where the ML estimation procedure is outlined more fully), A codominant and B dominant, and then both A and B dominant. Finally we consider the case where haploids are identified, either by isolation of chromosomes or by appropriate test crossing. In all cases the numbers of each type identified are assumed to be multinomially distributed.

(i) *Diploid identification: both* A *and* B *to dominant*

When all three genotypes can be identified at both loci, but the coupling and repulsion heterozygotes can not be separated, there are nine phenotypic classes. The expected frequencies (y_{ij}, where $y_{11} = f_{11}^2$, for example), the

TABLE 1

Expected frequencies and observed numbers for different genetic models

(a) *Definitions of frequencies; chromosome identification*

Chromosome	*AB*	*Ab*	*aB*	*ab*	Total
Expected frequency	f_{11} $pq+D$	f_{12} $p(1-q)-D$	f_{21} $(1-p)q-D$	f_{22} $(1-p)(1-q)+D$	
Observed numbers	n_{11}	n_{12}	n_{21}	n_{22}	n

(b) A *codominant*, B *codominant: expected frequencies* (y_{ij})

	BB	*Bb*	*bb*
AA	f_{11}^2	$2f_{11}f_{12}$	f_{12}^2
Aa	$2f_{11}f_{21}$	$2f_{11}f_{22}+2f_{12}f_{21}$	$2f_{12}f_{22}$
aa	f_{21}^2	$2f_{21}f_{22}$	f_{22}^2

(c) A *codominant*, B *codominant: observed numbers*

	BB	*Bb*	*bb*	Total
AA	N_{11}	N_{12}	N_{13}	$N_{1\cdot}$
Aa	N_{21}	N_{22}	N_{23}	$N_{2\cdot}$
aa	N_{31}	N_{32}	N_{33}	$N_{3\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	$N_{\cdot 3}$	N

Derived totals

$X_{11} = 2N_{11}+N_{12}+N_{21}$; $X_{12} = 2N_{13}+N_{12}+N_{23}$
$X_{21} = 2N_{31}+N_{21}+N_{31}$; $X_{22} = 2N_{33}+N_{23}+N_{32}$

(d) A *codominant*, B *dominant: observed numbers* (expected frequencies are obtained by summing columns 1 and 2 in (b))

	B–	*bb*	Total
AA	N_{11}	N_{12}	$N_{1\cdot}$
Aa	N_{21}	N_{22}	$N_{2\cdot}$
aa	N_{31}	N_{32}	$N_{3\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	N

(e) A *dominant*, B *dominant: observed numbers* (expected frequencies are obtained by summing rows 1 and 2 and columns 1 and 2 in (b))

	B–	*bb*	Total
A–	N_{11}	N_{12}	$N_{1\cdot}$
aa	N_{21}	N_{22}	$N_{2\cdot}$
Total	$N_{\cdot 1}$	$N_{\cdot 2}$	N

observed numbers (N_{ij}) and some functions of them (X_{ij}) are given in table 1 (*b*) and (*c*). The logarithm of the likelihood (L) is

$$\log L = \sum_{i,j=1}^{3} N_{ij} \log y_{ij} + \text{constant}$$
$$= \sum_{ij} X_{ij} \log f_{ij} + N_{22} \log (f_{11}f_{22}+f_{12}f_{21}) + \text{constant}, \quad (1)$$

which has been given by Bennett (1965). The parameter estimates can be obtained by differentiating log L, and finding the zero values by trial and error, as Bennett (1965) showed. Alternatively we can use the "gene-counting" method of Ceppellini, Siniscalco and Smith (1955) and described

by Elandt-Johnson (1971, p. 400), which gives identical solutions to maximum likelihood. Since it is applied in this paper to chromosomes we shall call it the " chromosome counting " method, and it appears to have been used by Webster (1973). Each phenotypic class is apportioned into the expected number of each chromosome type; thus an *AABb* individual comprises one *AB* and one *Ab* chromosome, while *AaBb* individuals have an expected proportion of $f_{11}f_{22}/(f_{11}f_{22}+f_{12}f_{21})$ *AB* and *ab* chromosomes and $f_{12}f_{21}/(f_{11}f_{22}+f_{12}f_{21})$ *Ab* and *aB* chromosomes. The equations are then

$$\begin{aligned}\hat{f}_{ij} &= [X_{ij}+N_{22}\hat{f}_{11}\hat{f}_{22}/(\hat{f}_{11}\hat{f}_{22}+\hat{f}_{12}\hat{f}_{21})]/2N, \quad i=j \\ \hat{f}_{ij} &= [X_{ij}+N_{22}\hat{f}_{12}\hat{f}_{21}/(\hat{f}_{11}\hat{f}_{22}+\hat{f}_{12}\hat{f}_{21})]/2N, \quad i\neq j.\end{aligned} \tag{2}$$

By summing equations (2) we find that the gene frequency estimates are given by the marginal frequencies:

$$\begin{aligned}\hat{p} &= \hat{f}_{11}+\hat{f}_{12} = (X_{11}+X_{12}+N_{22})/2N = (N_{1.}+\tfrac{1}{2}N_{2.})/N, \\ \hat{q} &= \hat{f}_{11}+\hat{f}_{21} = (N_{.1}+\tfrac{1}{2}N_{.2})/N;\end{aligned} \tag{3}$$

but $\hat{D}$ has no explicit solution. A suitable method is to replace $\hat{f}_{12}$ by $\hat{p}-\hat{f}_{11}$, $\hat{f}_{21}$ by $\hat{q}-\hat{f}_{11}$ and $\hat{f}_{22}$ by $1-\hat{p}-\hat{q}+\hat{f}_{11}$ in the equation (2) for $\hat{f}_{11}$, to give a single equation

$$\hat{f}_{11} = \{X_{11}+N_{22}\hat{f}_{11}(1-\hat{p}-\hat{q}+\hat{f}_{11})/[\hat{f}_{11}(1-\hat{p}-\hat{q}+\hat{f}_{11})+(\hat{p}-\hat{f}_{11}) \times(\hat{q}-\hat{f}_{11})]\}/2N. \tag{4}$$

The only unknown in (4) is $\hat{f}_{11}$, and it is solved by choosing a value of $\hat{f}_{11}$ for the right-hand side, evaluating the expression and using this as the next trial value of $\hat{f}_{11}$. The iterative process is continued until stability is reached and $\hat{D}$ obtained as $\hat{f}_{11}-\hat{p}\hat{q}$. A suitable starting value for iteration is

$$\hat{f}_{11} = \frac{1}{4N}(X_{11}-X_{12}-X_{21}+X_{22})+\tfrac{1}{2}-(1-\hat{p})(1-\hat{q}), \tag{5}$$

which is obtained by assuming that the genotype frequency of the double heterozygote class is exactly that computed from the other classes.

The sampling variances of the ML estimators can be obtained for large samples in the usual way from the inverse of the matrix of expected values of the log likelihood. Let $t_1 = p$, $t_2 = q$ and $t_3 = D$. From (1)

$$\frac{\partial^2 \log L}{\partial t_k \partial t_l} = \sum_{i,j=1}^{3} N_{ij}\left(y_{ij}\frac{\partial^2 y_{ij}}{\partial t_k \partial t_l} - \frac{\partial y_{ij}}{\partial t_k}\frac{\partial y_{ij}}{\partial t_l}\right)\Big/ y_{ij}^{\,2}$$

We have $E(N_{ij}) = Ny_{ij}$, and note that $\sum_{i,j} \partial^2 y_{ij}/\partial t_k \partial t_l = 0$, since $\sum_{i,j} y_{ij} = 1$ (Elandt-Johnson, 1971, p. 317). Letting

$$m_{kl} = -E(\partial^2 \log L/\partial t_k \partial t_l)$$

we obtain

$$m_{kl} = N\sum_{i,j=1}^{3} \frac{\partial y_{ij}}{\partial t_k}\frac{\partial y_{ij}}{\partial t_l}/y_{ij}. \tag{6}$$

The variance-covariance matrix of the estimates is given by M^{-1}, where M is a 3×3 matrix with elements m_{kl}. The necessary derivatives, $\partial y_{ij}/\partial t_k$, are given in table 2, and these can be used in (6).

TABLE 2

Derivatives of genotypic frequencies (y_{ij}) *for diploid model with both loci codominant with respect to the frequency of* A(p), B(q) *and* D

	BB	*Bb*	*bb*
		$\frac{1}{2}\partial y_{ij}/\partial p$	
AA	qf_{11}	$qf_{12}+(1-q)f_{11}$	$(1-q)f_{12}$
Aa	$q(f_{21}-f_{11})$	$q(f_{22}-f_{12})+(1-q)(f_{21}-f_{11})$	$(1-q)(f_{22}-f_{12})$
aa	$-qf_{21}$	$-qf_{22}-(1-q)f_{21}$	$-(1-q)f_{22}$
		$\frac{1}{2}\partial y_{ij}/\partial q$	
AA	pf_{11}	$p(f_{12}-f_{11})$	$-pf_{12}$
Aa	$pf_{21}+(1-p)f_{11}$	$p(f_{22}-f_{21})+(1-p)(f_{12}-f_{11})$	$-pf_{22}-(1-p)f_{12}$
aa	$(1-p)f_{21}$	$(1-p)(f_{22}-f_{21})$	$-(1-p)f_{22}$
		$\frac{1}{2}\partial y_{ij}/\partial D$	
AA	f_{11}	$f_{12}-f_{11}$	$-f_{12}$
Aa	$f_{21}-f_{11}$	$f_{11}-f_{12}-f_{21}+f_{22}$	$f_{12}-f_{22}$
aa	$-f_{21}$	$f_{21}-f_{22}$	f_{22}

The above method for finding the variances and covariances provides a simple way of computing $V(\hat{D})$ in this codominant-codominant model, and is useful in the other models for parameters which do not have explicit ML estimators. However, for those that do, a direct approach can be used; for example $\hat{p}$ is given by (3) and is binomially distributed. We obtain

$$V(\hat{p}) = p(1-p)/2N, \quad V(\hat{q}) = q(1-q)/2N \tag{7}$$

$$\text{cov}(\hat{p}, \hat{q}) = D/2N, \quad \text{cov}(\hat{p}, \hat{D}) = (1-2p)D/2N, \quad \text{cov}(\hat{q}, \hat{D}) = (1-2q)D/2N.$$

The variances in (7) are, of course, the same as for a single gene situation. When $D = 0$, we see that the covariances are zero, and also find that the equation (6) simplifies, to give

$$V(\hat{D}) = p(1-p)q(1-q)/N. \tag{8}$$

More generally, for $D \neq 0$ it is clear that $V(\hat{D})$ can not be expressed as a linear function of the terms obtained subsequently in (22) for the haploid model.

In any experiment only estimates of p, q and D are available, and these have to be used instead of the parameters in table 2, (6) and (7). Alternatively the second derivatives of the log likelihood can be obtained numerically and used as the elements of M.

Using the large sample assumption of normality, a test for $D = 0$ can be made using (8). This is equivalent to the likelihood ratio test for, under the null hypothesis that $D = 0$, the quantity given by

$$k = -2 \log [L(p, q, D)/L(p, q)] \tag{9}$$

has the chi-square distribution asymptotically with 1 d.f., where $L(p, q, D)$,

$L(p, q)$ are the likelihoods (1) obtained by fitting only the specified parameters. It can be shown that, ignoring terms of order D^3 or higher,

$$k = N\hat{D}^2/\hat{p}(1-\hat{p})\hat{q}(1-\hat{q})$$
$$= N\hat{r}^2, \tag{10}$$

where r^2 is the squared correlation of gene frequencies. The chi-square test proposed by Sinnock and Sing (1972*b*) is equivalent except theirs is obtained by using goodness-of-fit rather than likelihood arguments.

(ii) *Diploid identification:* A *codominant*, B *dominant*

There are now six phenotypes, with the observed numbers shown in table 1 (d) and expected frequencies obtained by summing the appropriate frequencies for B codominant in table 1 (b) (*i.e.* columns 1 and 2). The likelihood equation can be written down using these frequencies but, for solving the equation, we again adopt the chromosome counting method. The equations are (ignoring " hats " on estimates)

$$f_{11} = \frac{1}{2N}\left[\frac{2N_{11}(f_{11}^2+f_{11}f_{12})}{f_{11}^2+2f_{11}f_{12}} + \frac{N_{21}(f_{11}f_{21}+f_{11}f_{22})}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}}\right] \tag{11a}$$

$$f_{12} = \frac{1}{2N}\left[\frac{2N_{11}f_{11}f_{12}}{f_{11}^2+2f_{11}f_{12}} + 2N_{12} + \frac{N_{21}f_{12}f_{21}}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}} + N_{22}\right] \tag{11b}$$

$$f_{21} = \frac{1}{2N}\left[\frac{N_{21}(f_{11}f_{21}+f_{12}f_{21})}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}} + \frac{2N_{31}(f_{21}^2+f_{21}f_{22})}{f_{21}^2+2f_{21}f_{22}}\right] \tag{11c}$$

$$f_{22} = \frac{1}{2N}\left[\frac{N_{21}f_{11}f_{22}}{f_{11}f_{21}+f_{11}f_{22}+f_{12}f_{21}} + N_{22} + \frac{2N_{31}f_{21}f_{22}}{f_{21}^2+2f_{21}f_{22}} + 2N_{32}\right]. \tag{11d}$$

Summing equations (11a) and (11b), we find that for the codominant gene, A, the estimated frequency, $\hat{p}$, is given by the marginal frequencies,

$$\hat{p} = (N_{1.} + \tfrac{1}{2}N_{2.})/N. \tag{12}$$

But we notice that the sum of (11a) and (11c) does not simplify in this way, so we obtain the rather surprising result that the ML estimator of gene frequency of a dominant gene suspected of being in disequilibrium with a codominant gene is not given by the marginal frequencies. Similarly, $\hat{D}$ is not obtained explicitly, so we need to retain two of the equations (11), for example (11a) and (11c) and express $\hat{f}_{12}$ and $\hat{f}_{22}$ in terms of $\hat{p}$, $\hat{f}_{11}$ and $\hat{f}_{21}$. These equations are iterated to obtain a solution for $\hat{f}_{11}$ and $\hat{f}_{21}$ and consequently $\hat{q}$ and $\hat{D}$. Since $\hat{q}$ is unlikely to depart far from the estimate given by the marginal frequencies, a suitable starting value for the iterations is obtained using $1-\hat{q} = (N_{.2}/N)^{\frac{1}{2}}$ and $\hat{f}_{22} = (N_{32}/N)^{\frac{1}{2}}$.

The sampling variances of all of the estimators can be found as before, using (6), but with the subscript j taking only two values. The appropriate frequencies y_{ij} and derivatives $\partial y_{ij}/\partial t_k$ are given by summing the first two columns in tables 1 (b) and 2, respectively. Explicit formulae for the variances

or covariances involving the codominant gene A can be given, however. These are the same as when B is codominant also, *i.e.*

$$V(\hat{p}) = p(1-p)/2N \tag{13}$$

$$\text{cov}\,(\hat{p}, \hat{q}) = D/2N, \quad \text{cov}\,(\hat{p}, \hat{D}) = (1-2p)D/2N.$$

When $D = 0$, all covariances are zero and

$$V(\hat{q}) = q(2-q)/4N, \quad V(\hat{D}) = p(1-p)q(2-q)/2N; \tag{14}$$

and we note that $V(\hat{q})$ is that for a single dominant gene.

The likelihood ratio criterion (9) for testing $D = 0$ is, approximately,

$$k = 2N\hat{D}^2/[\hat{p}(1-\hat{p})\hat{q}(2-\hat{q})]. \tag{15}$$

(iii) *Diploid identification: both* A *and* B *dominant*

There are only four phenotypic classes (table 1 (e)), so the ML estimators are the obvious ones, namely

$$\hat{p} = 1-(N_{2.}/N)^{\frac{1}{2}}, \quad \hat{q} = 1-(N_{.2}/N)^{\frac{1}{2}} \quad \text{and} \quad \hat{f}_{22} = (N_{22}/N)^{\frac{1}{2}} \tag{16}$$

giving

$$\hat{D} = (N_{22}/N)^{\frac{1}{2}}-(N_{2.}N_{.2})^{\frac{1}{2}}/N \tag{17}$$

(Turner, 1968; Cavalli-Sforza and Bodmer, 1971).

The sampling variances of the estimators can be found using (6), but after summing the first two rows and columns in tables 1 (b) and 2. The only explicit formulae not involving a large number of terms are

$$V(\hat{p}) = p(2-p)/4N, \quad V(\hat{q}) = q(2-q)/4N \tag{18}$$

and the estimators are correlated. When $D = 0$, $\hat{p}$, $\hat{q}$ and $\hat{D}$ are uncorrelated and

$$V(\hat{D}) = p(2-p)q(2-q)/4N. \tag{19}$$

The likelihood ratio criterion (9) is, approximately,

$$k = 4N\hat{D}^2/[\hat{p}(2-\hat{p})\hat{q}(2-\hat{q})], \tag{20}$$

which differs from that given by Cavalli-Sforza and Bodmer (1971, p. 285) in that a term in D^3 has been ignored.

(iv) *Haploid identification*

A sample of n chromosomes is taken from the population and identified by an appropriate method (*e.g.* by test crossing or making an isogenic line) with the observed numbers shown in table 1 (a). The observed chromosome frequencies are their ML estimators, *i.e.* $\hat{f}_{ij} = n_{ij}/n$, so

$$\hat{p} = n_{1.}/n, \quad \hat{q} = n_{.1}/n, \quad \hat{D} = (n_{11}n_{22}-n_{12}n_{21})/n^2. \tag{21}$$

The sampling variances of the estimators can be found directly from the multinomial distribution, with that for $V(\hat{D})$ being obtained from formulae given by Hill and Robertson (1968):

$$\left.\begin{aligned} &V(\hat{p}) = p(1-p)/n, \quad V(\hat{q}) = q(1-q)/n \\ &V(\hat{D}) = [p(1-p)q(1-q)+(1-2p)(1-2q)D-D^2]/n \\ &\text{cov}\,(\hat{p}, \hat{q}) = D/n, \quad \text{cov}\,(\hat{p}, \hat{D}) = (1-2p)D/n, \quad \text{cov}\,(\hat{q}, \hat{D}) = (1-2q)D/n \end{aligned}\right\} \tag{22}$$

We note that, when $D = 0$, the estimates are uncorrelated and

$$V(\hat{D}) = p(1-p)q(1-q)/n. \tag{23}$$

The likelihood ratio criterion (9) is, approximately,

$$k = n\hat{r}^2$$

and k is the usual chi-square statistic in a 2×2 contingency table (Hill and Robertson, 1968).

3. Example

Suitable data for diploid models have been given by Cleghorn (1960) on the *M/N*, *S/s* blood systems in man, and these were also used by Bennett (1965). The data are given in table 3 (a), and we note that both loci are codominant.

Table 3(a)

Cleghorn's data on numbers observed for the M/N *and* S/s *loci and the designation of the alleles in this paper*

Genotype	Designation	*SS* *BB*	*Ss* *Bb*	*ss* *bb*	Total
MM	*AA*	57	140	101	298
MN	*Aa*	39	224	226	489
NN	*aa*	3	54	156	213
	Total	99	418	483	1000

$X_{11} = 293$ $X_{12} = 568$ $X_{21} = 99$ $X_{22} = 592$

Data in 3(a) *reallocated:*

3(b) B *dominant*

	B–	*bb*	Total
AA	197	101	298
Aa	263	226	489
aa	57	156	213
Total	517	483	1000

3(c) A *and* B *dominant*

	B–	*bb*	Total
A–	460	327	787
aa	57	156	213
Total	517	483	1000

(i) A *and* B *codominant*

From (3), $\hat{p} = 0{\cdot}5425$ and $\hat{q} = 0{\cdot}3080$, and with these values inserted into (4) we obtain the chromosome counting formula for iteration

$$\hat{f}_{11} = 0{\cdot}1465 + 0{\cdot}112\hat{f}_{11}(0{\cdot}1495 + \hat{f}_{11})/(0{\cdot}16709 - 0{\cdot}701\hat{f}_{11} + 2\hat{f}_{11}^2)$$

The starting value (5) is $\hat{f}_{11} = 0{\cdot}23791$. After 11 iterations successive values of $\hat{f}_{11}$ differed by less than 10^{-8}, giving a solution of $\hat{f}_{11} = 0{\cdot}2370976$; and from that $\hat{D} = 0{\cdot}0700076$, agreeing with Bennett's value of $\hat{D} = 0{\cdot}07001$. The estimates, together with their standard errors and correlations (computed by replacing the parameter values by their estimates in (6), or in (7) where possible), are summarised in table 4. More figures than are significant are shown for comparison with estimates from the other models. We see in table 4 that D differs significantly ($P < 0{\cdot}001$) from zero, using the likelihood ratio (9) or the approximation to it (10). As Bennett (1965) showed with this data, there is a good fit to Hardy-Weinberg equilibrium: the residual chi-square (from likelihood ratio test) after fitting p, q and D is 3·3 with

5 d.f.). Bennett (1965) gave the standard error of $\hat{D}$ as 0·00596; this value differs slightly from that in table 4, largely because Bennett ignored covariances between the estimators: he assumed $V(\hat{D}) = m_{33}^{-1}$, which he computed by differentiating the likelihood directly.

TABLE 4

Results of analysis of data of table 3

Loci codominant		*A*, *B*	*A*	—
dominant		—	*B*	*A*, *B*
Estimates	p	0·54250	0·54250	0·53848
	q	0·30800	0·30474	0·30502
	D	0·07001	0·07048	0·07422
Standard errors	p	0·01114	0·01114	0·01403
	q	0·01032	0·01135	0·01137
	D	0·00617	0·00712	0·00763
Correlations	p, q	0·3044	0·2788	0·2596
	p, D	−0·0433	−0·0378	−0·1170
	q, D	0·2111	0·1656	0·1725
$-2\ log\ [L(p, q, D)/L(p, q)]$		101·9	79·7	69·3
k (equation 20)		92·6	77·5	54·2

(ii) A *codominant*, B *dominant*

We assume *BB* and *Bb* can not be distinguished in the data in table 3 (a), so by summing the first and second columns we obtain table 3 (b). For gene *A*, $\hat{p} = 0{\cdot}5425$ as before (12). Using (11a) and (11c) and writing $\hat{f}_{12} = \hat{p} - \hat{f}_{11}, \hat{f}_{22} = 1 - \hat{p} - \hat{f}_{21}$ we have

$$\hat{f}_{11} = \frac{0{\cdot}106872}{1{\cdot}0850 - \hat{f}_{11}} + \frac{0{\cdot}060161\hat{f}_{11}}{K}, \quad \hat{f}_{21} = \frac{0{\cdot}026077}{0{\cdot}9150 - \hat{f}_{21}} + \frac{0{\cdot}071338\hat{f}_{21}}{K},$$

where $K = 0{\cdot}4575\,\hat{f}_{11} + 0{\cdot}5425\,\hat{f}_{21} - \hat{f}_{11}\hat{f}_{21}$. Suitable starting values for the iterations are $q = 1 - \sqrt{(483/1000)} = 0{\cdot}3050$ from the marginal totals and $f_{22} = \sqrt{(156/1000)} = 0{\cdot}3950$, equivalent to $\hat{D} = 0{\cdot}0770$, $\hat{f}_{11} = 0{\cdot}2425$, $\hat{f}_{21} = 0{\cdot}0625$. After 22 iterations both $\hat{f}_{11}$ and $\hat{f}_{21}$ changed by less than 10^{-8} in successive iterations, giving, as final values $\hat{q} = 0{\cdot}30474$ and $\hat{D} = 0{\cdot}07048$ (table 4). Notice that the ML estimate of q departs slightly from that computed from marginal frequencies. The data still show a highly significant departure from linkage equilibrium.

(iii) *Both* A *and* B *dominant*

Further reduction of table 3 (a) gives the necessary data for the example in table 3 (c). The ML estimates from (16) and (17) and their sampling variances are listed in table 4. The departure from linkage equilibrium is shown to be significant.

Since the computations are so simple, no example for the chromosomal analysis will be given.

4. DISCUSSION AND CONCLUSIONS

The main object of this analysis was to compare the relative efficiencies of the alternative methods of estimating D. Formally, we measure efficiency as $E = [V(\hat{D})$ from n haploids$]/[V(\hat{D})$ from N diploids$]$, so that $E > 1$ if

the diploid method gives a lower variance for the same number of observations, and $E<1$ if the haploid method gives a lower variance. We recall that a single observation is either the identification of one diploid individual, or the identification of the allelic content of one chromosome, which may be one observation on an isogenic line or one test cross progeny.

The case of most interest is where the population is near linkage equilibrium, or we wish to test the null hypothesis that $D = 0$, and fortunately this has given us the simplest solutions. The results can be summarised as follows:

Haploid identification:

$$V(\hat{D}) = p(1-p)q(1-q)/n = nV(\hat{p})V(\hat{q}).$$

Diploid identification:

$$V(\hat{D}) = 4NV(\hat{p})V(\hat{q})$$

and the efficiencies for the different models are related to the accuracy of gene frequency estimation:

A, B codominant	$E = 1$
A codominant, B dominant	$E = (1-q)/(1-\frac{1}{2}q)$
A, B dominant	$E = [(1-p)/(1-\frac{1}{2}p)][1-q)/(1-\frac{1}{2}q)]$

If both loci are codominant, typical for biochemical variants, we see that $\hat{D}$ has the same variance when estimated from diploids directly as from a sample of the same size of extracted chromosomes or test crosses, which requires much more labour. Some examples have also been computed for $D \neq 0$ for the double codominant case, with p, $q = 0{\cdot}1$, $0{\cdot}25$, $0{\cdot}5$ and $q<p$. It turns out that $E \leqq 2$, only approaching $E = 2$ with $p = q = 0{\cdot}5$ and $D \to \pm 0{\cdot}25$, but $E>1$ over most combinations of p, q and D. The only cases with $E<1$ are listed below, together with the lowest values attained:

$(p, q) = (0{\cdot}1, 0{\cdot}1)$, $-0{\cdot}010 < D < 0$, minimum $E = 0{\cdot}74$
$(p, q) = (0{\cdot}25, 0{\cdot}1)$, $-0{\cdot}018 < D < 0$, minimum $E = 0{\cdot}91$
$(p, q) = (0{\cdot}25, 0{\cdot}25)$, $-0{\cdot}031 < D < 0$, minimum $E = 0{\cdot}97$.

Therefore, even when $D \neq 0$, the diploid method is likely to give better estimates, $\hat{D}$, for a given input of labour.

Returning to the case of $D = 0$ and considering dominant genes, we see that the diploid and haploid models have similar efficiencies if the dominant genes are at low frequency; but if they are at high frequency, the chromosome or test cross method may be worth while, just as it would be if we were interested in estimating gene frequencies.

This analysis has been restricted to two loci, but some preliminary studies have been carried out with more. It appears that, if all loci are codominant, the efficiency of the diploid relative to haploid method of estimating the disequilibrium between c loci, under the null hypothesis of equilibrium, is equal to 2^{2-c}. This equals 1 for 2 loci, $\frac{1}{2}$ for 3 loci, $\frac{1}{4}$ for 4 loci, and so on. Thus for three loci the haploid method would be justified only if it required less than twice the labour, per individual scored, than the diploid method. It is interesting to note that the diploid method is twice as efficient for estimating gene frequencies, since two genes are scored per individual, and this efficiency of 2 is obtained by setting $c = 1$ in the above formula. In effect we lose half

the information on D in the two locus diploid cases because we cannot distinguish between the coupling and repulsion heterozygotes, and a greater proportion with more loci when there are several multiple heterozygote classes.

5. References

ALLARD, R. W., BABBEL, G. R., CLEGG, M. T , AND KAHLER, A. L. 1972. Evidence for coadaptation in *Avena barbata*. *Proc. Nat. Acad. Sci. U.S.A., 69*, 3043-3048.

BENNETT, J. H. 1965. Estimation of the frequencies of linked gene pairs in random mating populations. *Amer. J. Hum. Genet., 17*, 51-53.

CAVALLI-SFORZA, L. L., AND BODMER, W. F. *The Genetics of Human Populations*. Freeman, San Francisco.

CEPPELLINI, R., SINISCALCO, M., AND SMITH, C. A. B. 1955. The estimation of gene frequencies in a random-mating population. *Ann. Eugen., 20*, 97-115.

CHARLESWORTH, B., AND CHARLESWORTH, D. 1973. A study of linkage disequilibrium in populations of *Drosophila melanogaster*. *Genetics, 73*, 351-359.

CLEGHORN, T. E. 1960. MNSs gene frequencies in English blood donors. *Nature, 187*, 701.

ELANDT-JOHNSON, R. C. 1971. *Probability Models and Statistical Methods in Genetics*. Wiley, New York.

FRANKLIN, I. R. 1973. Selection, migration and genetic drift in natural populations of *D. melanogaster*. *Genetics, 74*, s84.

HILL, W. G., AND ROBERTSON, A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet., 38*, 226-231.

KOJIMA, K., GILLESPIE, J., AND TOBARI, Y. N. 1970. A profile of *Drosophila* species' enzymes assayed by electrophoresis. I. Number of alleles, heterozygosities and linkage disequilibrium in glucose-metabolising systems and some other enzymes. *Biochem. Genet., 4*, 626-637.

MUKAI, T., METTLER, L. E., AND CHIGUSA, S. I. 1971. Linkage disequilibrium in a local population of *Drosophila melanogaster*. *Proc. Nat. Acad. Sci. U.S.A., 69*, 2474-2478.

PRAKASH, S., AND LEWONTIN, R. C. 1968. A molecular approach to the study of genetic heterozygosity in natural populations. III. Direct evidence of coadaptation in gene arrangements of *Drosophila*. *Proc. Nat. Acad. Sci. U.S.A., 59*, 398-405.

PRAKASH, S., AND LEWONTIN, R. C. 1971. A molecular approach to the study of genetic heterozygosity in natural populations. V. Further direct evidence of coadaptation in inversions of *Drosophila*. *Genetics, 69*, 405-408.

SINNOCK, P., AND SING, C. F. 1972*a*. Analysis of multilocus genetic systems in Tecumseh, Michigan. I. Definition of the data set and tests for goodness-to-fit to expectations based on gene, gamete and single locus phenotypic frequencies. *Amer. J. Hum. Genet., 24*, 381-392.

SINNOCK, P., AND SING, C. F. 1972*b*. Analysis of multilocus genetic systems in Tecumseh, Michigan. II. Consideration of the correlation between non-alleles in gametes. *Amer. J. Hum. Genet., 24*, 393-415.

TURNER, J. R. G. 1968. On supergenes. II. The estimation of gametic excess in natural populations. *Genetica, 39*, 82-93.

WEBSTER, T. P. 1973. Adaptive linkage disequilibrium between two esterase loci of a salamander. *Proc. Nat. Acad. Sci. U.S.A., 70*, 1156-1160.

ZOUROS, E., AND KRIMBAS, C. B. 1972. Linkage disequilibrium in natural populations of *Drosophila subobscura* maintained by selection. *Genetics, 71*, s71.