On Measures of Gametic Disequilibrium

R. C. Lewontin

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138

Manuscript received June 24, 1988
Accepted July 23, 1988

ABSTRACT

Various measures have been proposed for characterizing the statistical association that arises between alleles at different loci. Hedrick has compared these measures with the standardized measure D' proposed by Lewontin on the grounds that this latter measure is independent of allele frequency. Although D' has the same range for all allelic frequencies, in fact, D' is not "independent" of allele frequency, and no measure with that general property is possible for the multilocus association problem. The insolubility of this problem arises from the ill-defined nature of general "association."

IT is now generally understood that, as a consequence of selection, random genetic drift, co-ancestry, or gene flow, alleles at different loci may not be randomly associated with each other in a population. While this effect is generally regarded as a consequence of linkage, even genes on different chromosomes may be held temporarily or permanently out of random association by forces of selection, drift and nonrandom mating.

For simplicity, let us consider only two loci with two alleles each, say, A,a and B,b. We denote, for the population frequencies of the four gametic types

g_{11} = frequency of AB gametes
g_{10} = frequency of Ab gametes
g_{01} = frequency of aB gametes
g_{00} = frequency of ab gametes

and

p_1 = frequency of allele A at locus A
q_1 = frequency of allele a at the A locus
p_2 = frequency of allele B at the B locus
q_2 = frequency of allele b at the B locus.

Then, if the loci are associated at random (*gametic equilibrium* or *linkage equilibrium*) we expect that $g_{11} = p_1p_2$, $g_{10} = p_1q_2$, etc.

If we write dynamical equations for changes in gametic frequencies as a consequence of recombination, selection, drift, gene flow or any combination of these, there appears in these equations the quantity

$$g_{11}g_{00} - g_{01}g_{10}.$$

[See, for example, LEWONTIN and KOJIMA (1960) or any subsequent paper on this subject.] This quantity has come to be symbolized by D, the *linkage disequilibrium* or *gametic disequilibrium* parameter. Obviously, if the alleles at the loci are randomly associated, then

$$D = g_{11}g_{00} - g_{10}g_{01} = (p_1p_2)(q_1q_2) - (p_1q_2)(p_2q_1) = 0$$

so it seems tempting to use D as a measure of the degree of nonrandom association. The temptation becomes all the greater when one notices that dynamical models aside, in a 2×2 table representing observed proportions of gametes in a population sample of size N gametes (Table 1), the χ^2 test for association between the loci can be written simply as:

$$\chi^2 = \frac{D^2N}{p_1p_2q_1q_2}.$$

From the beginning of work on gametic disequilibrium, however, it was noticed that the possible values of D are constrained by the marginal allelic frequencies, p_1 and p_2. By reference to the 2×2 table just given, it is easy to verify that the largest positive value D can take is either p_1q_2 or p_2q_1, whichever is smaller, while the most negative value D can take is either p_1p_2 or q_1q_2, whichever is smaller. This allele frequency limitation of D disqualifies it as a general measure of association, since one would like to compare the gametic disequilibria for the same loci in different populations, or different pairs of loci in the same population, that have different allelic frequencies. As a consequence, many measures related to D have been proposed that are supposed to normalize D for allelic frequencies. In a recent paper, HEDRICK (1987) has examined these various measures and shown them all to be severely tainted by the marginal allelic frequencies. In his examination, HEDRICK used, as a standard of comparison, a measure D' introduced by LEWONTIN (1964).

$$D' = \frac{D}{D_{\max}} \tag{1}$$

where

$$D_{\max} = \min(p_1q_2, q_1p_2) \text{ when } D > 0$$

where

$$D_{\max} = \min(p_1p_2, q_1q_2) \text{ when } D < 0.$$

TABLE 1

Two by two array relating gametic frequencies to allele frequencies at two loci

	A	a	
B	g_{11}	g_{01}	p_2
b	g_{10}	g_{00}	q_2
	p_1	q_1	

HEDRICK gives as the reason for using D' as a standard that "it is independent of allelic frequencies" (p. 333). Unfortunately, this is not true, and the situation is even messier than he suspected. So one must be even more cautious than suggested by HEDRICK's title, "Gametic disequilibrium measures: proceed with caution."

Obviously, the *range* of D' is independent of the p_i since it is precisely defined as a proportion of the maximum possible value of D. So it must vary between -1 and $+1$, independent of the p_i. It is not, however, "independent" of the p_i in any other general sense. To see the problem, we must first consider what we might *mean* by a measure of association that is independent of other properties of a distribution. Somehow, there is involved the concept of "association" as an independent pure property of a pair of variables so that the "same" association can be exhibited by variable pairs taken from all sorts of populations with different single variable distributions. But there is no standpoint from which we can define such a completely general concept, although we may do so in certain specific cases.

GENERAL RELATIONSHIPS

Consider the concept of *correlation*. If two variables x and y have a bivariate normal distribution, then distribution function $f(x, y)$ is given by the familiar

$$f(x, y)\,dxdy = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot\exp\left\{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)\right]\right\} dxdy \quad (2)$$

where μ_1, μ_2, σ_1, and σ_2 are the means and standard deviations of the two variables and ρ, the correlation coefficient, is defined by the relationship

$$\rho^2 = \frac{(\sigma_{12})^2}{\sigma_1^2\sigma_1^2}$$

and σ_{12} is the covariance of x and y.

Let us consider another population in which the values of the variates, x' and y', are related to x and y by linear relationships.

$$x' = ax + b$$
$$y' = cx + d. \quad (3)$$

Noting that the new variable x' has mean $a\mu_1 + b$ and standard deviation $a\sigma$, and similarly for y' and that $\sigma_{12}' = ac\sigma_{12}$ we can find the density function of x', y' by substituting the transformation (3) into (2). We then get

$$f(x', y')\,dx'dy' = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot\exp\left\{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x'-\mu_1'}{\sigma_1'}\right)^2 + \left(\frac{y'-\mu_2'}{\sigma_2'}\right)^2 - 2\rho\left(\frac{x'-\mu_1'}{\sigma_1'}\right)\left(\frac{y'-\mu_2'}{\sigma_2'}\right)\right]\right\} dx'dy'. \quad (4)$$

From (4) and (2) we then see that the bivariate distribution for the old and new variates have exactly the same form, with the parameters μ', σ', etc., replacing the old parameters μ, σ, etc., *but with an unchanged parameter*, ρ. The invariance of ρ follows simply from its definition

$$\rho^2 = \frac{\rho_{12}^2}{\sigma_1\sigma_2} = \frac{ac\sigma_{12}^2}{a\sigma_1 c\sigma_2} = \rho'^2$$

and its appearance in the exact same form in the transformed distribution is a consequence of the particular distribution, the bivariate normal, which transforms into itself under a linear transformation of the variables. If the underlying variables are not bivariate normally distributed, or if the transformation relating x, y to x', y' is not linear, this invariance does not apply, and ρ loses its value as an invariant measure of association. For example, if the transformation

$$x' = (x-a)^2$$
$$y' = (y-b)^2$$

is applied to normal variates, the new variates, x' and y', are not normally distributed (each has a χ^2 distribution) and $\rho(x', y') \neq \rho(x, y)$. Thus, correlation is not, in general, a measure of "pure" association. Its value in the case of normal distributions arises from its *forced* invariance under linear transformations of the variables.

In fact, several measures compared by Hedrick are closely related to a "correlation" measure that has been used in linkage disequilibrium studies:

$$r = \frac{D}{\sqrt{p_1p_2q_1q_2}} = \frac{\mathrm{cov}(A, B)}{\sqrt{\sigma_A^2\,\sigma_B^2}} \quad (5)$$

which looks like a correlation measure because r, assigning the random variables 0 and 1 to the two allelic states at locus, is the ratio of the covariance between the loci to the square root of the product of their variances. In fact, it has none of the invariant properties of ρ in the bivariate normal distribution.

TABLE 2

Contingency table relating the gamete disequilibrium parameter D, to the allelic frequencies at two loci

	1	0	
1	$p_1p_2 + D$	$p_2q_1 - D$	p_2
0	$p_1q_2 - D$	$q_1q_2 + D$	q_2
	p_1	q_1	

TABLE 3

Fitnesses for a hypothetical case of selection at two loci

	AA	*Aa*	*aa*
BB	1.00	0.90	0.80
Bb	0.95	0.80	0.75
bb	0.50	0.45	0.30

For example, it does not vary between -1 and $+1$, unless $p_1 = p_2 = 0.5$.

We now turn from correlation measures in normal populations to the simple bivariate 2 × 2 contingency table. Identifying the A locus with i and the B locus with j, then the pair i,j can take only four values, 1,1; 1,0; 0,1 and 0,0, with general probabilities given in the 2 × 2 contingency table (Table 2). In algebraic terms, the bivariate distribution of i,j is given in terms of the allele frequencies and D as:

$$P(i,j) = [ip_1 + (1 - i)q_1]\,[jp_2 + (1 - j)q_2] + (-1)^{i+j}D \tag{6}$$

which we represent as

$$P = [I]\,[J] + (-1)^{i+j}D$$

If we now change the marginals to

$$p_1' = p_1 + \epsilon \quad \text{and} \quad p_2' = p_2 + \sigma$$

and denoting the new value of D as D^*, we obtain

$$P'(i,j) = [I']\,[J'] + (-1)^{i+j}D^* \tag{7}$$

where I' and J' are the same as I and J with the new values of p_1' and p_2' substituted.

We note that D^* is not forced to be equal to D because there is one degree of freedom in filling the 2 × 2 table. Thus we are at liberty to substitute any value of D^*, subject only to the constraint that

$$P'(i, j) \geq 0 \qquad \text{for all } i,j.$$

It is this constraint that causes the *limits* of D^* to be determined by the p_i, but only the limits. But if D^* is indeterminate for a given change of p_i, then any closed function of D^* and the p_i' is also indeterminate (except for trivial functions in which D^* cancels out entirely). Thus, there is no measure that includes D, or any parametric representation of p_1, p_2 and D, that is invariant with arbitrary changes in the p_1. Therefore, if we are to seek for some measure to compare the "pure" associations in two populations, we must appeal to some *a priori* notion of association in each case.

SPECIFIC EXAMPLES

In an evolutionary context, we might demand that a measure of pure association between loci, independent of allele frequency, ought to have the property that two populations starting with the same association and subject to identical forces of evolutionary change ought to have equal association values after the change. As we now show, D' fails this test in two different cases, in two different ways.

Assume two loci each with two alleles, subject to selection according to the fitness values given in Table 3. The fitnesses have been chosen to favor *AABB* and some epistasis has been introduced. Assume the two loci have 5% recombination. We contrast the results of a single generation of selection in three different populations, each beginning at a different set of allele frequencies in the initial generation, but with $D = 0$ in all three cases. The initial compositions and result of 1 generation of selection are shown in Table 4. As the table shows, neither D nor D' are the same in the three populations after selection, although the only difference is in the starting allele frequencies, and all started from $D = 0$. In the sense of equal results of equal forces, D' is clearly not "independent" of the p_i.

As a second case, assume two populations which exchange, without selection, a proportion m of their

TABLE 4

Outcome of a single generation of selection according to the fitnesses in Table 1, for three populations of different initial composition

Population	q_{11}	q_{10}	q_{01}	q_{00}	p_1	p_2	D	D'
Pop. 1								
Gen. 0	0.2500	0.2500	0.2500	0.2500	0.5000	0.5000	0	0
Gen. 1	0.3067	0.2269	0.2731	0.1933	0.5336	0.5798	−0.00268	−0.01379
Pop. 2								
Gen. 0	0.0900	0.2100	0.2100	0.4900	0.3000	0.3000	0	0
Gen. 1	0.1316	0.2041	0.2767	0.3876	0.3357	0.4083	−0.00550	−0.04013
Pop. 3								
Gen. 0	0.0400	0.1600	0.1600	0.6400	0.2000	0.2000	0	0
Gen. 1	0.0666	0.1678	0.2426	0.5229	0.1344	0.2292	−0.00587	−0.08216

individuals and then form gamete pools. Assume the two populations have different gametic frequencies before migration designated by G_{ij} for population 1 and g_{ij} for population 2. We designate by D and d the gametic disequilibrium in populations 1 and 2 respectively, and m, the migrant proportion exchanged.

Then in population 1, the gametic frequencies after migration are:

$$G_{ij}' = mg_{ij} + (1 - m)G_{ij}$$

so that the value of D after one generation

$$D^* = m^2d + (1 - m)^2D + m(1 - m) \\ [g_{11}G_{00} + g_{00}G_{11} - g_{01}G_{10} - g_{10}G_{01}]. \quad (8)$$

But, if $d = D$, then (8) is completely symmetrical in m and $(1 - m)$, so (8) also gives the new disequilibrium in population 2. That is, if $D = d$, then $D^* = d^*$, so after one generation of reciprocal exchange, two populations that begin with equal D end up with equal D, even if their initial *allele* frequencies were different. In that case, the D' values will not be equal. Nor, even if $D' = d'$, would $D^{*\prime} = d^{*\prime}$. So, D' is not an invariant of the evolutionary process for equal reciprocal migration, although D is!

Finally, if, instead of exchanging equal proportions of migrants, the two populations receive equal proportions of migrants from a third population, then neither D nor D' remain equal even though they start as equal.

Our conclusion is that there are no generally gene frequency independent measures of association between loci, and that, indeed, the concept itself is an ill-defined one. In any particular case, we *may* be able to find a measure of association that is preserved under particular conditions, but the search for a "pure" measure of gametic disequilibrium is doomed to failure.

Research was carried out under grant GM 21179-14 from the National Institutes of Health and grant DMB 8801057 of the National Science Foundation.

LITERATURE CITED

HEDRICK, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

LEWONTIN, R. C., and K. KOJIMA, 1960 The evolutionary dynamics of complex polymorphisms. Evolution **14:** 458–472.

Communicating editor: B. S. WEIR