Heredity **60** (1988) 435–438
© The Genetical Society of Great Britain

Received 17 August 1987

Inference of recombinational hotspots using gametic disequilibrium values

Philip W. Hedrick

Division of Biological Sciences, University of Kansas,
Lawrence, Kansas 66045, U.S.A.

Gametic disequilibrium between RFLP sites have been used to infer recombinational hotspots. However, some measures of gametic disequilibrium, such as the correlation coefficient, are quite dependent upon allelic frequencies. The normalized disequilibrium measure of Lewontin, on the other hand, is independent of allelic frequencies. Data from the β-globin gene cluster are used as an illustration of such potential misinterpretations and the use of these measures.

INTRODUCTION

Population data on DNA sequences or restriction fragment length polymorphisms (RFLP) promise to give new insights into the organization of genetic variation. Evolutionary interpretation of these data will obviously be a difficult, but hopefully rewarding, task in the coming years. It is my intent here to illustrate a situation in which the interpretation of RFLP data may be dependent upon the measure of genetic variation used, specifically the inference of a recombinational hotspot from gametic disequilibrium values (*e.g.*, Chakravarti *et al.*, 1984). I should note that the criteria for choosing the measure of gametic disequilibrium may vary (see Hedrick, 1987), and that other measures may be appropriate in other evolutionary situations. In addition, there is substantial direct evidence indicating the presence of recombinational hot spots (for a review, see Steinmetz *et al.*, 1987).

MEASURING GAMETIC DISEQUILIBRIUM

Let me begin by a brief discussion of two possible scenarios by which a two-locus polymorphism may originate *via* mutation or mutation and genetic drift, *i.e.*, no recombination is involved. In fig. 1, the original gamete (or haplotype) is $--$, *i.e.*, it is missing restriction sequences at both sites (the opposite, $++$, could be the original type as well). In the first instance as given in fig. 1(a), mutation at the second site results in a population having the two gametes $--$ and $-+$. A second mutation which occurs on a $--$ gamete results in both sites being polymorphic and the population having the three gametes $--$, $-+$, and $+-$. In the second scenario (fig. 1(b)) the first step begins as in fig. 1(a). However, the second mutation occurs on gamete $-+$ leading to a $++$ gamete so that the population is polymorphic at both sites. Finally, by genetic drift, the intermediate $-+$ gamete is lost and the population has only gametes $--$ and $++$, still being polymorphic at both sites.

In these scenarios, mutation or mutation and genetic drift generated the disequilibrium. The latter case, when there are only $--$ and $++$ gametes (it could be only $-+$ and $+-$ with a different scenario), can occur only when the frequencies of the alleles at the two loci are identical. These two scenarios result in two different kinds of association (Kendall and Stuart, 1961; Clegg *et al.*, 1976). First, *absolute association* (or absolute disequilibrium in this case) as shown in fig. 1(b) requires that an allele at one locus must always be associated with a particular allele at another locus, *e.g.*, only $++$ and $--$ or only $+-$ and $-+$ gametes are present. Second, when there is *complete association* (or disequilibrium) as shown in fig. 1(a), the allelic frequencies at the two loci may differ but there is still the *maximum* association possible given differences in allelic frequencies. Important for our purposes is that both absolute and complete disequilibrium can occur when there is no recombi-

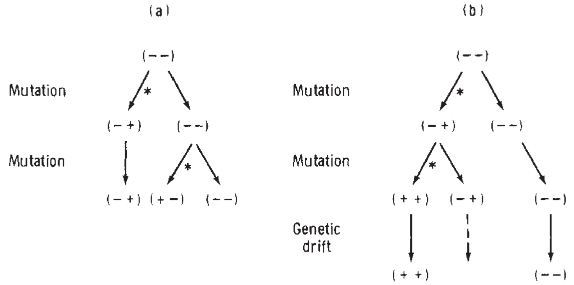

Figure 1 Diagrams of two scenarios resulting in disequilibrium between two restriction sites without recombination.

nation and that absolute disequilibrium occurs only when allelic frequencies at the two loci are equal. Therefore, it seems when trying to identify regions of low (or high) recombination from population data, we need a measure of disequilibrium that incorporates both absolute and complete disequilibrium.

In analyzing restriction site polymorphisms within the β-globin gene cluster, Chakravarti *et al.* (1984) use the "standardized" gametic (or linkage) disequilibrium measure

$$\Delta = \frac{D}{(p_1 p_2 q_1 q_2)^{1/2}} \qquad (1)$$

where $D = x_{11} - p_1 q_1$, x_{11} is frequency of gamete $++$ (presence of restriction sequences at two sites), p_1 and p_2 are the frequency of the restriction sequence ($+$) or its absence ($-$), respectively, at the first site and q_1 and q_2 are analogous frequencies at the second site (Hill and Robertson, 1968). Chakravarti *et al.* state that Δ is "independent of the true gene frequencies of the RSPs." However, Δ is not independent of the observed allelic frequencies and can have different ranges depending upon allelic frequencies (Lewontin, 1964; Hedrick, 1987). This fact suggests that any conclusions about recombinational hot spots based on Δ values (or other values derived from them) should be carefully evaluated.

The measure Δ (actually the product moment correlation) is a measure of absolute association, *i.e.*, it only assumes its extreme values when $x_{11} = x_{22} \neq 0$ and $x_{12} = x_{21} = 0$ or vice versa (Clegg *et al.*, 1976). In fact, Nei (1987) states that "this measure cannot be used for comparing the extent of nonrandom association for different loci". For our purposes, we would like a measure that assumes its extreme values when there is either absolute or complete association. A disequilibrium measure that incorporates both absolute and complete association is that suggested by Lewontin (1964) which is

$$D' = \frac{D}{D_{\max}} \qquad (2)$$

where if $D < 0$, $D_{\max}$ is the lesser of $p_1 q_1$ and $p_2 q_2$ and if $D > 0$, $D_{\max}$ is the lesser of $p_1 q_2$ and $p_2 q_1$. Unlike Δ, D' has the same range, from -1 to 1, for all combinations of allelic frequencies at the two loci. (A number of different measures (Clegg *et al.*, 1976; Donald and Bell, 1985; Litt and Jorde, 1986; Nei and Li, 1980) have ranges that are independent of allelic frequencies. In fact, Δ divided by the maximum Δ possible for the observed allelic frequencies may be a useful measure in this respect.)

To illustrate the frequency dependence of Δ, let us assume that D' has a given value and then calculate the resulting value of Δ for various allelic frequencies at the two loci. If we assume $D' > 0$ and $p_1 q_2 < p_2 q_1$, then

$$D = D' p_1 q_2$$

and

$$\Delta = \frac{D' p_1 q_2}{(p_1 p_2 q_1 q_2)^{1/2}}. \qquad (3)$$

If $D' = 1$, $p_1 = 0{\cdot}1$, $q_2 = 0{\cdot}5$, then $D = 0{\cdot}05$ ($x_{11} = 0{\cdot}1$, $x_{12} = 0{\cdot}0$, $x_{21} = 0{\cdot}4$, and $x_{22} = 0{\cdot}5$) and $\Delta = 0{\cdot}111$. In this case, there is complete association of the alleles at the two loci, the maximum possible given different allelic frequencies at the two loci, but the Δ value is only $0{\cdot}111$.

To give an overall picture of the dependence of Δ on allelic frequencies, fig. 2 plots Δ when

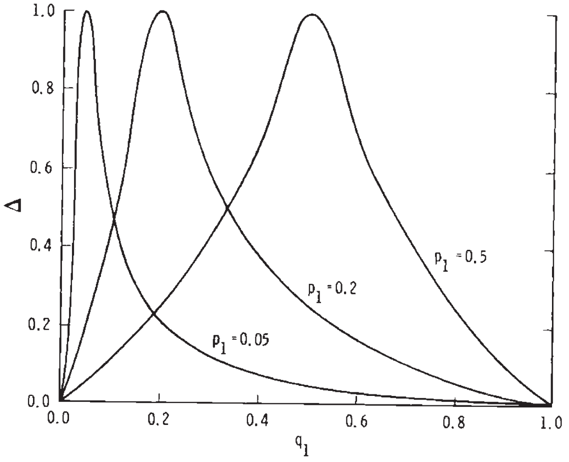

Figure 2 The magnitude of Δ given that $D' = 1$ for different values of p_1 and q_1.

$D' = 1$ for different allelic frequencies (when $D' = -1$, complementary results occur and when $|D'| < 1$, then similar but not as extreme results occur). Notice that when one locus has alleles at intermediate frequency, say $q_1 = q_2 = 0{\cdot}5$ and the other locus has one allele with a lower frequency, say $p_1 = 0{\cdot}05$ and $p_2 = 0{\cdot}95$, then the values of Δ are much lower than if the alleles at both loci were intermediate in frequency. The maximum value of Δ occurs when the allelic frequencies at the two loci are equal.

As an example, let us assume that restriction sites, *A*, *B*, *C*, and *D*, are tightly linked and that there is the maximum disequilibrium possible present for the observed allelic frequencies. Two such gametic arrays are given in table 1 with the calculated Δ and D' disequilibrium values. These arrays were chosen so that the frequencies of the alleles at loci *A* and *B* were equal and those at loci *C* and *D* were equal but different from *A* and *B*. For example, for the array given on left, the frequency of + at sites *A* and *B* is 0·2 and that of + at sites *C* and *D* is 0·4. Notice that the magnitude of Δ is smaller between sites *B* and *C* than for sites *A* and *B* or *C* and *D* while the value of D' is unity in all cases. If one did not know that Δ was dependent upon allelic frequencies, then it would appear that the disequilibrium between *B* and *C* was actually lower than between *A* and *B* or *C* and *D*.

Table 1 The value of pairwise disequilibrium for Δ and D' for the frequencies of four-locus gametes (*A B C D*) given in parentheses

	(++++, ++−−, −−++, −−−−)					
Frequency of gametic array	(0·2, 0·0, 0·2, 0·6)			(0·2, 0·0, 0·6, 0·2)		
Locus pair	*A–B*	*B–C*	*C–D*	*A–B*	*B–C*	*C–D*
Δ	1·0	0·062	1·0	1·0	0·375	1·0
D'	1·0	1·0	1·0	1·0	1·0	1·0

THE β-GLOBIN GENE CLUSTER

How could the frequency-dependence of Δ have affected the conclusions of Chakravarti *et al.* concerning the presence of a recombinational hotspot? From their analysis, they conclude that there is a 5′ cluster, 5′ to TaqI-5′δ, a 3′ cluster 3′ to HgiAI-β, and a recombinational hotspot of 9·1 kb between these sites that also includes site HinfI-5′β. In their description of these areas, they state that "(RsaI-5′δ and) TaqI-5′δ are strongly associated with all polymorphisms 5′ to their location but not with RSPs 3′ to their location ... HinfI-5′β RSP does not show significant associations with any marker 5′ to its location .... Thus, for this data set, TaqI-5′δ is, in fact 3′ terminus of the 5′ cluster. Interestingly, the HinfI-5′β RSP does not show significant associations with any RSP 3′ to its location either".

Table 2 gives the numbers of haplotypes for the two pairs of sites TaqI-5′δ, HinfI-5′β and HinfI-5′β, HgiAI-β along with the calculated Δ and D' values. First, notice that in five of the six combinations involving HinfI-5′β (first six rows of table 2), there is the maximum disequilibrium possible for the given D' values. The Δ values which Chakravarti *et al.* based their conclusions on are much lower because the frequencies of the restriction sites at the pairs of positions are quite different.

Furthermore, Chakravarti *et al.* state that a new site, TaqI-5′ε which they found in American blacks "fails to associate significantly with any other RSP". Table 2 gives the number of haplotypes for this site and HincII-5′ε, the site immediately 3′. For this pair of sites, Δ is 0·16 but D' is the maximum possible with these allelic frequencies. In addition, they state that "the HpaI-3′β site, polymorphic only in blacks, shows no significant associations with any 3′ RSP". The bottom row in table 2 is this site and the site immediately 3′, HindIII-3′β. Here again there is the maximum disequilibrium possible for these allelic frequencies.

It is important to note that there are independent pieces of evidence (Chakravarti *et al.*, 1984, 1986) that there is a region of relatively higher recombination in the β-globin gene cluster. It would be useful to analyze the complete data set of Chakravarti *et al.* using an allelic-frequency

Table 2 The number of haplotypes for five pairs of sites and the Δ and D' values calculated from them

	++	+−	−+	−−	Δ	D'
TaqI-5′β, HinfI-5′→						
Greek	10	1	7	0	−0·18	−1·0
Italian	6	1	20	0	−0·33	−1·0
Black	4	2	6	2	−0·09	−0·13
HinfI-5′β, HgiAI-β						
Greek	26	7	1	0	−0·09	−1·0
Italian	37	8	2	0	−0·10	−1·0
Black	27	7	8	0	−0·22	−1·0
TaqI-5′ε, HincII-5′ε						
Black	2	18	0	6	0·16	1·0
HpaI-3′β, HindIII-3′β						
Black	11	4	0	1	0·38	1·0

independent measure such as D' or its multilocus analogues (*e.g.*, Thomson and Baur, 1984) to determine if the results of the indirect approach are consistent with the direct experimental data. Another approach to understand these disequilibrium patterns might be to attempt a parsimonious reconstruction of events that could result in the observed sample of haplotypes (*e.g.*, Hudson and Kaplan, 1985).

DISCUSSION

Molecular data may allow the evaluation of both the importance of new phenomena and the relative effect of known factors in evolution. However, in utilizing such data we need to carefully apply population genetic measures and techniques to avoid conclusions that are partially a function of the technique or measure used. As discussed above, some measures of gametic disequilibrium are strongly influenced by alleleic frequencies, thereby making comparisons using these measures between pairs of loci with different allelic frequencies invalid. On the other hand, the disequilibrium measure of Lewontin (1964) is frequency-independent and incorporates both absolute and complete association between alleles at different loci. Obviously, Lewontin's measure I or others with similar properties, are preferable to frequency-dependent measures in many evolutionary contexts (*e.g.*, Hedrick, 1987).

Acknowledgement This work was supported by NIH grants GM 35326 and HD 12731.

REFERENCES

CHAKRAVARTI, A., BUETOW, K. H. AND ANTONARAKIS, S. E. *et al.* 1984. Nonuniform recombination within the human β-globin gene cluster. *Am. J. Hum. Genet.*, *36*, 1239-1258.

CHAKRAVARTI, A., BUETOW, K. H. AND ANTONARAKIS S. E. *et al.* 1986. Nonuniform recombination within the human β-globin gene cluster: A reply to B. S. Weir and W. G. Hill. *Am. J. Hum. Genet.*, *38*, 779-781.

CLEGG, M. T., KIDWELL J. F., KIDWELL, M. G. AND DANIEL, N. J. 1976. Dynamics of correlated genetic systems. I. Selection in the region of the Glued locus of *Dropsophila melanogaster. Genetics*, *83*, 793-810.

DONALD, J. A. AND BALL, S. P. 1984. Approximate linkage equilibrium between two polymorphic sites within the gene for human complement component 3. *Ann. Hum. Genet.*, *48*, 269-273.

HEDRICK, P. W. Gametic disequilibrium measures: Proceed with caution. *Genetics*, *117*, 331-341.

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

HUDSON, R. R. AND KAPLAN, N. L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, *222*, 147-164.

KENDALL, M. G. AND STUART, A. 1961. *The Advanced Theory of Statistics*, Vol. II. Charles Griffin and Co., London.

LEWONTIN, R. C. 1984. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*, *49*, 49-67.

LITT, M. AND JORDE, L. B. 1986. Linkage disequilibria between pairs of loci within a highly polymorphic region of chromosome 2Q. *Am. J. Hum. Genet.*, *39*, 166-178.

NEI, M. 1987. *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York.

NEI, M. AND LI, W-H. 1980. Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet. Res.*, *35*, 65-83.

STEINMETZ, M., UEMATSU, Y. AND LINDAHL, K. F. 1983. Hot spots of homologous recombination in mammalian genomes. *Trends Genet.*, *3*, 7-10.

THOMSON, G. AND BAUR, M. P. 1984. Third order linkage disequilibrium. *Tissue Antigens*, *24*, 250-255.