Ann. Hum. Genet. (2001), **65**, 395–406
Printed in Great Britain

Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci

C. ZAPATA[1], C. CAROLLO[2] and S. RODRIGUEZ[1]

[1] *Departamento de Biología Fundamental, Universidad de Santiago, Santiago de Compostela, Spain*
[2] *Departamento de Estadística e I. O., Universidad de Santiago, Santiago de Compostela, Spain*

(Received 6.11.00. Accepted 30.3.01)

SUMMARY

The development of the theory of estimation of gametic disequilibrium for multiallelic systems is particularly necessary, since a large number of the genetic markers available at present are highly polymorphic multiallelic systems. The D' coefficient is one of the most commonly used measures of the extent of overall disequilibrium between all possible pairs of alleles at two multiallelic loci. Nevertheless, the sampling properties of this measure of overall disequilibrium, are to date, unknown. In this work, we have derived explicit expressions by large-sample theory to compute the approximate sampling variance of $\hat{D}'$ between pairs of multiallelic loci, when samples of haplotypes are taken from populations. Formulae for calculating the asymptotic sampling variance were checked by Monte Carlo simulation. In addition, the magnitude of the sampling variance of $\hat{D}'$ was investigated under different scenarios of disequilibrium between multiallelic loci. Extensive simulations were also carried out for describing the sampling distribution of $\hat{D}'$, conditioned on the sample size, number of alleles and their frequencies, and disequilibrium components. It was found that the sampling distribution of $\hat{D}'$ generally approaches well the theoretical normal distribution for experimental sample sizes, particularly when loci have many alleles. Disequilibrium data between microsatellite loci of human chromosome 11p are used for illustration. These investigations increase substantially our knowledge about this widely used measure of overall disequilibrium, which is relevant to evaluate disequilibrium between multiallelic loci in populations.

INTRODUCTION

The study of non-random association of alleles at different loci, or gametic disequilibrium, is useful for revealing the location and relationships of the genes along the chromosomes, the relative influence of different evolutionary forces, and the history of populations. Searching for gametic disequilibrium between pairs of multiallelic loci often makes use of the theory of estimation for the two-allele, two-locus model. The alleles of multiallelic loci are frequently reduced to diallelic systems, by combining all rare alleles at each locus into a single class. However, this method of condensing the information generally tends to underestimate disequilibrium, especially as the number of pooled alleles increases (Weir & Cockerham, 1978; Sham & Curtis, 1995; Terwilliger, 1995). It may also obscure potentially relevant information for discriminating between the evolutionary forces generating disequilibrium in populations (Hedrick & Thomson, 1986; Hedrick, 1987; Klitz & Thomson, 1987; Thomson & Klitz, 1987; Klitz *et al.* 1992; Slatkin, 1994). In addition, pooling may decrease the ability of disequilibrium mapping to refine the location of a disease gene (Nakamura *et al.* 1987; Watkins *et al.* 1994; Chapman & Wijsman, 1998). Therefore, it is necessary to apply a proper theory of estimation of disequilibrium

Correspondence: Dr Carlos Zapata, Departamento de Biología Fundamental, Area de Genética, Facultad de Biología, Universidad de Santiago, 15782 Santiago de Compostela, Spain. Tel: (34) 981563100 Ext. 13257; Fax: (34) 981596904 (Faculty). E-mail: bfcazaba@usc.es

for multiallelic loci, as an alternative to that for diallelic loci. This need is especially acute with the advent of polymorphic markers with many alleles, such as microsatellite loci.

Recently, considerable progress has been made in the development of statistical procedures to detect significant deviations from random association for multiallelic two-locus systems (Weir & Cockerham, 1978; Excoffier & Slatkin, 1995; Long *et al.* 1995; Zaykin *et al.* 1995; Slatkin & Excoffier, 1996). There are difficulties in reliably evaluating and comparing the strength of disequilibrium across pairs of multiallelic loci. Probabilities resulting from significance tests are frequently used as a measure of the extent of disequilibrium, and for comparisons among pairs of loci and populations (Peterson *et al.* 1995; Laan & Pääbo, 1997; Huttley *et al.* 1999). Nevertheless, probabilities indicate only whether there is some real gametic disequilibrium, not whether this is weak, moderate or strong. It must also be recognized that the power of statistical tests to detect nonrandom associations depends on the sample size, the statistical tests, the number of alleles, their frequencies and whether the association is positive or negative (Brown, 1975; Weir & Cockerham, 1978; Zapata & Alvarez, 1992, 1993, 1997*a*; Slatkin, 1994; Ott & Rabinowitz, 1997; Zapata *et al.* 1997). Therefore, probabilites do not appear to be the best tools for comparing disequilibrium among pairs of loci; coefficients or indices of association should be used as well (Zapata & Alvarez, 1993; Kidd *et al.* 1998; Kruglyak, 1999; Ott, 1999).

Although different coefficients can be used for measuring the magnitude of overall disequilibrium between all possible pairs of alleles at two multiallelic loci (Karlin & Piazza, 1981; Hedrick, 1987; Klitz *et al.* 1995; Kidd *et al.* 1998; Zhao *et al.* 1999), there has been, in general, little systematic evaluation of their properties. A widely used measure of overall disequilibrium is the D' coefficient introduced by Hedrick (1987), which is a multiallelic extension of Lewontin's (1964) standardized measure of disequilibrium. The D' coefficient of overall disequilibrium has the advantageous property that its range is quite independent of the polymorphisms at the loci, thus allowing comparisons across loci or populations (Zapata, 2000). Nevertheless, the sampling properties of this measure of overall disequilibrium have not been investigated. Naturally, understanding the sampling properties of D' is useful for any descriptive or inferential analysis and can give some insight into the important factors affecting the estimation of disequilibrium between pairs of multiallelic loci, such as the number of alleles at the loci and the sample size. Formulae based on asymptotic theory for calculating the sampling variance of $\hat{D}'$ in a two-allele system have recently been obtained (Zapata *et al.* 1997), but the multiallelic case remains unresolved. It would also be desirable to know the sampling distribution of D' to assess what are the more convenient statistical tests for testing differences of the extent of overall disequilibrium over pairs of loci.

The distribution of D' for two-allele and multiallelic systems has already been characterized for samples from populations at equilibrium under neutrality (Hudson, 1985; Hedrick & Thomson, 1986; Hedrick, 1987), which has allowed comparison of the observed disequilibria to the expectations of neutrality (Hedrick & Thomson, 1986; Klitz & Thomson, 1987; Thomson & Klitz, 1987). However, the variation of disequilibrium is the sum of two sampling processes (Weir & Hill, 1980; Devlin *et al.* 1996; Weir, 1996). Firstly, the evolutionary sampling process due to the transmission of haplotypes from parent to offspring, which is a function of the effective population size each generation. Secondly, the statistical sampling process due to the sampling of a finite number of haplotypes to estimate the population disequilibrium, which is dependent on sample size. Therefore, the statistical or sampling variance is a relevant force of variation to be taken into account, along with evolutionary or stochastic variance, for interpreting disequilibrium patterns under the neutral model (Devlin *et al.* 1996; Weir, 1996). In addition, understanding the sampling variance of $\hat{D}'$ would allow us to compare disequilibrium intensities across loci without specifying any particular population model. Such comparisons may provide a valuable tool in interpreting observations in populations. Comparisons of the intensities of disequilibrium can be made among locus pairs that differ with respect to a factor that is expected to cause those disequilibria. For instance, we will be able to determine whether differences in disequilibrium intensity between pairs of loci parallel geographic variation of allele frequencies at the loci; or whether greatest disequilibria are restricted to functionally related loci. In these cases, we would have some evidence for migration and selection, respectively. On the other hand, it may also be interesting to test for differences in the extent of disequilibrium, irrespective of the evolutionary force (s) generating them. This is the case, for example, when disequilibrium is used as a tool of fine-scale disease-gene localization, because it only uses the principle that alleles at loci nearest a particular disease-influencing locus will show stronger gametic disequilibrium with the disease than alleles at distant loci (Ott, 1999).

In this paper, we develop analytical formulae, by large-sample theory, for calculation of the approximate sampling variance of $\hat{D}'$ between multiallelic loci, for a sample of haplotypes taken from a population. Monte Carlo simulations were used to check the sampling variance of $\hat{D}'$ and to investigate the sampling distribution of this disequilibrium measure.

SAMPLING VARIANCE OF $\hat{D}'$ BETWEEN MULTIALLELIC LOCI

There are several ways of describing multiple-allele gametic disequilibrium that provide very useful information depending on the particular interest of the study. The analysis of disequilibrium for each pair of alleles or haplotype shows which haplotypes are in excess and which are deficient, relative to the expectations of random association. There is a global disequilibrium analysis that condenses the information of disequilibrium between all the alleles at two loci.

Consider two polymorphic loci, A and B. Let A_i be an allele of locus A ($i = 1, \ldots, m$), B_j an allele of locus B ($j = 1, \ldots, n$), and $\hat{p}_{ij}$ the relative frequency of gamete A_iB_j in N haplotypes sampled from a population. Then $\hat{p}_{i.} = \Sigma_j \hat{p}_{ij}$ and $\hat{p}_{.j} = \Sigma_i \hat{p}_{ij}$, giving the estimated frequency of the alleles A_i and B_j, respectively. There are a total of mn coefficients of gametic disequilibrium between alleles A_i and B_j, which can be defined as

$$D_{ij} = p_{ij} - p_{i.}p_{.j} \quad \text{and estimated as} \quad \hat{D}_{ij} = \hat{p}_{ij} - \hat{p}_{i.}\hat{p}_{.j}$$

The approximate variance of $\hat{D}_{ij}$ is

$$\text{Var}\ (\hat{D}_{ij}) \approx \frac{p_{i.}(1-p_{i.})p_{.j}(1-p_{.j}) + D_{ij}(1-2p_{i.})(1-2p_{.j}) - D_{ij}^2}{N}$$

(Weir, 1979).

A more useful measure of the strength of disequilibrium for each pair of alleles is the standardized disequilibrium measure D'_{ij}, which can be defined as

$$D'_{ij} = \frac{D_{ij}}{D_{max}} \quad \text{and estimated as} \quad \hat{D}'_{ij} = \frac{\hat{D}_{ij}}{\hat{D}_{max}}$$

$$\text{where } \hat{D}_{\max} = \min\ [\hat{p}_{i.}\hat{p}_{.j},\ (1-\hat{p}_{i.})(1-\hat{p}_{.j})] \quad \text{when} \quad \hat{D}_{ij} < 0 \text{ or}$$

$$\hat{D}_{max} = \min\ [\hat{p}_{i.}(1-\hat{p}_{.j}),\ (1-\hat{p}_{i.})\hat{p}_{.j}] \quad \text{when} \quad \hat{D}_{ij} > 0.$$

(Lewontin, 1964; Hedrick, 1987). If (p_{ij} and/or $1-p_{i.}-p_{.j}+p_{ij}) = 0$ then $D'_{ij} = -1$, and if ($p_{i.}-p_{ij}$ and/or $p_{.j}-p_{ij}) = 0$ then $D'_{ij} = 1$ (Zapata, 2000). The formulae of the asymptotic sampling variance of $\hat{D}'$ for two-allele systems (Zapata *et al.* 1997) can be adapted for calculation of the variance of $\hat{D}'_{ij}$ as

$$\text{Var}(\hat{D}'_{ij}) \approx \frac{1}{ND_{\max}^2}\{(1-|D'_{ij}|)[N\text{Var}(\hat{D}_{ij}) - |D'_{ij}|D_{\max}(ap_{i.} + b(1-p_{i.}) - 2|D_{ij}|)] + |D'_{ij}|X_{ij}(1-X_{ij})\}$$

where

$$a = 1-p_{.j},\ b = p_{.j} \quad \text{for } D'_{ij} < 0 \quad \text{and} \quad a = p_{.j},\ b = 1-p_{.j} \quad \text{for } D'_{ij} > 0;\ X_{ij} \text{ is}$$

$$p_{ij},\ p_{i.}-p_{ij},\ p_{.j}-p_{ij} \quad \text{and } 1-p_{i.}-p_{.j}+p_{ij} \quad \text{when } \hat{D}_{max} \text{ is}$$

$$p_{i.}p_{.j},\ p_{i.}(1-p_{.j}),\ (1-p_{i.})\ p_{.j} \quad \text{and } (1-p_{i.})(1-p_{.j}), \text{ respectively.}$$

A global disequilibrium measure of the extent of disequilibrium between all the alleles at two loci can be defined as

$$D' = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{i.}p_{.j}|D'_{ij}| \quad \text{and estimated as} \quad \hat{D}' = \sum_{i=1}^{m}\sum_{j=1}^{n} \hat{p}_{i.}\hat{p}_{.j}|\hat{D}'_{ij}|$$

which makes use of the absolute values of $\hat{D}'_{ij}$ weighted by the frequencies of the gametes expected at gametic equilibrium (Hedrick, 1987). The D' coefficient varies from 0 to a maximum value equal or very close to 1, depending on the number of alleles and their frequencies (Zapata, 2000). This disequilibrium measure can be alternatively estimated as

$$\hat{D}' = \sum_{D^+}\hat{p}_{i.}\hat{p}_{.j}\hat{D}'_{ij} - \sum_{D^-}\hat{p}_{i.}\hat{p}_{.j}\hat{D}'_{ij},$$

where

$$\hat{D}^+ = \{ij/\hat{D}'_{ij} > 0\}$$
$$\hat{D}^- = \{ij/\hat{D}'_{ij} < 0\}.$$

The variance of $\hat{D}'$ becomes

$$\mathrm{Var}(\hat{D}') = \sum_{i}\sum_{j}\mathrm{Var}(\hat{p}_{i.}\hat{p}_{.j}\hat{D}'_{ij}) + \sum_{D^+}\sum_{D^+}\mathrm{Cov}(\hat{p}_{i.}\hat{p}_{.j}\hat{D}'_{ij},\hat{p}_{k.}\hat{p}_{.l}\hat{D}'_{kl}) + \sum_{D^-}\sum_{D^-}\mathrm{Cov}(\hat{p}_{i.}\hat{p}_{.j}\hat{D}'_{ij},\hat{p}_{k.}\hat{p}_{.l}\hat{D}'_{kl})$$

$$-\sum_{D^+}\sum_{D^-}\mathrm{Cov}(\hat{p}_{i.}\hat{p}_{.j}\hat{D}'_{ij},\hat{p}_{k.}\hat{p}_{.l}\hat{D}'_{kl}) - \sum_{D^-}\sum_{D^+}\mathrm{Cov}(\hat{p}_{i.}\hat{p}_{.j}\hat{D}'_{ij},\ \hat{p}_{k.}\hat{p}_{.l}\hat{D}'_{kl}).$$

To calculate this expression, we make use of the delta-method for deriving standard errors for large-sample inferences (Kendall & Stuart, 1977; Agresti, 1990). Let Φ denote a differentiable function of $\{\mathrm{p}_{ij}\}$, and let $\hat{\Phi}$ denote the sample value of Φ for a multinomial sample, then as $\mathrm{N} \to \propto$, the distribution of $\sqrt{N}(\hat{\Phi}-\Phi)/\sigma$ converges to a standard normal,

$$\sigma^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}(\Phi^{ij})^2 - \left(\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\Phi^{ij}\right)^2, \quad \text{where} \quad \Phi^{ij} = \frac{\partial\Phi}{\partial p_{ij}}.$$

The asymptotic variance depends on the cell probabilities $\{p_{ij}\}$ and the partial derivatives of the measure with respect to $\{p_{ij}\}$. In practice, we replace $\{p_{ij}\}$ and Φ^{ij} by their sample values, yielding a ML estimate $\hat{\sigma}^2$ of σ^2.

We make use of the following notations:

(i) $$\alpha_i = 1 \quad \text{and} \quad \beta_j = 1 \quad \text{if} \quad \hat{D}_{max} = \hat{p}_{i.}\hat{p}_{.j}$$

$$\alpha_i = 1 \quad \text{and} \quad \beta_j = \frac{p_{.j}}{1-p_{.j}} \quad \text{if} \quad \hat{D}_{max} = \hat{p}_{i.}(1-\hat{p}_{.j})$$

$$\alpha_i = \frac{p_i}{1-p_{i.}} \quad \text{and} \quad \beta_j = 1 \quad \text{if} \quad \hat{D}_{max} = (1-\hat{p}_{i.})\hat{p}_{.j}$$

$$\alpha_i = \frac{p_{i.}}{1-p_{i.}} \quad \text{and} \quad \beta_j = \frac{p_{.j}}{1-p_{.j}} \quad \text{if} \quad \hat{D}_{max} = (1-\hat{p}_{i.})(1-\hat{p}_{.j})$$

where $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{D}_{ij}$ are ML estimators of α_i, β_j and D_{ij}, respectively.

(ii) $$(D_{ij})^{i'j'} = \frac{\partial D_{ij}}{\partial p_{i'j'}} \quad (\alpha_i)^{i\cdot} = \frac{\partial \alpha_i}{\partial p_{i.}} \quad (\beta_j)^{\cdot j} = \frac{\partial \beta_j}{\partial p_{.j}}$$

(iii) $$E_{ij} = \sum_{i'=1}^{m} \sum_{j'=1}^{n} p_{i'j'} (D_{ij})^{i'j'}.$$

With these notations, all the variances and covariances included in the formula of the Var $(\hat{D}')$ can be expressed as $\text{Cov}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}, \hat{\alpha}_k\hat{\beta}_l\hat{D}_{kl})$, which, after applying the delta-method result in

$$\text{NCov}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}, \hat{\alpha}_k\hat{\beta}_l\hat{D}_{kl}) \approx \sum_{i'=1}^{m} \sum_{j'=1}^{n} p_{i'j'} (\alpha_i\beta_j D_{ij})^{i'j'} (\alpha_k\beta_l D_{kl})^{i'j'} - E_{ij}E_{kl}.$$

From here, we can distinguish four possible cases:

(A) $i = k, j = l$; (B) $i = k, j \neq l$; (C) $i \neq k, j = l$; (D) $i \neq k, j \neq l$.

The covariance for each case is given by,

(*A*) $i = k, j = l$

$$\begin{gathered} \text{NCov}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij},\ \hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}) = \text{NVar}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}) \approx \\ p_{ij}\{[(\alpha_i)^{i\cdot}\beta_j + \alpha_i(\beta_j)^{\cdot j}]D_{ij} + \alpha_i\beta_j[1-(p_{i.}+p_{.j})]\}^2 + (p_{i.}-p_{ij})\{(\alpha_i)^{i\cdot}\beta_j D_{ij} - \alpha_i\beta_j p_{.j}\}^2 \\ + (p_{.j}-p_{ij})\{\alpha_i(\beta_j)^{\cdot j} D_{ij} - \alpha_i\beta_j p_{i.}\}^2 - (E_{ij})^2 \end{gathered}$$

(*B*) $i = k, j \neq l$

$$\begin{gathered} \text{NCov}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}, \hat{\alpha}_i\hat{\beta}_l\hat{D}_{il}) \approx p_{ij}\{[(\alpha_i)^{i\cdot}\beta_j + \alpha_i(\beta_j)^{\cdot j}]D_{ij} + \alpha_i\beta_j[1-(p_{i.}+p_{.j})]\} \\ \{(\alpha_i)^{i\cdot}\beta_l D_{il} - \alpha_i\beta_l p_{.l}\} + p_{il}\{(\alpha_i)^{i\cdot}\beta_j D_{ij} - \alpha_i\beta_j p_{.j}\} \\ \{[(\alpha_i)^{i\cdot}\beta_l + \alpha_i(\beta_l)^{\cdot l}]D_{il} + \alpha_i\beta_l[1-(p_{i.}+p_{.l})]\} + (p_{i.}-p_{ij}-p_{il})\ \{(\alpha_i)^{i\cdot}\beta_j D_{ij} - \alpha_i\beta_j p_{.j}\} \\ \{(\alpha_i)^{i\cdot}\beta_l D_{il} - \alpha_i\beta_l p_{.l}\} - E_{ij}E_{il} \end{gathered}$$

(*C*) $i \neq k, j = l$

$$\begin{gathered} \text{NCov}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}, \hat{\alpha}_k\hat{\beta}_j\hat{D}_{kj}) \approx p_{ij}\{[(\alpha_i)^{i\cdot}\beta_j + \alpha_i(\beta_j)^{\cdot j}]D_{ij} + \alpha_i\beta_j[1-(p_{i.}+p_{.j})]\}\ \{\alpha_k(\beta_j)^{\cdot j}D_{kj} - \alpha_k\beta_j p_{k.}\} \\ + p_{kj}\{\alpha_i(\beta_j)^{\cdot j}D_{ij} - \alpha_i\beta_j p_{i.}\}\ \{[(\alpha_k)^{k\cdot}\beta_j + \alpha_k(\beta_j)^{j}]D_{kj} + \alpha_k\beta_j[1-(p_{k.}+p_{.j})]\} \\ + (p_{.j}-p_{ij}-p_{kj})\ \{\alpha_i(\beta_j)^{\cdot j}D_{ij} - \alpha_i\beta_j p_{i.}\}\ \{\alpha_k(\beta_j)^{\cdot j}D_{kj} - \alpha_k\beta_j p_{k.}\} - E_{ij}E_{kj} \end{gathered}$$

(*D*) $i \neq k, j \neq l$

$$\begin{gathered} \text{NCov}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}, \hat{\alpha}_k\hat{\beta}_l\hat{D}_{kl}) \approx p_{kj}\{(\alpha_k)^{k\cdot}\beta_l D_{kl} - \alpha_k\beta_l p_{.l}\}\ \{\alpha_i(\beta_j)^{\cdot j}D_{ij} - \alpha_i\beta_j p_{i.}\} \\ + p_{il}\{(\alpha_i)^{i\cdot}\beta_j D_{ij} - \alpha_i\beta_j p_{.j}\}\ \{\alpha_l(\beta_l)^{\cdot l}D_{kl} - \alpha_k\beta_l p_{k.}\} - E_{ij}E_{kl}, \end{gathered}$$

where

$$E_{ij} = \{p_{i.}(\alpha_i)^{i\cdot}\beta_j + p_{.j}\alpha_i(\beta_j)^{\cdot j}\}D_{ij} + \alpha_i\beta_j(D_{ij} - p_{i.}p_{.j})$$

Estimates of the asymptotic variance for $\hat{D}'$ are obtained by replacing all frequencies used in the definitions with the corresponding observed values. When $\alpha_i = \beta_j = 1$ and $i = k, j = l$, the formula of $\text{NCov}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}, \hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij}) = \text{NVar}(\hat{\alpha}_i\hat{\beta}_j\hat{D}_{ij})$ coincides with that of $\text{NVar}(\hat{D}_{ij})$ mentioned above.

A program (2ld) to compute the sampling variance between multiallelic markers using the present approach is available from Jin Hua Zhao (http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GepiBSt/software.stm).

BEHAVIOUR OF THE SAMPLING VARIANCE AND DISTRIBUTION OF $\hat{D}'$

We conducted Monte Carlo simulations to check the performance of the asymptotic sampling variance of $\hat{D}'$, as well as to investigate its sampling distribution. In addition, we have explored in some detail the behaviour of the sampling variance and distribution of $\hat{D}'$ under different scenarios of disequilibrium between multiallelic loci.

Table 1. *Asymptotic sampling variance of $\hat{D}'$ and Monte Carlo simulations describing the distribution of $\hat{D}'$ between multiallelic loci with equifrequent alleles*

Examples of disequilibrium						Asymptotic variance	Variation coefficient	Monte Carlo simulations statistics describing the distribution of $\hat{D}'$				
$m = n$	$p_{i.} = p_{.j}$	D'^{a}_{ij}	D'^{b}_{ij}	D'	N	s^2	CV	$\bar{D}'$	s^2	g_1	g_2	D_{k-s}
4	0.250	0.200	−0.600	0.400	200	0.0017	0.10	0.410	0.0014	−0.204	−0.254	0.033*
					400	0.0009	0.08	0.405	0.0008	−0.081	−0.012	0.024^{ns}
					1000	0.0003	0.04	0.404	0.0003	−0.082	−0.037	0.019^{ns}
		0.147	−0.440	0.293	200	0.0021	0.16	0.301	0.0019	0.074	−0.060	0.019^{ns}
					400	0.0011	0.11	0.300	0.0009	0.000	−0.305	0.027^{ns}
					1000	0.0004	0.07	0.297	0.0004	0.044	0.013	0.015^{ns}
6	0.167	0.127	−0.636	0.382	198	0.0016	0.10	0.394	0.0011	−0.112	0.274	0.025^{ns}
					394	0.0008	0.07	0.386	0.0006	−0.005	0.062	0.020^{ns}
					990	0.0003	0.05	0.384	0.0002	−0.024	−0.159	0.021^{ns}
8	0.125	0.095	−0.667	0.381	192	0.0016	0.10	0.412	0.0008	−0.042	0.100	0.014^{ns}
					384	0.0008	0.07	0.390	0.0005	−0.139	0.191	0.021^{ns}
					960	0.0003	0.05	0.384	0.0002	−0.042	0.028	0.020^{ns}
10	0.100	0.056	−0.500	0.278	200	0.0020	0.16	0.365	0.0006	−0.022	−0.065	0.016^{ns}
					400	0.0010	0.11	0.313	0.0004	−0.026	0.116	0.013^{ns}
					1000	0.0004	0.07	0.284	0.0002	−0.058	0.023	0.015^{ns}

[a] $i+j = 2k$.
[b] $i+j = 2k+1$; $k \in IN$.
*$p < 0.05$; ns, non significant.

Undoubtedly, the sampling variance and distribution of $\hat{D}'$ are potentially affected by a large number of factors. These include sample size, number and frequencies of alleles at the loci, and components of disequilibrium (overall and inter-allelic disequilibria). Consequently, to explore all possible variations of multiallelic systems becomes prohibitively large, and to ascertain what are the effects attributable to each factor is not straightforward. Nevertheless, an exhaustive analysis of all factors and combination of involved factors was performed, although it would be too tedious to be presented here. For the sake of brevity, some representative examples illustrating the most important conclusions from the present analysis are shown.

Let us first consider what happens when there are the same number of equifrequent alleles at both loci ($m = n$ and $p_{i.} = q_{.j} = 1/m$) and arrays of haplotype frequencies are constructed to give only two different D'_{ij} values. These conditions can be obtained if $m = n = 2k$ ($k \in IN$) and there are only two different haplotype frequencies (X, X'), verifying that if $i+j = 2k$ and $i+j = 2k+1$, then the relative haplotype frequencies, p_{ij}, are X and X', respectively. It is clear that these assumptions can be easily violated in real studies. However, we have begun by using this simplified scenario because it will facilitate investigation of the impact that different factors have on the sampling variance and distribution of $\hat{D}'$. It should also be noted that, for the same number of equifrequent alleles, the D' coefficient always ranges from 0 to 1 (Zapata, 2000). Table 1 shows the D' values along with their asymptotic sampling variances and the corresponding coefficients of variation (CV), for numerical cases involving different numbers of alleles (m = 4, 6, 8 and 10), several combinations of D'_{ij}, $p_{i.}$ and $p_{.j}$ values, and sample sizes (N ≈ 200, 400 and 1000). We have examined a minimum sample size of around 200, which seemed large enough to avoid the absence of haplotypes with low expectations, and yet small enough to be realistic. The problem of smaller sample sizes and haplotype classes with too low expectations will be examined below. Monte Carlo simulations were carried out by taking 1000 randomly drawn haplotype samples of size N from populations with a given set of haplotype frequencies and disequilibrium values, and obtaining the distribution of $\hat{D}'$ that results. Table 1 also gives the statistics used to describe the sampling distribution of $\hat{D}'$. These statistics were the mean

($\bar{D}'$), variance (s^2), skewness and kurtosis (g_1 and g_2, respectively) and the Kolmogorov–Smirnov test (D_{k-s}) for goodness of fit to a normal distribution (Sokal & Rohlf, 1995).

Overall, the results are very satisfactory, and several points can be made. First, as shown in Table 1 the asymptotic sampling variances of $\hat{D}'$ are generally in good agreement with the empirical variances in the computer simulation, which demonstrates that formulae for the asymptotic sampling variance of $\hat{D}'$ were indeed well derived. Second, our observations show that the sampling variance decreases as the sample size increases (under otherwise equivalent conditions), but sampling variances associated with the estimates of D' are not too large. Coefficients of variation ranged from 0.04 to 0.16 and averaged 0.09 ± 0.01. In addition, it appears that the variance does not necessarily decline as the number of alleles increases, judged by the values of the coefficients of variation given in Table 1. By way of illustration, note that the coefficients of variation are the same when the number of alleles increases from six to eight for similar D' values and sample sizes. Third, Monte Carlo simulations show that the empirical distribution of $\hat{D}'$ approaches the normal distribution with the exception of the four-allele case. Thus, empirical distributions of $\hat{D}'$ for the four-allele case do not always fit a normal distribution when the sample size is equal to 200 haplotypes, although they are normally distributed for larger sample sizes.

The above conclusions refer only to equifrequent alleles and low diversity of interallelic disequilibria. As this is not very realistic, we now consider the more general situation of nonequifrequent alleles at the loci, and greater heterogeneity of interallelic disequilibria. We have also examined a range of sample sizes that includes smaller values as well as a higher number of alleles. Interestingly, Table 2 shows that the aforementioned conclusions also apply to this more general situation. First, the sampling variances are, in general, quite close to the empirical variances except when the number of alleles is high in comparison with the sample size. In addition, the asymptotic sampling variance is always higher than the empirical variance. Assuming Monte Carlo results represent a better estimate of the uncertainty associated with the estimated parameter, tests of disequilibrium hypotheses based on the asymptotic variance will be conservative. However, it may be that the Monte Carlo approach does not provide more exact estimates of the true variance than large sample theory. Thus, the magnitude of the sampling variance can be underestimated by Monte Carlo if certain haplotypes have low expectations such that the probability of obtaining samples showing reduced variability is high. In fact, for typical microsatellite datasets, many haplotypes carrying alleles at very low frequencies are unlikely detected due to their low expectations (see Peterson *et al.* 1995). On the other hand, ML estimators are asymptotically efficient and unbiased (Elandt-Johnson, 1971). It should also be noted that the mean value of $\hat{D}'$ over the Monte Carlo replicates differ substantially from the real value of D', for smaller sample sizes. Second, the sampling variance undergoes conspicuous oscillations across disequilibrium examples, although its magnitude is not, in general, too large. Coefficients of variation ranged from 0.08 to 0.97 and averaged 0.30 ± 0.03. There is no apparent trend for sampling variance decreasing with an increasing number of alleles. Third, distributions of $\hat{D}'$ generally fit to the normal curve especially when one increases the number of alleles at the loci. We found no evidence for deviations from normality when $m \times n > 18$.

We have not yet considered in our analyses that distributions of $\hat{D}'_{ij}$, obtained either from experimental disequilibrium studies or under neutrality equilibrium models, typically exhibit large tails of $\hat{D}'_{ij} = \pm 1$ values (notably $\hat{D}'_{ij} = -1$), most probably due to the absence of haplotypes with low expectations (Hedrick & Thomson, 1986). Consequently, it was found that those distributions of $\hat{D}'_{ij}$ deviate greatly from normal distribution unless $\hat{D}'_{ij} = \pm 1$ values are excluded. Therefore, we decided to explore the performance of the sampling distribution of $\hat{D}'$, when a large number of $\hat{D}'_{ij} = \pm 1$ values is found, on the basis of a real data set.

Table 2. *Asymptotic sampling variance of $\hat{D}'$ and Monte Carlo simulations describing the distribution of $\hat{D}'$ between multiallelic loci with nonequifrequent alleles*

Examples of disequilibrium							Asymptotic variance	Variation coefficient	Monte Carlo simulations statistics describing the distribution of $\hat{D}'$				
		Range of											
m	n	$p_{i.}$	$p_{.j}$	$\lvert D'_{ij}\rvert$	D'	N	s^2	CV	$\bar{D}'$	s^2	g$_1$	g$_2$	D_{k-s}
3	3	0.30–0.40	0.32–0.36	0.220–0.580	0.436	50	0.0123	0.25	0.459	0.0099	−0.043	−0.022	0.016^{ns}
						100	0.0061	0.18	0.444	0.0053	0.023	−0.169	0.013^{ns}
						200	0.0031	0.13	0.438	0.0027	−0.073	−0.246	0.025^{ns}
						400	0.0015	0.08	0.436	0.0013	−0.055	0.256	0.024^{ns}
3	6	0.28–0.30	0.12–0.24	0.007–0.260	0.108	50	0.0109	0.97	0.290	0.0044	0.197	−0.088	0.022^{ns}
						100	0.0054	0.68	0.218	0.0028	0.363	0.115	0.037*
						200	0.0027	0.48	0.173	0.0017	0.242	0.017	0.028^{ns}
						400	0.0014	0.35	0.145	0.0009	0.253	−0.036	0.026^{ns}
3	10	0.23–0.44	0.04–0.24	0.005–0.640	0.224	100	0.0054	0.32	0.332	0.0030	0.050	0.165	0.018^{ns}
						200	0.0027	0.23	0.280	0.0016	0.099	−0.031	0.018^{ns}
						400	0.0014	0.17	0.252	0.0009	0.104	0.132	0.018^{ns}
4	4	0.24–0.26	0.12–0.38	0.050–0.700	0.228	50	0.0059	0.34	0.327	0.0048	0.169	−0.270	0.024^{ns}
						100	0.0029	0.24	0.274	0.0023	−0.016	0.004	0.028^{ns}
						200	0.0015	0.17	0.248	0.0011	0.042	−0.002	0.017^{ns}
						400	0.0007	0.12	0.235	0.0006	−0.120	−0.207	0.026^{ns}
4	6	0.22–0.28	0.12–0.26	0.007–0.490	0.174	50	0.0091	0.55	0.328	0.0040	0.126	−0.080	0.022^{ns}
						100	0.0032	0.33	0.260	0.0024	0.083	−0.277	0.024^{ns}
						200	0.0023	0.28	0.219	0.0012	0.125	−0.170	0.022^{ns}
						400	0.0011	0.19	0.194	0.0007	−0.048	−0.103	0.020^{ns}
4	10	0.19–0.35	0.08–0.15	0.010–0.600	0.144	100	0.0046	0.47	0.309	0.0018	0.123	−0.137	0.021^{ns}
						200	0.0023	0.33	0.242	0.0010	0.155	−0.034	0.023^{ns}
						400	0.0012	0.24	0.202	0.0006	0.142	0.223	0.019^{ns}
						800	0.0006	0.17	0.174	0.0003	0.015	−0.099	0.026^{ns}
6	6	0.08–0.24	0.09–0.24	0.010–0.630	0.100	100	0.0047	0.69	0.264	0.0016	0.171	0.223	0.023^{ns}
						200	0.0024	0.49	0.197	0.0010	0.145	0.010	0.023^{ns}
						400	0.0012	0.35	0.156	0.0005	0.044	−0.236	0.023^{ns}
						800	0.0006	0.24	0.130	0.0003	0.174	−0.256	0.025^{ns}
6	10	0.13–0.23	0.07–0.15	0.007–0.530	0.178	200	0.0023	0.27	0.291	0.0009	0.068	−0.059	0.023^{ns}
						400	0.0012	0.19	0.242	0.0006	0.144	0.104	0.023^{ns}
						800	0.0006	0.14	0.213	0.0003	0.092	0.042	0.015^{ns}
10	10	0.07–0.13	0.07–0.17	0.000–0.620	0.160	200	0.0021	0.29	0.323	0.0006	0.062	0.014	0.018^{ns}
						400	0.0010	0.20	0.254	0.0004	0.174	0.017	0.024^{ns}
						800	0.0005	0.14	0.212	0.0002	0.082	−0.071	0.027^{ns}
10	14	0.05–0.15	0.05–0.13	0.001–0.430	0.098	400	0.0011	0.34	0.257	0.0003	0.159	0.153	0.016^{ns}
						800	0.0005	0.22	0.195	0.0002	0.138	0.014	0.008^{ns}
14	14	0.05–0.15	0.05–0.12	0.000–0.565	0.133	400	0.0010	0.23	0.308	0.0003	0.071	0.011	0.018^{ns}
						800	0.0005	0.17	0.238	0.0002	0.032	−0.076	0.009^{ns}

*$p < 0.05$; ns, non significant.

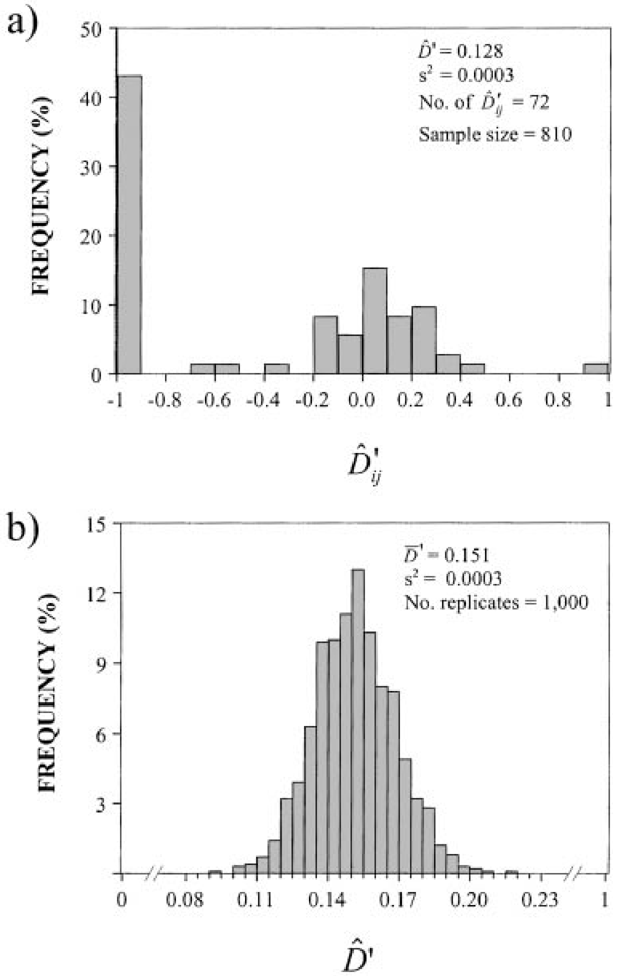

Fig. 1. (*a*) Frequency distribution of $\hat{D}'_{ij}$ values for D11S926 and D11S4124 microsatellite loci in a sample of 810 haplotypes from the Galician population (Spain); s^2 is the asymptotic variance of $\hat{D}'$. (*b*) Frequency distribution of $\hat{D}'$ for D11S926 and D11S4124 was obtained by bootstrap simulation from 1000 replicate random haplotype samples of size 810; s^2 is the empirical variance of $\hat{D}'$.

Let us consider disequilibrium data between D11S926 and D11S4124 microsatellite loci located on the 11p human chromosome for a sample of 810 haplotypes taken from the Spanish population (Zapata & Rodríguez, unpublished results). The number of alleles detected in that sample, for D11S926 and D11S4124, was eight and nine, respectively. Allele frequencies ranged from 0.005 to 0.457 for D11S926 and from 0.003 to 0.367 for D11S4124 (averages 0.125 ± 0.054 and 0.111 ± 0.050, respectively). Figure 1*a* shows the observed distribution of D'_{ij} values. As expected, it can be seen that the distribution of $\hat{D}'_{ij}$ values between all alleles at the two microsatellite loci contains a substantial proportion of $\hat{D}'_{ij} = \pm 1$ (32/72), and therefore, it clearly deviates from the theoretical normal distribution ($D_{k-s} = 0.228$; $p < 0.01$). A closer inspection of the data shows that those $\hat{D}'_{ij} = \pm 1$ are exclusively explained by haplotypes, absent from the sample, bearing alleles at very low frequency at the two loci (data not shown). The behaviour of the resulting distribution of $\hat{D}'$ under these extreme circumstances was also investigated by bootstrap simulation. We constructed a population at the observed haplotype frequencies and 1000 replicate random haplotype samples of size 810 were drawn, with replacement, from the population. Finally, the $\hat{D}'$ value was obtained for each of the 1000 random samples. As shown in Figure 1*b*, the resulting empirical distribution of $\hat{D}'$ values approached a normal distribution very closely ($D_{k-s} = 0.024$, non significant). This result can be easily understood, taking into account that $\hat{D}'$ is a weighted mean of the $\hat{D}'_{ij}$ values by their corresponding expected haplotype frequencies at gametic equilibrium. Then, maximum values of $\hat{D}'_{ij} = \pm 1$ are outweighed greatly by their low expectations and they have, therefore, a small repercussive effect on the D' measure of overall disequilibrium.

DISCUSSION

We have obtained the expressions for estimating the approximate sampling variance of the D' measure of overall disequilibrium between pairs of multiallelic loci. This allows us to define the degree of accuracy of D' estimates by means of their corresponding asymptotic standard errors, as well as to investigate the relative influence of the different factors associated with its estimation.

It is often assumed that the variances of disequilibrium estimates tend to be too large, but disequilibrium can be more easily detected to the extent to which the number of alleles at loci increases. In fact, it seems that the expected variance of disequilibrium under neutrality tends to be quite large and decreases as the number of alleles increases. In addition, the statistical power to detect disequilibrium increases when there are more alleles (Hedrick & Thomson, 1986; Hedrick, 1987; Slatkin, 1994). Nevertheless, those conclusions concerning the evolutionary variance cannot be transferred automatically to the sampling variance, because they are obtained under very different scenarios (see Introduction). Our observations suggest that the sampling variances of D' estimates are not too large for realistic sample sizes. This provides good opportunities for testing hypotheses concerning differences in disequilibrium intensity. Furthermore, it appears that the sampling variance of $\hat{D}'$ does not necessarily decline with an increasing number of alleles. Fluctuations in the magnitude of the sampling variance depend not only on the number of alleles at the two loci, but also on other factors and combinations of factors, such as the allelic frequencies and the intensity of the interallelic disequilibria, for given D' values and sample sizes.

Monte Carlo simulations show that generally $\hat{D}'$ is normally distributed, especially when the number of alleles at the loci increases. This result is not surprising since $\hat{D}'$ is defined as a weighted mean of $\hat{D}'_{ij}$ values (in absolute value) and, according with the central limit theorem, the sampling distribution of the means of random samples of any distribution, will approach the normal distribution if the sample size is sufficiently large (Sokal & Rohlf, 1995). In addition, ML estimators are asymptotically normally distributed (Elandt-Johnson, 1971). Interestingly, the assumption of normality of $\hat{D}'$ for a high number of alleles is also demonstrated to be appropriate, even when there are a high proportion of $\hat{D}'_{ij} = \pm 1$. This is illustrated by disequilibrium data between microsatellites of the human chromosome 11p. An assumption of normality allows us to apply parametric standard statistical procedures for testing differences in the intensity of disequilibrium across loci. Using the sampling variance of $\hat{D}'$, confidence intervals can be rapidly constructed, without the need to use more-time consuming statistical methods such as resampling (Efron & Tibshirani, 1993; Good, 1994; Weir, 1996). In addition, parametric statistical procedures are preferred in comparison to resampling methods because they maximize the statistical power (Crowley, 1992; Good, 1994). It must be noted that a lack of statistical power of tests used for detecting disequilibrium has traditionally been one of the most important factors causing underestimation of the importance of disequilibrium in populations (Zapata & Alvarez, 1992, 1993, 1997a).

On the other hand, our observations show that $\hat{D}'$ does not always follow a normal distribution when loci have a more reduced number of alleles ($m \times n < 20$). However, tests of goodness of fit to a normal distribution can be performed for distributions of $\hat{D}'$ generated by Monte Carlo simulation, under any particular experimental conditions, as shown in the present paper. When the distribution of $\hat{D}'$ is inadequate, resampling statistical techniques can be carried out for testing disequilibrium hypotheses, as was suggested for the two-locus, two allele case (Zapata & Alvarez, 1992, 1993, 1997*b*). Further research will be necessary to determine the relative merits of the different resampling methods (see Crowley, 1992; Good, 1994) for testing differences in the intensity of disequilibrium between pairs of multiallelic loci.

We thank two anonymous reviewers for useful comments. This work was supported by grants from the Xunta de Galicia (XUGA-20001B97) and the Ministerio de Educación y Cultura (PB96-0948), Spain, to C.Z.

REFERENCES

Agresti, A. (1990). *Categorical data analysis.* New York: John Wiley and Sons.

Brown, A. H. D. (1975). Sample sizes required to detect linkage disequilibrium between two or three loci. *Theor. Pop. Biol.* **8**, 184–201.

Chapman, E. W. & Wijsman, E. M. (1998). Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. *Am. J. Hum. Genet.* **63**, 1872–1885.

Crowley, P. H. (1992). Resampling methods for computation-intensive data analysis in ecology and evolution. *Annu. Rev. Ecol. Syst.* **23**, 405–447.

Devlin, B., Risch, N. & Roeder, K. (1996). Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**, 1–16.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall.

Elandt-Johnson, R. C. (1971). *Probability models and statistical methods in genetics.* New York: Wiley.

Excoffier, L. & Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Evol. Biol.* **12**, 921–927.

Good, P. (1994). *Permutation tests.* New York: Springer-Verlag.

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

Hedrick, P. W. & Thomson, G. (1986). A two-locus neutrality test: applications to humans, *E. coli* and lodgepole pine. *Genetics* **112**, 135–156.

Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631.

Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. (1999). A scan for linkage disequilibrium across the human genome. *Genetics* **152**, 1711–1722.

Karlin, S. & Piazza, A. (1981). Statistical methods for assessing linkage disequilibrium at the HLA-A, B, C loci. *Ann. Hum. Genet.* **45**, 70–94.

Kendall, M. & Stuart, A. (1977). *The advanced theory of statistics.* Vol. 1. London: Charless Griffin.

Kidd, K. K., Morar, B., Castiglioni, C. M., Zhao, H., Pakstis, A. J., *et al.* (1998). A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum. Genet.* **103**, 211–227.

Klitz, W. & Thomson, G. (1987). Disequilibrium pattern analysis. II. Applications to Danish HLA-A and B locus data. *Genetics* **116**, 633–643.

Klitz, W., Thomson, G., Borot, N. & Cambon-Thomsen, A. (1992). Evolutionary and population perspectives of the human HLA complex. *Evol. Biol.* **26**, 35–72.

Klitz, W., Stephens, J. C., Grote, M. & Carrington, M. (1995). Discordant patterns of linkage disequilibrium of the peptide-transporter loci within the HLA class II region. *Am. J. Hum. Genet.* **57**, 1436–1444.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144.

Laan, M. & Pääbo, S. (1997). Demographic history and linkage disequilibrium in human populations. *Nature Genet.* **17**, 435–438.

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

Long, J. C., Williams, R. C. & Urbanek, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**, 799–810.

Nakamura, Y., Leppert, M., O'Conell, P., Wolff, R., Holm, T., *et al.* (1987). Variable number of tandem repeat VNTR markers for human gene mapping. *Science* **235**, 1616–1622.

Ott, J. (1999). *Analysis of human genetic linkage.* Baltimore and London: Johns Hopkins University Press.

Ott, J. & Rabinowitz, D. (1997). The effect of marker heterozygosity on the power to detect linkage disequilibrium. *Genetics* **147**, 927–930.

Peterson, A. C., Di Rienzo, A., Lehesjoki, A. E., de la Chapelle, A., Slatkin, M., *et al.* (1995). The distribution of linkage disequilibrium over anonymous genome regions. *Hum. Mol. Genet.* **4**, 887–894.

Sham, P. C. & Curtis, D. (1995). Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann. Hum. Genet.* **59**, 97–105.

Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.

Slatkin, M. & Excoffier, L. (1996). Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. *Heredity* **76**, 377–383.

Sokal, R. R. & Rohlf, F. J. (1995). *Biometry.* New York: W. H. Freeman.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* **56**, 777–787.

Thomson, G. & Klitz, W. (1987). Disequilibrium pattern analysis. I. Theory. *Genetics* **116**, 623–632.

Watkins, W. S., Zenger, R., O'Brien, E., Nyman, D., Eriksson, A. W., *et al.* (1994). Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. *Am. J. Hum. Genet.* **55**, 348–355.

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

Weir, B. S. (1996). *Genetic data analysis II*. Massachusetts: Sinauer Associates, Inc. Publishers.

Weir, B. S. & Cockerham, C. C. (1978). Testing hypothesis about linkage disequilibrium with multiple alleles. *Genetics* **88**, 633–642.

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

Zapata, C. (2000). The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. *Evolution* **54**, 1809–1812.

Zapata, C. & Alvarez, G. (1992). The detection of gametic disequilibrium between allozyme loci in natural populations of *Drosophila*. *Evolution* **46**, 1900–1917.

Zapata, C. & Alvarez, G. (1993). On the detection of nonrandom associations between DNA polymorphisms in natural populations of *Drosophila*. *Mol. Biol. Evol.* **10**, 823–841.

Zapata, C. & Alvarez, G. (1997*a*). On Fisher's exact test for detecting gametic disequilibrium between DNA polymorphisms. *Ann. Hum. Genet.* **61**, 71–77.

Zapata, C. & Alvarez, G. (1997*b*). Testing for homogeneity of gametic disequilibrium among populations. *Evolution* **51**, 606–607.

Zapata, C., Alvarez, G. & Carollo, C. (1997). Approximate variance of the standardized measure of gametic disequilibrium D'. *Am. J. Hum. Genet.* **61**, 771–774.

Zaykin, D., Zhivotovsky, L. & Weir, B. S. (1995). Exact tests for association between alleles at arbitrary numbers of loci. *Genetica* **96**, 169–178.

Zhao, H., Pakstis, A. J., Kidd, J. R. & Kidd, K. K. (1999). Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. *Ann. Hum. Genet.* **63**, 167–179.