Genet. Res., Camb. (1981), **38**, *pp.* 209–216
Printed in Great Britain

Estimation of effective population size from data on linkage disequilibrium[1]

By WILLIAM G. HILL

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27650, U.S.A. and Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN[2]

(*Received* 15 *April* 1980 *and in revised form* 23 *March* 1981)

SUMMARY

A method is proposed for estimating effective population size (N) from data on linkage disequilibrium among neutral genes at several polymorphic loci or restriction sites. The efficiency of the method increases with larger sample size and more tightly linked genes; but for very tightly linked genes estimates of N are more dependent on long-term than on recent population history. Two sets of data are analysed as examples.

1. INTRODUCTION

Linkage disequilibrium can, in theory, be produced by a number of factors operating separately or together: epistatic selection, migration, hitchhiking or random drift in finite populations (reviewed by Hedrick, Jain & Holden, 1978). Many estimates of linkage disequilibrium between allozymes at polymorphic loci have been made both in laboratory and natural populations, particularly in Drosophila; for references and summaries see, for example, Langley, Smith & Johnson (1978) and Hedrick *et al.* (1978). Usually little or no significant disequilibrium has been found, except where associated with inversions, or in laboratory populations maintained with small size (Langley *et al.* 1978; Laurie-Ahlberg & Weir, 1979), or between very closely linked loci such as those of the HLA system in man (e.g. Bodmer, 1973). Much additional data on disequilibrium are likely to come from direct studies on the DNA, from restriction enzyme sites or DNA sequences directly. Disequilibrium has been demonstrated between a nearby restriction site and the sickle-cell variant of the β-globin structural gene in man (Kan & Dozy, 1978).

For neutral genes or sites, the disequilibrium can be used to estimate population size because the variance of the disequilibrium or correlation of gene frequencies

[1] Paper No. 6428 of the Journal Series of the North Carolina Agricultural Research Service, Raleigh, North Carolina. This investigation was supported in part by NIH Research Grant No. GM 11546 from the National Institute of General Medical Sciences.

[2] Present address.

is a known function of population size (Langley, 1977; Laurie-Ahlberg & Weir, 1979). However, Langley did not go into much detail and Laurie-Ahlberg & Weir considered only unlinked loci. In this report the methodology is taken further.

2. ANALYSIS

(i) *Variance of disequilibrium*

Consider first a pair of loci, each with two alleles, neutral with respect to fitness. At the first locus the frequency of allele A is p and at the second the frequency of allele B is q. The linkage disequilibrium is $D = \text{freq.}(AB) - pq$ and the correlation of gene frequencies is $r = D/(p(1-p)q(1-q))^{\frac{1}{2}}$. The recombination fraction between the loci is c.

The population is assumed to be closed, and random mating with constant effective size N to be of sufficiently long standing that founder effects can be ignored. Therefore $E(D) = E(r) = 0$ and, although $V(D)$ declines as homozygosity increases, $V(r)$ approaches a steady value over populations remaining segregating (Hill and Robertson, 1968). Although $V(r)$ cannot be predicted exactly, it is well approximated as a ratio of moments. The variance of r comprises two parts; the first is a function of the effective population size (N) and c, and reflects the past finite population history; the second derives from sampling a limited sample of individuals from the population for estimating gene frequencies and disequilibrium, and is a function of the sample size (n). The latter contribution is the same whether a sample of n chromosomes are extracted and identified (possible in Drosophila) or whether n diploids are analysed in which coupling and repulsion heterozygotes cannot be distinguished (Hill, 1974); and the method of analysis of diploids, whether by maximum likelihood or the simpler Burrows' procedure, makes no difference to variance in random mating populations (Weir, 1979). Putting the terms together,

$$E(r^2) = V(r) = \frac{(1-c)^2+c^2}{2Nc(2-c)} + \frac{1}{n} \tag{1}$$

(Weir & Hill, 1980). Equation (1) strictly holds only for monoecious populations or dioecious with no permanent matings; for monogamy (1) should be increased by $1/(2Nc(2-c))$ (Weir & Hill, 1980), but this increment is mostly small and is not considered here. For very small Nc values (1) is not adequate because $E(r^2) \leqslant 1$. A fully relevant formulation is not available, but (1) almost certainly can be improved by replacing $2Nc(2-c)$ by $1+2Nc(2-c)$ in the denominator of the first term following Sved & Feldman (1973). In the following analyses Nc values near unity are not encountered, and this additional term is ignored.

Generalizing to l loci, there are $k = l(l-1)/2$ pairs; and let r_i, c_i and n_i, $i = 1, \ldots, k$, denote the correlation, recombination fraction and number of observations taken on the ith pair. Defining

$$\gamma_i = ((1-c_i)^2 + c_i^2)/(2c_i(2-c_i))$$

equation (1) reduces to

$$E(r_i^2) = \gamma_i/N + 1/n_i, \quad i = 1, \ldots, k. \tag{2}$$

https://doi.org/10.1017/S0016672300020553 Published online by Cambridge University Press

The variance–covariance structure of r_i^2 has been little studied, but some information will be needed here for pooling data from different pairs of loci. Some data are given by Hill (1977), and further analysis and discussion are given in the Appendix. These analyses can be summarized as follows (a) $V(r_i^2) \sim 2E^2(r_i^2)$ or $CV(r_i^2) \sim 1{\cdot}4$, such as might be expected if $r_i^2/E(r_i^2)$ were a chi-square deviate with one degree of freedom; (b) the correlations among the r_i^2 are zero or essentially so when different loci are involved, e.g. r_{AB}^2 and r_{CD}^2, and not more than about 0·25 when there are loci in common, e.g. r_{AC}^2, r_{BC}^2. Thus in the following analysis the r_i^2 will be assumed to be uncorrelated with variance double their expected value. Although more efficient estimates of N could be obtained if the full variance–covariance structure of the r_i^2 were known, in view of the small correlations the exact error structure is not critical since the figures are being used mainly to weight different unbiased estimates of population size or its inverse.

(ii) *Estimation of population size*

Equation (2) can be used to estimate N by replacing $E(r_i^2)$ by its observed value. To avoid inverting r_i^2 before pooling over loci, it is better to estimate N^{-1} by

$$\hat{\alpha}_i = (r_i^2 - 1/n_i)/\gamma_i \tag{3}$$

for the ith pair of loci. The $\hat{\alpha}_i$ can be combined most efficiently using their variance–covariance structure which, from the preceding arguments, is given approximately by

$$V(\hat{\alpha}_i) = 2E^2(r_i^2)/\gamma_i^2, \quad \mathrm{Cov}\,(\hat{\alpha}_i, \hat{\alpha}_j) = 0, \quad i \neq j.$$

Thus the weighted estimate of N^{-1} is, using (2) and (3),

$$\hat{N}^{-1} = \frac{\Sigma_i(\hat{\alpha}_i/V(\hat{\alpha}_i))}{\Sigma_i(1/V(\hat{\alpha}_i))} = \frac{\Sigma_i\left[\gamma_i(r_i^2 - 1/n_i)\Big/\left(\frac{\gamma_i}{N}+\frac{1}{n_i}\right)^2\right]}{\Sigma_i\left[1\Big/\left(\frac{1}{N}+\frac{1}{\gamma_i n_i}\right)^2\right]} \tag{4}$$

with variance

$$V(\hat{N}^{-1}) = 1/\sum_i(1/V(\hat{\alpha}_i)) = 2/\sum_i(1/N + 1/\gamma_i n_i)^{-2}. \tag{5}$$

The estimate of N is obtained by inverting (4) and, to first-order terms,

$$V(\hat{N}) = N^4 V(\hat{N}^{-1}) = 2N^2/\sum_i\left(1+\frac{N}{\gamma_i n_i}\right)^{-2}. \tag{6}$$

Alternatively, the coefficient of variation is

$$CV(\hat{N}) = CV(\hat{N}^{-1}) = \left[2/\sum_i\left(1+\frac{N}{\gamma_i n_i}\right)^{-2}\right]^{\frac{1}{2}} \tag{7}$$

which, for illustration, if $n_i = n$, $\gamma_i = \gamma$ for all $i = 1, \ldots, k$, reduces to

$$CV(\hat{N}) = \left[1+\frac{N}{\gamma n}\right]\sqrt{\frac{2}{k}}. \tag{8}$$

https://doi.org/10.1017/S0016672300020553 Published online by Cambridge University Press

Equation (8) demonstrates that precise estimates of population size can be obtained only when the sample size, n, is large relative to the ratio $N/\gamma \sim 4Nc$. Some values are given in Table 1. For example, with 18 pairs of loci each around 0·1 map units apart (impossible, of course, but only for illustration), $\sqrt{2/k} = \frac{1}{3}$, so for a sample size of $n = N/10$, from Table 1 $CV(\hat{N}) = 5{\cdot}63/3 = 1{\cdot}9$. This is obviously a very imprecise estimate, but if n were about the same size as N, $CV(\hat{N}) = 0{\cdot}5$. Note that, even with very large sample sizes, $CV(\hat{N}) > \sqrt{2/k}$, or 0·33 for $k = 18$. Table 1 also shows that, unless the sample size is very much larger than the population size, unlinked loci ($c = 0{\cdot}5$) give little information.

Table 1. *Effect of sample size* (n) *and recombination fraction* (c) *on sampling error of estimation of effective population size* (N)

(For k pairs of loci, $CV(\hat{N}) = [1+N/(n\gamma)]\sqrt{(2/k)}$, where $\gamma = [(1-c)^2+c^2]/[2c(2-c)]$.)

		N/n			
c	γ	0·1	1	10	100
		$1+N/(n\gamma)$			
0·5	0·33	1·30	4·00	31·00	301·0
0·1	2·16	1·05	1·46	5·63	47·3
0·02	12·13	1·01	1·08	1·82	9·2
$\rightarrow 0$	$1/(4c)$	$1+0{\cdot}4c$	$1+4c$	$1+40c$	$1+400c$

3. EXAMPLES

The analysis will be illustrated for two data sets on *Drosophila melanogaster*. The first, published by Langley, Ito & Voelker (1977), is on the September collection from the wild made in North Carolina (Table 2). Second and third chromosomes were extracted and analysed from 198 flies, with six loci on the second and five on the third chromosomes. Thus for these data $n_i = 198$ for all i, and there are 15 pairs of loci on the second and 10 pairs on the third, giving $k = 25$. From such data there is no information on unlinked loci. The second set is from the Maine cage population of Langley *et al.* (1978), but original data were made available by C. H. Langley (Table 3). The following analysis is on pooled genotypic data from 634 to 756 flies, depending on the pair of loci, and values of r were estimated using Burrows' method (Cockerham & Weir, 1977). The data in Table 3 are only part of that given by Langley *et al.* (1978) and refer to a single rather than several time periods of sampling of the cage; also their published values were of $r_i/2$ rather than r_i. Whilst no inversions were found in the Maine cage, they were present in the wild population, and recombination fractions have been adjusted in Table 2 in proportion to the inversion heterozygosity, following Langley *et al.* (1977).

For the wild population (Table 2) the estimate of N^{-1} from (4) was negative. This implies that the best estimate of the population size is infinitely large, the disequilibrium observed being slightly less than that expected by chance (the

Table 2. *Isozyme data on extracted chromosomes from North Carolina wild population autumn collection* (*Langley* et al. *1977*)

(Sample size (n_i) = 198 for all pairs of loci. Values of c_i computed using observed karyotypic heterozygosities.)

Chromosome II: No.	1	2	3	4	5	6
Locus	α-Gpdh	Mdh	Adh	Dip-A	Hex-C	Amy
Allele frequency	0·823	0·975	0·742	0·914	0·934	0·929
Chromosome III: No.	7	8	9	10	11	
Locus	Est-6	Pgm	Odh	Lap-D	Acph	
Allele frequency	0·621	0·909	0·929	0·687	0·960	

Pair	c_i	γ_i	r_i	Pair	c_i	γ_i	r_i
1 2	0·058	4·0	0·094	4 6	0·081	2·7	−0·085
1 3	0·102	2·1	0·000	5 6	0·013	18·9	−0·073
1 4	0·113	1·9	0·000	7 8	0·029	8·3	−0·162
1 5	0·153	1·3	−0·069	7 9	0·049	4·7	0·069
1 6	0·158	1·2	−0·076	7 10	0·148	1·4	0·034
2 3	0·058	4·0	−0·095	7 11	0·152	1·3	0·051
2 4	0·072	3·1	−0·049	8 9	0·022	11·0	−0·042
2 5	0·123	1·7	−0·043	8 10	0·135	1·5	−0·016
2 6	0·130	1·6	−0·044	8 11	0·140	1·5	−0·076
3 4	0·019	12·8	−0·016	9 10	0·124	1·7	−0·016
3 5	0·085	2·6	0·030	9 11	0·129	1·6	0·043
3 6	0·094	2·3	0·063	10 11	0·011	22·4	0·027
4 5	0·072	3·1	−0·008				

Table 3. *Isozyme data on genotypes from Maine cage collection* (*Langley* et al., *1978*)

(Range of sample size (n_i) values 634 to 756.)

Chromosome	II			III			
Locus no.	1	2	3	4	5	6	7
Locus	α-Gpdh	Mdh	Adh	Est-6	Pgm	Est-C	Odh
Allele frequency*	0·930	0·927	0·589	0·579	0·920	0·945	0·969

Linked loci				Unlinked loci ($c_i = 0·5$, $\gamma_i = 0·33$)	
Pair	c_i	γ_i	r_i	Pair	r_i
1 2	0·071	3·2	−0·085	1 4	0·015
1 3	0·112	1·9	−0·049	1 5	0·001
2 3	0·057	4·0	−0·093	1 6	0·021
4 5	0·033	7·2	0·125	1 7	0·089
4 6	0·067	3·4	0·128	2 4	0·014
5 7	0·057	4·0	−0·028	2 5	−0·014
5 6	0·040	5·9	0·291	2 6	−0·063
5 7	0·029	8·3	−0·003	2 7	−0·083
6 7	0·013	18·9	0·247	3 4	−0·034
				3 5	0·026
				3 6	−0·024
				3 7	0·017

* Frequencies varied slightly from pair to pair because of different sample sizes.

https://doi.org/10.1017/S0016672300020553 Published online by Cambridge University Press

chi-square value for testing $r_i = 0$ from Table 2, $\Sigma n_i r_i^2$, is 19·3 with 25 d.f.). For the Maine cage population (Table 3), equation (4) gave $\hat{N}^{-1} = 0{\cdot}002755$ and thus an estimate of population size of 363. From (6), $SD(\hat{N}) = 170$ or $CV(\hat{N}) = 0{\cdot}47$, approximately. The estimate is below the census figure of 1000 (Langley *et al.* 1978), but has a very large standard error.

4. DISCUSSION

There are clearly difficulties in getting estimates of effective population size with much precision, as the examples show. As seen in Table 1, most information comes from tightly linked loci, but the problem with these is that any disequilibrium could be due to long-past founder effects or migration. (Indeed, if precision were not limiting, it would be possible to estimate N separately from loosely linked or unlinked loci and from tightly linked loci, and thereby get some information on population history, for example the presence of annual bottlenecks during overwintering.) In practice, it should be possible to obtain reasonably reliable estimates from laboratory populations, where sample sizes of the same order as effective population sizes can be obtained, or from isolated small colonies, but not from natural populations of effective size in the tens of thousands or more.

There is additional information available which could be exploited, notably on loci taken in threes, fours and so on. Some results for predicting random disequilibrium among multi-locus neutral models are available (Hill, 1976), but they lack the precision of the analyses of loci in pairs. There are also clearly correlations between, for example, r^2_{AB} and the relevant quantity involving loci A, B and C, so the extra information may not be in proportion to the number of additional terms, but the problem needs investigation. Similarly multiple allelic information needs to be incorporated; this might occur not only with isozymes but with multiple restriction sites in a region or with complete DNA sequencing.

Some assumptions and limitations of the analysis need to be emphasized. Populations are assumed to be random mating. (There was some evidence of non-random mating in the Maine Cage data (Langley *et al.* 1978), but this has been ignored in using the data for illustration). The variance–covariance structure of the r_i^2 is not known precisely; and if their correlations are higher than indicated in the examples of the appendix, variances of population size estimates are increased. Simulation (T. Maruyama and W. G. Hill, unpublished) shows that, if gene frequencies are very close to zero or one, the correlation (r) of gene frequencies conditional on the observed frequencies does not have a mean of zero. Thus biased estimates of population size would be obtained, but the relevant analysis has not yet been done. The data used in the examples have several gene frequencies in excess of 0·9, so such biases may have occurred in the analysis.

The basic model used has been of neutrality at each locus. If, however, loci are maintained segregating in the population at intermediate frequencies by heterozygote superiority, without epistasis, r has a mean of zero, and variance essentially the same as in the neutral case (Felsenstein, 1974; Avery, 1978). If the joint

distribution of the r_i were known more precisely in the neutral case, perhaps normal as suggested by the Appendix, it would be possible to test whether they were compatible with neutrality.

I am grateful to Dr C. H. Langley for access to unpublished data, to him and Dr B. S. Weir for helpful comments, and to an anonymous referee for valuable, albeit destructive, criticism of an earlier version.

REFERENCES

AVERY, P. J. (1978). The effects of finite population size on models of linked overdominant loci. *Genetical Research* **31**, 239–254.

BODMER, W. F. (1973). Population genetics of the HL-A system: retrospect and prospect. In *Histocompatibility Testing 1972* (ed. J. Dauset and J. Colombani). Copenhagen: Munksgaard.

COCKERHAM, C. C. & WEIR, B. S. (1977). Digenic descent measures for finite populations. *Genetical Research* **30**, 121–147.

FELSENSTEIN, J. (1974). Uncorrelated genetic drift of gene frequencies and linkage disequilibrium in some models of linked overdominant polymorphisms. *Genetical Research* **24**, 281–294.

HEDRICK, P., JAIN, S. & HOLDEN, L. (1978). Multilocus systems in evolution. *Evolutionary Biology* **11**, 101–184.

HILL, W. G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**, 229–239.

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. (1977). Correlation of gene frequencies between neutral linked genes in finite populations. *Theoretical Population Biology* **11**, 239–248.

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

KAN, Y. W. & DOZY, A. M. (1978). Polymorphism of DNA sequence adjacent to human β-globulin structural gene: relationship to sickle mutation. *Proceedings of the National Academy of Sciences of USA* **75**, 5631–5635.

LANGLEY, C. H. (1977). Nonrandom associations between allozymes in natural populations of *Drosophila melanogaster*. In *Lecture Notes in Biomathematics*. 19. *Measuring Selection in Natural Populations* (ed. F. B. Christiansen and T. M. Fenchel), pp. 265–273. New York: 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.

LANGLEY, C. H., SMITH, D. B. & JOHNSON, F. M. (1978). Analysis of linkage disequilibrium between allozyme loci in natural populations of *Drosophila melanogaster*. *Genetical Research* **32**, 215–229.

LAURIE-AHLBERG, C. & WEIR, B. S. (1979). Allozyme variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster*. *Genetics* **92**, 1295–1314.

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. (1979). Inferences about linkage disequilibrium. *Biometrics* **35**, 235–254.

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

APPENDIX

Variance–covariance structure of r_i^2 over pairs of loci

To investigate this, Monte Carlo simulation of random mating monoecious populations with four loci of two alleles were undertaken. Initial gene frequencies were taken as 0·5 at each locus, but simulation was continued until $E(r_i^2)$ in segregating populations remained constant. Several runs were made, but those used for illustration had map lengths in the ratios 1:3:2 between adjacent loci *AB*, *BC*, *CD* which also gave distances of 4:5:6 between *AC*, *BD* and *AD*. Results are given in the Appendix Table and these are typical of others. Note that the coefficients of variation are close to $\sqrt{2}$ (averaging 1·34 in these data). The standard errors of the correlations equal 0·06 for $r = 0$.

Appendix Table 1. *Coefficient of variation* (*CV*) *and correlations* (*corr*) *of* r^2 *from Monte Carlo simulation with* $N = 50$ *and map length l between loci specified*

(400 replicates initially, observations at generation 40 over replicates segregating at all four loci (296 and 292 replicates for runs with Nl = ½ and 2 respectively for AB.)

Loci	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*
Nl	½	2	3	1½	2½	1	2	8	12	6	10	4
						$CV(r_i^2)$ %						
	113	133	133	155	134	131	130	135	135	131	134	150
						Corr (r_i^2, r_j^2) %						
AB		9	11	1	3	9	—	4	10	8	4	−1
AC		—	15	27	2	2	—	—	9	5	−2	7
AD		—	—	1	23	5	—	—	—	0	18	14
BC		—	—	—	11	2	—	—	—	—	4	2
BD		—	—	—	—	9	—	—	—	—	—	−1

https://doi.org/10.1017/S0016672300020553 Published online by Cambridge University Press