

PERSON

the change in ranking would be a
esent example was chosen specifically
ly as well to be aware of the circum-
cur, namely

ts is small;
nment interaction;
ironments is either well above or well

the Division of Plant Industry, New South
on to use the data from their annual wheat
L. Jinks for his valuable comments on a first

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

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

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} $(1-p)(1-q) + D$	n
Observed numbers	n_{11}	n_{12}	n_{21}	n_{22}	

(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}	$N_{1.}$
Aa	N_{21}	N_{22}	N_{23}	$N_{2.}$
aa	N_{31}	N_{32}	N_{33}	$N_{3.}$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$	N

Derived totals

$$X_{11} = 2N_{11} + N_{12} + N_{21}; \quad X_{12} = 2N_{12} + N_{13} + N_{23}$$

$$X_{21} = 2N_{21} + N_{22} + N_{31}; \quad X_{22} = 2N_{22} + N_{23} + N_{32}$$

(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}	$N_{1.}$
Aa	N_{21}	N_{22}	$N_{2.}$
aa	N_{31}	N_{32}	$N_{3.}$

ly applicable to *Drosophila* is the iso-
 natural populations against crossover-
 somes may thus be made homozygous
 (e.g. Kojima *et al.*, 1970; Mukai *et al.*,
 to test cross individuals against a
 chromosome isolation, in particular,
 ervation, *i.e.* a diploid or a haploid
 and we may ask whether this labour is
 y of estimation of the disequilibrium.
 Dr D. A. Briscoe, and an attempt is
 er by predicting the sampling variance
 l by the alternative methods.
 likelihood (ML) estimation be used
 here numerical solutions are required
 relevant computer programs. (A
 analysis of designs discussed in this
 Whilst the main results of this paper
 s, it has been extended to include
 examples to help the experimentalist.
 an ML procedure has been given by
 theod is presented here; and the ML
 given by Turner (1968) and Cavalli-
 ated for completeness.

random mating and to be in Hardy-
 At the first locus there are two
 and $1-p$, and at the second locus two
 and $1-q$. The frequencies of the
 f_{11} , f_{12} , f_{21} and f_{22} respectively are

ch gives identical solutions to maxi-
 this paper to chromosomes we shall
 method, and it appears to have been
 atypic class is apportioned into the
 e type; thus an *AABb* individual
 me, while *AaBb* individuals have an
 $(f_{12}f_{21})$ *AB* and *ab* chromosomes and
 osomes. The equations are then

$$[f_{22} + \hat{f}_{12}\hat{f}_{21}]/2N, \quad i = j \quad (2)$$

$$[f_{22} + \hat{f}_{12}\hat{f}_{21}]/2N, \quad i \neq j.$$

at the gene frequency estimates are

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	$-qf_{21}$	$-qf_{22} - (1-q)f_{21}$	$-(1-q)f_{22}$

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

$$\begin{aligned} k &= N\hat{D}^2/\hat{p}(1-\hat{p})\hat{q}(1-\hat{q}) \\ &= N\hat{r}^2, \end{aligned} \quad (10)$$

where r^2 is the squared correlation of gene frequencies. The chi-square test proposed by Sinnock and Sing (1972b) 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)

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

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

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

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

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

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

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 f_{12} and f_{22} in terms of \hat{p} , f_{11} and f_{21} . These equations are iterated to obtain a solution for f_{11} and 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 $f_{22} = (N_{32}/N)^{\frac{1}{2}}$.

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

d by fitting only the specified para-
g terms of order D^3 or higher,

$$-\hat{p})\hat{q}(1-\hat{q}) \tag{10}$$

ene frequencies. The chi-square test
s equivalent except theirs is obtained
elihood arguments.

A codominant, B dominant

th the observed numbers shown in
obtained by summing the appropriate
1 (b) (i.e. columns 1 and 2). The
wn using these frequencies but, for
the chromosome counting method.
n estimates)

$$\left[\frac{f_{11}f_{21} + f_{11}f_{22}}{f_{11}f_{22} + f_{12}f_{21}} \right] \tag{11a}$$

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

$$\left[\frac{f_{21}^2 + f_{21}f_{22}}{f_{21} + 2f_{21}f_{22}} \right] \tag{11c}$$

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

end that for the codominant gene,
y the marginal frequencies,

$$N_{e.})/N. \tag{12}$$

U (1c) does not simplify in this way,
ut that the ML estimator of gene
d of being in disequilibrium with a
arginal frequencies. Similarly, \hat{D} is
tain two of the equations (11). for

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

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,

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

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	Ss	ss	Total
		BB	Bb	bb	
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				3(c) A and B dominant			
	B-	bb	Total		B-	bb	Total
AA	197	101	298	A-	460	327	787
Aa	263	226	489	aa	57	156	213
aa	57	156	213	Total	517	483	1000
Total	517	483	1000				

(i) A and B codominant

From (3), $\hat{p} = 0.5425$ and $\hat{q} = 0.3080$, and with these values inserted into (4) we obtain the chromosome counting formula for iteration

$$\hat{f}_{11} = 0.1465 + 0.112\hat{f}_{11}(0.1495 + \hat{f}_{11}) / (0.16709 - 0.701\hat{f}_{11} + 2\hat{f}_{11}^2)$$

The starting value (5) is $\hat{f}_{11} = 0.23791$. After 11 iterations successive values of \hat{f}_{11} differed by less than 10^{-8} , giving a solution of $\hat{f}_{11} = 0.2370976$; and from that $\hat{D} = 0.0700076$, agreeing with Bennett's value of $\hat{D} = 0.07001$. 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 ($P < 0.001$) 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.).
differs
variance
comput

We
so by s
A, \hat{p}
 $\hat{f}_{12} =$

\hat{f}_{11}

where
iterati
 $\hat{f}_{22} =$
 $\hat{f}_{21} =$
in suc
(table
puted
depar

Ft
in tab
variab
showr
Si
analy

T
of the
as E

M. G. HILL

estimates are uncorrelated and

$$-p)q(1-q)/n. \quad (23)$$

approximately,

$$= n\hat{p}^2$$

in a 2×2 contingency table (Hill and

EXAMPLE

have been given by Cleghorn (1960) and these were also used by Bennett

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		<i>A, B</i>	<i>A</i>	—
	dominant	—	<i>B</i>	<i>A, B</i>
Estimates	<i>p</i>	0.54250	0.54250	0.53848
	<i>q</i>	0.30800	0.30474	0.30502
	<i>D</i>	0.07001	0.07048	0.07422
Standard errors	<i>b</i>	0.01114	0.01114	0.01403

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 $p, q = 0.1, 0.25, 0.5$ and $q < p$. It turns out that $E \leq 2$, only approaching $E = 2$ with $p = q = 0.5$ and $D \rightarrow \pm 0.25$, 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:

$(p, q) = (0.1, 0.1),$	$-0.010 < D < 0,$	minimum $E = 0.74$
$(p, q) = (0.25, 0.1),$	$-0.018 < D < 0,$	minimum $E = 0.91$
$(p, q) = (0.25, 0.25),$	$-0.031 < D < 0,$	minimum $E = 0.97.$

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

variance for the same number of observations gives a lower variance. We recall the identification of one diploid individual, content of one chromosome, which may be or one test cross progeny.

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