Theor Appl Genet (1985) 70:491–496

A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients*

W. C. Black, IV and E. S. Krafsur
Department of Entomology, Iowa State University, Ames, IA 50011, USA

Received September 17, 1984; Accepted March 18, 1985
Communicated by A. Robertson

Summary. A FORTRAN program was written that calculates composite linkage disequilibrium coefficients from genotypic data. Chi-square tests determine whether coefficients calculated for allele and locus pairs are significantly greater than zero. A subroutine is provided that partitions the variance in linkage disequilibrium into within- and between-subpopulation components. Output obtained from analysis of allozyme data collected from natural subpopulations of the house fly (*Musca domestica* L.) are included to illustrate features of the program.

Key words: FORTRAN – Linkage disequilibrium – Electrophoresis – Breeding structure

Introduction

The term "linkage disequilibrium" was introduced in a discussion of two-locus selection theory by Lewontin and Kojima (1960) to refer to nonrandom association of alleles at two loci. Hill and Robertson (1968) and Ohta and Kimura (1969) demonstrated that disequilibrium may be indicative of genetic drift in populations. Lewontin (1974) stressed that linkage disequilibrium is theoretically a sensitive indicator of natural selection. Hill (1981) developed a statistical method for the estimation of effective population size from disequilibrium data. Recently Ohta (1982b) devised a method for partitioning the variance in disequilibrium into within- and between-subpopulation components.

Linkage disequilibrium has been monitored in field and laboratory populations of various species. Artificial disequilibrium was created in laboratory colonies of *Drosophila* by stocking them with many individuals homozygous at different loci (Clegg et al. 1980; Laurie-Ahlberg and Weir 1979). A natural "stocking effect" can occur in field populations after catastrophic reductions in population size. Disequilibrium generated through homozygote stocking is temporary and decays at a rate proportional to the amount of recombination between loci. Some intriguing examples of stable disequilibrium have been noted in field populations of salamanders, *Plethodon cinereus* (Webster 1973); *Drosophila subobscura* (Zouros and Krimbas 1973); mussels, *Mytilus edulis* (Mitton and Koehn 1973) and the Yanomama American Indians (Smouse and Neel 1977). Stable linkage disequilibria exist at the major histocompatibility gene complex in man and the mouse, *Mus musculus* (Bodmer 1979). Robertson and Hill (1983) noted that a stable disequilibrium also exists among restriction endonuclease sites in the β-globin gene cluster in humans (Orkin et al. 1982) and flanking regions of the *ADH* gene in *Drosophila* (Langley et al. 1982). But it is more typical to find little or no linkage disequilibrium in natural populations (Langley et al. 1974, 1977, 1978).

The calculation of linkage disequilibrium coefficients requires counting the number of times that the same two alleles appear together in individuals in a population. It is a tedious, error-prone, and very time consuming process when done by hand. A computer program was therefore developed that calculates two-locus linkage disequilibrium coefficients from genotypic data. The program is organized as a series of subroutines which may run by themselves or as part of a larger program "Genestats" (Black and Krafsur 1985). A printed copy of the program, a test data set, and instructions for its use are available at cost from the authors. The program, "Linkdis", is the only one currently available which calculates linkage disequilibrium coefficients for multiple alleles, performs various tests of significance and partitions the variance in disequilibrium coefficients.

* Journal Paper No. J-11345 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. 2411

Method

In the following discussion, the frequency of individuals formed by the union of gametes A_iB_j and A_kB_h is denoted P_{ij}^{kh} ($= P_{kh}^{ij}$), where i and k are alleles at locus A and j and h are alleles at locus B. Dots will denote sums of genotypic frequencies. So, the frequency with which A_iB_j gametes join with other gametes to form zygotes is denoted by:

$$P_{ij}^{\cdot\cdot} = \sum_k \sum_h P_{ij}^{kh}. \quad (1)$$

The frequency with which gametes bearing i and gametes bearing j join to form zygotes is denoted by:

$$P_{\cdot j}^{i\cdot} = \sum_k \sum_h P_{kj}^{ih}. \quad (2)$$

The frequencies of i and j are denoted p_i and p_j.

Assume for discussion that gametes can be sampled from a population and that all zygotes are formed by random union of gametes. Linkage equilibrium exists if the frequency of individuals bearing alleles i and j acquired from a single parent $P_{ij}^{\cdot\cdot}$ is equal to the frequency of individuals bearing i and j from different parents $P_{\cdot j}^{i\cdot}$.

If through epistatic natural selection (Lewontin 1974) or genetic drift (Ohta and Kimura 1969), i and j appear together more often in parents than is predicted by their independent frequencies, then equilibrium does not obtain. A statistical dependency D_{ij}^{w} (within individual disequilibrium component) (Cockerham and Weir 1977) exists between the alleles and

$$D_{ij}^{w} = P_{ij}^{\cdot\cdot} - P_{\cdot j}^{i\cdot}. \quad (3)$$

We assumed that union of gametes in the population is random. Therefore, with linkage equilibrium:

$$P_{ij}^{\cdot\cdot} = P_{\cdot j}^{i\cdot} = p_i p_j. \quad (4)$$

If linkage equilibrium does not exist then:

$$P_{ij}^{\cdot\cdot} \neq P_{\cdot j}^{i\cdot} = p_i p_j. \quad (5)$$

When, in addition, union of gametes is not random:

$$P_{\cdot j}^{i\cdot} \neq p_i p_j \quad (6)$$

and a second statistical dependency D_{ij}^{b} (between individual disequilibrium component) exists between i and j where (Cockerham and Weir 1977):

$$D_{ij}^{b} = P_{\cdot j}^{i\cdot} - p_i p_j. \quad (7)$$

D_{ij}^{b} and D_{ij}^{w} sum to the coefficient of linkage disequilibrium D_{ij}. Therefore:

$$D_{ij} = P_{ij}^{\cdot\cdot} - p_i p_j. \quad (8)$$

Cockerham and Weir observed that, for D_{ij} to be an accurate measure of linkage disequilibrium, there must be random union of gametes. Furthermore, estimation of D_{ij} requires that the frequency of A_iB_j gametes in the population be known. When working with genotypic data from natural populations, it is not reasonable to assume that random mating obtains, and gamete frequencies are seldom known. Cockerham and Weir suggested that a composite linkage disequilibrium coefficient, Δ_{ij}, be used instead of D_{ij}. Their formulation is:

$$\Delta_{ij} = D_{ij}^{w} + 2D_{ij}^{b} = P_{ij}^{\cdot\cdot} + P_{\cdot j}^{i\cdot} - 2p_i p_j. \quad (9)$$

Δ_{ij} contains $P_{ij}^{\cdot\cdot}$ and $P_{\cdot j}^{i\cdot}$, so that an estimate of the frequency of either is not required. Furthermore, Δ_{ij} is not biased by departures from random mating in the population. Note that, if D_{ij}^{b} is zero, $\Delta_{ij} = D_{ij}$. As calculated in "Linkdis" Δ_{ij} is unbiased. Weir (1979) showed with computer simulations that Δ_{ij} has a larger sampling variance than D_{ij} under random mating, but that under most conditions the differences are inconsequential. Therefore, Δ_{ij} is the best disequilibrium coefficient for analysis of field data.

Program

For illustrative purposes, "Linkdis" was used to analyse allozyme data from a series of house fly (*Musca domestica* L.) subpopulations in Iowa. Tables 1–4 are some of the outputs from this analysis.

The user lists the desired options (Table 1) and then provides the names of subpopulations, the names and number of loci examined, and the number of alleles at each locus. The genotype of each individual is then entered. The program sums the number, T_{ij} (Table 2), of times an allele i at one locus appears with another allele j at a second locus in the same individual. Where N is the number of individuals sampled,

$$T_{ij}/N = P_{ij}^{\cdot\cdot} + P_{\cdot j}^{i\cdot}. \quad (10)$$

The program uses composite frequencies because double heterozygotes cannot usually be separated into coupling and repulsion types. An unbiased estimate of Δ_{ij} can be calculated:

$$\Delta_{ij} = (N/(N-1))\,((T_{ij}/N) - 2p_i p_j) \quad \text{(Table 2)}. \quad (11)$$

A correlation coefficient R_{ij}, is determined (Weir 1979):

$$R_{ij} = \Delta_{ij}/((p_i(1-p_i) + C_i)\,(p_j(1-p_j) + C_j))^{1/2} \quad \text{(Table 2)}. \quad (12)$$

The "C" term corrects the correlation for departures from random mating.

$$C_i = H_{obs}(i) - p_i^2 \quad (12a)$$

where $H_{obs}(i)$ is the observed frequency of i homozygotes. A chi-square statistic and the corresponding level of significance are calculated for pairs of alleles to test the hypothesis that $R_{ij} = 0$ according to Weir's (1979) formula:

$$\chi^2_{(1\,d.f.)} = N R_{ij}^2 \quad \text{(Table 2)}. \quad (13)$$

Table 1. List of available options in "Linkdis"

Linkdis – options requested:
Linkage disequilibrium analysis
in subpopulations
condensed output
in total population
Ohta's analysis for subdivided populations
Table of allele frequencies

Table 2. Complete analysis of linkage disequilibrium at the *ADH* and *SOD* loci in a house fly subpopulation

Sheep 10/12/82

Analysis of disequilibrium between *ADH* and *SOD* loci (sample size = 50 individuals)

ADH Alleles		*SOD* Alleles	T_{ij}	Δ_{ij}	R_{ij}	Chi-square	*P*
1	and	1	1.00	0.00082	0.04299	0.09	0.7612
1	and	2	0.00	−0.00082	−0.04299	0.09	0.7612
2	and	1	67.00	0.01551	0.15341	1.18	0.2780
2	and	2	2.00	−0.01551	−0.15340	1.18	0.2781
3	and	1	15.00	−0.00735	−0.09321	0.43	0.5098
3	and	2	1.00	0.00735	0.09322	0.43	0.5098
4	and	1	12.50	0.00041	0.00539	0.00	0.9696
4	and	2	0.50	−0.00041	−0.00538	0.00	0.9696
5	and	1	0.50	−0.00939	−0.49434	12.22	0.0005
5	and	2	0.50	0.00939	0.49434	12.22	0.0005

Chi-square (4 d.f.) = 12.46
$P = 0.0143$

Table 3. Abbreviated table demonstrating results of analysis on each pair of loci for a house fly subpopulation

Analysis of linkage disequilibrium: Sheep 10/12/82

Loci compared	No. of comparisons	Chi-square	d.f.	*P*
AMY and *ADH*	50	19.02	16	0.2677
AMY and *PGM*: an allele at one or both loci is fixed				
AMY and *SOD*	50	2.82	4	0.5885
AMY and *ODH*	50	22.30	8	0.0044
AMY and *GOT*	50	5.14	8	0.7425
ADH and PGM: an allele at one or both loci is fixed				
ADH and *SOD*	50	12.46	4	0.0143
ADH and *ODH*	50	15.11	8	0.0570
ADH and *GOT*	50	4.40	8	0.8197
PGM and *SOD:* an allele at one or both loci is fixed				
PGM and *ODH*: an allele at one or both loci is fixed				
PGM and *GOT:* an allele at one or both loci is fixed				
SOD and *ODH*	50	2.76	2	0.2513
SOD and *GOT*	50	1.18	2	0.5549
ODH and *GOT*	50	0.65	4	0.9570

A second chi-square test indicates when significant disequilibrium exists between loci. The statistic is calculated and summed over all two-allele-interactions following Weir (1979):

$$\chi^2 = N \sum_i \sum_j (\Delta_{ij}^2 / p_i p_j) \quad \text{(Table 2)} \qquad (14)$$

with $(m-1)(n-1)$ degrees of freedom where m and n are the number of alleles at the loci.

Full output for each comparison will be lengthy. Upon request the program will print an abbreviated version (Table 3). Provided in the condensed output are the results of the chi-square analysis. Where significance is detected ($P \leqq 0.1$), the complete statistics (Table 2) are printed automatically.

Linkage disequilibrium in subdivided populations

We have discussed linkage disequilibrium in terms of a single population. In field studies allozyme data is often collected from a series of subpopulations. Linkage disequilibrium in a subdivided population can result in theory from epistatic natural selection within subpopulations (Lewontin 1974) and random drift among subpopulations (Ohta and Kimura 1969). Ohta (1982a) stated that linkage disequilibrium was too often attributed to epistatic natural selection. When migration among subpopulations is limited, random drift may generate different allele combinations and create disequilibrium. To differentiate between random and selective causes of disequilibrium, Ohta (1982a, b) devised a method for partitioning the variance of dilocus disequilibrium coefficients into within- and between-subpopulation components.

Ohta's theory is based on the hypothesis that when selection produces specific allele combinations, they should appear consistently among subpopulations. The variance in the observed frequency of allele combinations should therefore be small with respect to the variance expected on the hypothesis of random drift. Five "D-statistics" describe the variance in observed and expected frequencies of allele combinations. The statistics are calculated for locus pairs. We follow Ohta's notation and label the statistics with subscripts I, S, and T (I-Individuals, S-Subpopulations, T-Total population).

Table 4. Ohta's variance components of linkage disequilibrium for the entire house fly population

Variance components of linkage disequilibrium							
Loci compared			Within subpopulation components		Between subpopulation components		Total population component
			D^2_{IS}	D'^2_{IS}	D^2_{ST}	D'^2_{ST}	D^2_{IT}
AMY	and	*ADH*	0.00548	0.05319	0.00968	0.00066	0.05384
AMY	and	*PGM*	0.00091	0.04521	0.00960	0.00004	0.04525
AMY	and	*SOD*	0.00043	0.04129	0.00967	0.00019	0.04148
AMY	and	*GOT*	0.00126	0.04494	0.01081	0.00049	0.04543
ADH	and	*PGM*	0.00051	0.01332	0.00341	0.00003	0.01335
ADH	and	*SOD*	0.00035	0.01448	0.00362	0.00004	0.01451
ADH	and	*GOT*	0.00263	0.02730	0.00596	0.00036	0.02766
PGM	and	*SOD*	0.00036	0.00459	0.00101	0.00006	0.00465
PGM	and	*GOT*	0.00013	0.02904	0.00725	0.00005	0.02909
SOD	and	*GOT*	0.00022	0.02100	0.00570	0.00003	0.02103

The variance of disequilibrium in the total population, D^2_{IT} (Ohta 1982a), is,

$$D^2_{IT} = \sum_s \left(\sum_i \sum_j (T_{ij,s} - 2\bar{p}_i\bar{p}_j)^2\right)/x \quad \text{(Table 4)}, \qquad (15)$$

where $T_{ij,s}$ is the frequency that i and j appear together in individuals in subpopulation s, $\bar{p}_i$ and $\bar{p}_j$ are the weighted mean allele frequencies and x is the number of subpopulations. D^2_{IT} indicates when i and j appear together in individuals more often than is predicted by their independent frequencies in the overall population.

The variance of disequilibrium in subpopulations, D^2_{IS}, is calculated (Ohta 1982a) as:

$$D^2_{IS} = \sum_s \left(\sum_i \sum_j (T_{ij,s} - 2p_{is}p_{js})^2\right)/x \quad \text{(Table 4)} \qquad (16)$$

where p_{is} and p_{js} are the frequencies of i and j in subpopulation s. D^2_{IS} is $\Delta_{ij,s}$ squared and summed over all allele pairs. It is therefore a measure of the average squared disequilibrium in subpopulations.

The variance in the expected frequency of individuals bearing i and j, D^2_{ST} (Ohta 1982a) is,

$$D^2_{ST} = \sum_s \left(\sum_i \sum_j (p_{is}p_{js} - \bar{p}_i\bar{p}_j)^2\right)/x \quad \text{(Table 4)}. \qquad (17)$$

D^2_{ST} is a predictor of the variance of disequilibrium expected under the hypothesis of genetic drift. If subpopulations are differentiated, allele combinations will have been established independently among subpopulations and D^2_{ST} will be greater than D^2_{IS}. On the other hand, uniform epistatic selection among subpopulations will cause the frequency of allele combinations to converge so that D^2_{ST} is deflated and less than D^2_{IS}.

D'^2_{ST} is the variance of disequilibrium in the total population (Ohta 1982a),

$$D'^2_{ST} = \sum_i \sum_j (\bar{T}_{ij} - 2\bar{p}_i\bar{p}_j)^2 \quad \text{(Table 4)} \qquad (18)$$

where $\bar{T}_{ij}$ is the mean frequency with which i and j appear together in individuals in the population. Note that D'^2_{ST} is Δ_{ij} for the whole population squared and summed over all allele pairs. D'^2_{IS} is therefore an index of disequilibrium in the total population.

D'^2_{IS} is the variance in the observed frequency that i and j appear together in individuals in subpopulations. D'^2_{IS} (Ohta 1982a) is,

$$D'^2_{IS} = \sum_s \left(\sum_i \sum_j (T_{ij,s} - \bar{T}_{ij})^2\right)/x \quad \text{(Table 4)}. \qquad (19)$$

If random drift caused the variance in the frequency of allele combinations, D'^2_{IS} exceeds D'^2_{ST}. Uniform selection among subpopulations will increase D'^2_{ST} and deflate the variance D'^2_{IS}.

Ohta (1982a) showed that,

$$D^2_{IT} = D'^2_{IS} + D'^2_{ST} \quad \text{(Table 4)}. \qquad (20)$$

Equation (20) allows an evaluation of the relative contributions of random and selective processes to disequilibrium.

D-statistics identify disequilibrium patterns in a population. Three patterns of linkage disequilibrium can exist in a subdivided population. According to Ohta (1982a), these are nonsystematic disequilibrium, systematic disequilibrium, and unequal systematic disequilibrium. Disequilibrium is nonsystematic when caused by random drift. Under these conditions,

$$D^2_{IS} < D^2_{ST}, \quad \text{and} \quad D'^2_{IS} > D'^2_{ST}.$$

If disequilibrium arises through systematic epistatic selection for specific allele pairs in subpopulations then,

$$D^2_{IS} > D^2_{ST}, \quad \text{and} \quad D'^2_{IS} < D'^2_{ST}.$$

Unequal systematic disequilibrium arises if selection for specific allele pairs occurs only in a few sub-

populations. Then, D^2_{IS} is greater than D^2_{ST} because disequilibrium in subpopulations is greater than that expected by random drift. But D'^2_{IS} is greater than D'^2_{ST} because the variance in frequencies with which alleles appear together in individuals will be greater than disequilibrium in the total population.

"Linkdis" applied to a natural house fly population

Table 2 lists the number of joint occurrences of ADH alleles 1–5 with SOD alleles 1–2 T_{ij}, the corresponding coefficients of disequilibrium Δ_{ij}, the corrected correlation coefficients R_{ij}, and the chi-square values and significance levels. Examination of the *ADH-SOD* disequilibrium coefficients demonstrates that the only significant correlation was between *ADH* allele 5 and *SOD* alleles 1 and 2. These alleles were rare (frequency less than 0.02) in the population and so the probability was small of finding both alleles in a single fly.

The five loci studied in the same house fly subpopulation are summarily treated in Table 3, where the chi-square value, degrees of freedom and significance levels on each locus by locus comparison are printed. Significant disequilibrium was found between the *ADH* and *SOD* loci and the *AMY* and *ODH* loci.

Ohta's D-statistics are listed, locus by locus, in Table 4. These statistics represent samples from 5 spatially distributed fly subpopulations. Variances of disequilibrium among subpopulations (D^2_{ST} and D'^2_{IS}) were greater than the variances of disequilibrium within the overall population (D^2_{IS} and D'^2_{ST}). This demonstrates that correlations found between loci result from subpopulation differentiation and random drift rather than selection. D'^2_{IS} comprises 98–99% of the population variance of disequilibrium D^2_{IT}.

Discussion

Inference testing for linkage disequilibrium

Weir (1979) provided the most complete treatment to date of inference testing in linkage disequilibrium. The results of our analyses with house fly data suggest that chi-square tests were inadequate for multiple alleles especially when rare alleles were involved.

Chi-square values became inflated when expected frequencies were small. The problem would be alleviated by grouping the least frequent alleles into a single class, but information on specific alleles would be lost. The best alternative is to check allele frequencies (the program provides the option of printing allele frequencies) to determine when expected frequencies are small and likely to inflate the test statistic.

Linkage disequilibrium in subdivided populations

"Linkdis" was used to analyse data from 5 house fly subpopulations each sampled six times over the course of a season. Most statistically significant disequilibrium coefficients involved rare alleles and were probably artifacts. Disequilibrium in the total population was usually less than that found in subpopulations. D-statistics repeatedly revealed the same pattern of nonsystematic disequilibrium and no consistent pattern of linkage disequilibrium emerged. This agrees with earlier surveys of linkage disequilibrium in *Drosophila* (Langley et al. 1974, 1977, 1978).

Acknowledgement. We thank Bruce Weir for critically reviewing the manuscript.

References

Black WC, Krafsur ES (1985) A FORTRAN program for analysis of genotypic frequencies and description of the breeding structure of populations. Theor Appl Genet 70: 484–490

Bodmer WF (1979) Gene clusters and the HLA system. In: Human genetics: possibilities and realities, Ciba foundation series 66. Elsevier, North-Holland, pp 81–89

Clegg MT, Kidwell JF, Horch CR (1980) Dynamics of correlated genetic systems. 5. Rates of decay of linkage disequilibria in experimental populations of *Drosophila melanogaster*. Genetics 91:217–234

Cockerham CC, Weir BS (1977) Digenic descent measures for finite populations. Genet Res 30:121–147

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. Genet Res 38:209–216

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Langley CH, Tobari YN, Kojima K (1974) Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics 78:921–936

Langley CH, Ito K, Voelker RA (1977) Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Seasonal variation. Genetics 86:447–454

Langley CH, Smith DB, Johnson FM (1978) Analysis of linkage disequilibrium between allozyme loci in natural populations of *Drosophila melanogaster*. Genet Res 32: 215–229

Langley CH, Montgomery E, Quattlebaum W (1982) Restriction map variation in the *Adh* region of *Drosophila*. Proc Natl Acad Sci USA 79:5631–5635

Laurie-Ahlberg CC, Weir BS (1979) Allozymic variation and linkage disequilibrium in some laboratory populations of *Drosophila melanogaster*. Genetics 92:1295–1314

Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University Press, New York, pp 273–318

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Mitton JB, Koehn RK (1973) Population genetics of marine Pelecypods. 3. Epistasis between functionally related isozymes of *Mytilus edulis*. Genetics 73:487–496

Ohta T (1982a) Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc Natl Acad Sci USA 79:1940–1944

Ohta T (1982b) Linkage disequilibrium with the island model. Genetics 101:139–155

Ohta T, Kimura M (1969) Linkage disequilibrium due to random genetic drift. Genet Res 13:47–55

Orkin SH, Kazazian HH, Antonarakis SE, Goff SC, Boehm CD, Sexton JP, Waber PG, Giardina PJV (1982) Linkage of β-thalassemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. Nature 296:627–631

Robertson A, Hill WG (1983) Population and quantitative genetics of many linked loci in finite populations. Proc R Soc London, Ser B 219:253–264

Smouse PE, Neel JV (1977) Multivariate analysis of gametic disequilibrium in the Yanomama. Genetics 85:733–752

Webster TP (1973) Adaptive linkage disequilibrium between two esterase loci of a salamander. Proc Natl Acad Sci USA 70:1156–1160

Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 35:235–254

Zouros E, Krimbas CB (1973) Evidence for linkage disequilibrium maintained by selection in two natural populations of *Drosophila subobscura*. Genetics 73:659–674