Genet. Res., Camb. (2005), **86**, *pp.* 77–87. *With 2 figures.* © 2005 Cambridge University Press
doi:10.1017/S001667230500769X Printed in the United Kingdom

Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL

H. ZHAO[1], D. NETTLETON[2], M. SOLLER[3] AND J. C. M. DEKKERS[1]*
[1] *Department of Animal Science, 239 Kildee Hall, Iowa State University, Ames, IA 50011, USA*
[2] *Department of Statistics, 124 Snedecor Hall, Iowa State University, Ames, IA 50011, USA*
[3] *Hebrew University of Jerusalem, Jerusalem, Israel*

(*Received 24 January 2005 and in revised form 5 July 2005*)

Summary

Effectiveness of marker-assisted selection (MAS) and quantitative trait loci (QTL) mapping using population-wide linkage disequilibrium (LD) between markers and QTL depends on the extent of LD and how it declines with distance in a population. Because marker–QTL LD cannot be observed directly, the objective of this study was to evaluate alternative measures of observable LD between multi-allelic markers as predictors of usable LD of multi-allelic markers with presumed biallelic QTL. Observable LD between marker pairs was evaluated using eight existing measures and one new measure. These consisted of two pooled and standardized measures of LD between pairs of alleles at two markers based on Lewontin's LD measure, two pooled measures of squared correlations between alleles, one standardized measure using Hardy–Weinberg heterozygosities, and four measures based on the chi-square statistic for testing for association between alleles at two loci. In simulated populations with a range of LD generated by drift and a range of marker polymorphism, marker–marker LD measured by a standardized chi-square statistic (denoted $\chi^{2\prime}$) was found to be the best predictor of useable marker–QTL LD for a group of multi-allelic markers. Estimates of the level and decline of marker–marker LD with distance obtained from $\chi^{2\prime}$ were linearly and highly correlated with usable LD of those markers with QTL across population structures and marker polymorphism. Corresponding relationships were poorer for the other marker–marker LD measures. Therefore, when LD is generated by drift, $\chi^{2\prime}$ is recommended to quantify the amount and extent of usable LD in a population for QTL mapping and MAS based on multi-allelic markers.

1. Introduction

Linkage disequilibrium (LD) is the condition in which alleles at two loci are not independent. The extent of LD is a topic of great interest in both humans and livestock. Effectiveness of marker-assisted selection (MAS) and fine mapping of quantitative trait loci (QTL) using population-wide LD between markers and QTL depends on the extent of LD and how it declines with distance (Lande & Thompson, 1990; Terwilliger & Weiss, 1998; Dekkers & Hospital, 2002). Although population-wide LD can be created by crossing lines or breeds, here we focus on LD within outbreeding populations. Because QTL cannot be observed directly, LD between markers can be used to predict marker–QTL LD, in order to evaluate the extent of useful LD in a population (e.g. Farnir *et al.*, 2000; Pritchard & Przeworski, 2001).

The two most common LD measures for biallelic markers are D' and r^2 (Lewontin, 1964; Hill & Robertson, 1968; Ardlie *et al.*, 2002), although other measures have been used (Devlin & Risch, 1995; Morton *et al.*, 2001). Based on current research, the square of the correlation coefficient between markers, r^2, is preferred to detect markers that might correlate with the QTL of interest, because r^2 quantifies the amount of information about one locus provided by the other (Ardlie *et al.*, 2002; Flint-Garcia *et al.*,

* Corresponding author. Tel: +1 (515) 2947509. Fax: +1 (515) 2949150. e-mail: jdekkers@iastate.edu

2003), although other optimal measures have been proposed (Devlin & Risch, 1995; Morton *et al.*, 2001). For biallelic markers, the absolute value of LD is the same between any pair of alleles across two loci. However, this is not true when one or both markers have more than two alleles, as is the case for the still frequently used microsatellite markers. This makes assessing the degree of LD between multi-allelic markers more complicated.

A variety of statistics have been proposed to measure LD between multi-allelic markers (Yamazaki, 1977; Hedrick & Thomson, 1986; Hedrick, 1987; Sabatti & Risch, 2002). Hedrick's (1987) multi-allelic extension of Lewontin's (1964) normalized LD measure, D', is commonly used. Using D', extensive LD over a long range was observed in dairy cattle, sheep and pigs (Farnir *et al.*, 2000; McRae *et al.*, 2002; Tenesa *et al.*, 2003; Nsengimana *et al.*, 2004). However, it is known that LD measured by D' tends to be inflated with small sample sizes and/or low allele frequencies (Ardlie *et al.*, 2002; McRae *et al.*, 2002; Flint-Garcia *et al.*, 2003). A generally satisfactory measure of LD between multi-allelic markers has not been agreed upon, nor have alternate measures of LD among multi-allelic markers been compared for their ability to predict the extent of usable LD for QTL mapping or MAS (see, however, Devlin & Risch (1995), where disease and marker loci were both assumed to have two alleles).

Random drift plays an important role in generating LD in livestock breeding populations, which are typically of limited size (Flint-Garcia *et al.*, 2003). The objective of this study was, therefore, to evaluate, by simulation, alternative measures of LD between multi-allelic markers as predictors of usable LD of multi-allelic markers with QTL and, more generally, as predictors of LD of multi-allelic markers with biallelic single nucleotide polymorphisms (SNPs), when LD is generated by drift. The ability to use LD between multi-allelic markers to predict LD among SNPs or usable LD of SNPs with QTL will be addressed in a subsequent paper.

2. Materials and methods

(i) *Measures of marker–marker LD*

The standard measure of LD between two alleles at two different loci is

$$D_{ij} = p(A_iB_j) - p(A_i)p(B_j),$$

where $p(A_i)$ is the frequency of allele A_i at locus A, $p(B_j)$ the frequency of allele B_j at locus B, and $p(A_iB_j)$ the frequency of haplotype A_iB_j. For loci with two alleles, D_{ij} completely describes LD between all pairs of alleles. Because D_{ij} depends on gene frequencies, Lewontin (1964) suggested standardizing D_{ij} by the maximum absolute value it can attain, given the allele frequencies:

$$D'_{ij} = \frac{D_{ij}}{D_{ij}^{\max}},$$

where

$$D_{ij}^{\max} = \min\left[p(A_i)\ p(B_j),\ (1-p(A_i))\ (1-p(B_j))\right] \quad \text{when } D_{ij} < 0,$$

$$D_{ij}^{\max} = \min\left[p(A_i)\ (1-p(B_j)),\ (1-p(A_i))\ p(B_j)\right] \quad \text{when } D_{ij} \geqslant 0.$$

Hill & Robertson (1968) suggested using the square of the correlation between A_i and B_j, denoted by r_{ij}^2, as a standardized measure of LD between biallelic loci. This measure can be computed from D_{ij} and allele frequencies as follows:

$$r_{ij}^2 = \frac{D_{ij}^2}{p(A_i)\ (1-p(A_i))\ p(B_j)\ (1-p(B_j))}.$$

Measures $|D'_{ij}|$ and r_{ij}^2 range from 0 to 1 but $|D'_{ij}|$ is strongly inflated if some haplotypes are not observed, which can occur for haplotypes of low-frequency alleles in small samples (Flint-Garcia *et al.*, 2003). Compared with $|D'_{ij}|$, r_{ij}^2 is less inflated in small samples (Ardlie *et al.*, 2002) and quantifies the information one locus provides about the other. Current researchers appear to prefer r_{ij}^2 for finding biallelic markers that might correlate with QTL of interest (Ardlie *et al.*, 2002; Flint-Garcia *et al.*, 2003), although there are other viewpoints (Devlin & Risch, 1995; Morton *et al.*, 2001).

As noted above, when markers have more than two alleles, LD can differ between pairs of alleles and a combined measure of LD across alleles is needed. Several such measures have been proposed (Yamazaki, 1977; Hedrick & Thomson, 1986; Hedrick, 1987; Sabatti & Risch, 2002). In this study, we compared eight existing measures and one new measure of LD between multi-allelic markers. The first two measures are based on pooling and standardizing D_{ij} across loci based on allele frequencies, following Hedrick (1987):

$$D' = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_i)\ p(B_j)\left|\frac{D_{ij}}{D_{ij}^{\max}}\right|, \tag{1}$$

or based on haplotype frequencies, following Karlin & Piazza (1981):

$$D_{hap} = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_iB_j)\left|\frac{D_{ij}}{D_{ij}^{\max}}\right|, \tag{2}$$

where k and m are the numbers of alternate alleles at locus A and B, respectively.

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

The next two measures are based on pooling r_{ij}^2 based on allele frequencies:

$$r^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} p(A_i)\ p(B_j) r_{ij}^2, \tag{3}$$

or based on haplotype frequencies:

$$r_{hap}^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} p(A_i B_j)\ r_{ij}^2. \tag{4}$$

Using Hardy–Weinberg heterozygosities at two loci, the fifth measure is

$$D^* = \frac{D^2}{H_A H_B} \tag{5}$$

(Maruyama, 1982; Hedrick & Thomson, 1986; Hedrick, 1987), where

$$D^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} D_{ij}^2, \quad H_A = 1 - \sum_{i=1}^{k} p^2(A_i)$$

and

$$H_B = 1 - \sum_{j=1}^{m} p^2(B_j).$$

The final four measures are related to the chi-square statistic to test for independence between alleles at two loci. The chi-square statistic has been discussed by Hedrick (1987) and Hill (1975) as a measure of LD and is defined as

$$\chi^2 = 2N \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{D_{ij}^2}{p(A_i)\ p(B_j)}, \tag{6}$$

where N is the sample size and $2N$ is the number of haplotypes that occurs in the sample. Two standardized measures of χ^2 have been proposed to quantify LD with values between 0 and 1:

$$\chi_{df}^2 = \frac{\chi^2}{2N(k-1)(m-1)} \tag{7}$$

(Hedrick & Thomson, 1986; Hedrick, 1987), where $(k-1)(m-1)$ is equal to the degrees of freedom of χ^2, and

$$\chi^{2\prime} = \frac{\chi^2}{2N(l-1)} \tag{8}$$

(Yamazaki, 1977), where $l = \min(k, m)$. The quantity $2N(l-1)$ gives an upper bound for the maximum of χ^2 with given marginals (i.e. given allele frequencies) in a classical χ^2 contingency table. In most cases, however, $2N(l-1)$ is much higher than the true maximum of χ^2 (Kalantari *et al.*, 1993).

To standardize χ^2 by an upper bound closer to the maximum of χ^2 than $2N(l-1)$, we developed the ninth measure by casting maximization of χ^2 conditional on marginal frequencies as a transportation problem (see Appendix 1; Winston, 1991). The optimal solution to the transportation problem provides a sharper bound, $\chi_{\max}^2$, for the maximum of χ^2 (see Appendix 1; Kalantari *et al.*, 1993) and is used to standardize χ^2:

$$\chi_{tr}^2 = \frac{\chi^2}{\chi_{\max}^2}. \tag{9}$$

Note that for biallelic markers, these nine measures reduce to four because $D' = D_{hap}$ and $r^2 = r_{hap}^2 = D^* = \chi_{df}^2 = \chi^{2\prime}$.

(ii) *Simulation*

The nine measures of marker–marker LD were evaluated for their ability to quantify LD of multi-allelic markers with biallelic QTL in simulated populations. The following criteria were used to determine the most appropriate measure of LD between markers: (1) the measure should have easy interpretation with values between 0 and 1; (2) for a given population, the measure should give a trend of marker–marker LD across distance that is similar to that of marker–QTL LD; and (3) estimates of the level and decline of LD with distance obtained from marker–marker LD should be linearly and highly correlated with the level and decline of marker–QTL LD across population structures and degrees of marker polymorphism.

To allow generation of multiple comparisons between pairs of markers and between markers and QTL at different distances, multiple markers and QTL were simulated on a 100 cM chromosome. In generation zero, markers with 2, 4, 6, 8 or 10 equi-frequent alleles were simulated at 0, 2, …, 100 cM, and QTL with two equi-frequent alleles (Q and q) were simulated at 1, 3, …, 99 cM. A total of $2N$ haplotypes were randomly sampled by independently selecting alleles at each locus. Thus, all markers and QTL were in Hardy–Weinberg and linkage equilibrium in generation 0. Subsequent generations were produced by randomly selecting and mating N parents, allowing selfing. Recombination between loci was simulated using the Haldane mapping function (Haldane, 1919).

To generate populations with varying levels of LD, data were generated for 20 combinations of population size ($N = 50$, 100, 150 or 200) and number of marker alleles (2, 4, 6, 8 or 10) in generation 0. Population size was constant across generations and data on segregating loci in generation 100 were used for analysis. Each population was replicated 100 times. Sved (1971) showed that when the number of generations is large, the expected value of LD becomes steady as a function of the product of effective population size and distance between loci. We verified

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

that LD had reached a 'steady-state' condition in generation 100 by comparing the average amount of LD for combinations that resulted in the same product of effective population size and distance between loci. These were found to be similar.

(iii) *Quantification of marker–QTL LD*

Marker–QTL LD at a given distance d in the final generation of each simulated replicate was quantified based on the ability to predict the allele at a biallelic QTL from the observed allele at a linked marker at distance d. To measure marker–QTL LD, presence or absence of allele Q in a haplotype consisting of a marker and QTL was treated as a Bernoulli random variable with probability $p(Q)$ of 'success' (i.e. presence of Q), and usable marker–QTL LD was quantified as the R^2 of the regression of Q on alleles (A_i) at a single marker. An expression for this R^2 (derived in the Appendix 2) is:

$$R^2 = \sum_{i=1}^{k} p(A_i) \frac{[p(Q|A_i) - p(Q)]^2}{p(Q)(1-p(Q))}, \quad (10)$$

where $p(Q|A_i)$ is the frequency of Q given A_i. If marker A and the QTL are in linkage equilibrium, then $p(Q|A_i) = p(Q)$ and $R^2 = 0$. If $p(Q|A_i) \neq p(Q)$, the marker allele contains information about the QTL allele and $R^2 > 0$. Measure R^2 was used as the standard to evaluate the various LD measures between markers described in Section 2.i and quantified in the simulated populations by regressing each QTL allele separately on each marker. Note that algebraically, $R^2 = \chi^{2\prime}$ for LD between a multi-allelic and a biallelic locus, and $R^2 = \chi^{2\prime} = r^2$ when both loci are biallelic.

(iv) *Comparison of LD curves predicted from marker–QTL and marker–marker LD*

To assess and compare the decline in LD with distance ($\leqslant 20$ cM) for marker–QTL LD and marker–marker LD, the function

$$LD_d = 1/(1 + 4\beta d) \quad (11)$$

(Sved, 1971; Hayes *et al.*, 2003) was fitted to the LD data that were generated for each replicate, where LD_d is LD at distance d morgans, as measured by the marker–QTL R^2 or by a marker–marker LD measure, and β is a parameter that is related to effective population size (N_e = actual population size for the idealized populations that were simulated: Falconer & Mackay, 1996). Because the variance of LD tends to decline with distance, a weighted least squares regression, which took heterogeneity of variance of LD into account, was used to estimate β for each simulated data set. The LD data for loci separated by 20 cM or less were used for this purpose. At a given distance ($\leqslant 20$ cM), the weight used was the inverse of the LD variance, which was estimated from the LD data for each replicate by using the lowess function in R software (Cleveland, 1979) to fit a smooth curve through the scatterplot of the absolute difference of the observed LD from the median LD at a given distance. The fraction of data used for smoothing at each distance point was 0·6.

Two criteria were used to compare LD curves estimated from marker–marker LD with those estimated from marker–QTL LD. The first was a measure of the correlation of estimates of β obtained from marker–QTL LD ($\hat{\beta}_{MQ}$) with those from marker–marker LD ($\hat{\beta}_{MM}$) for the various simulation conditions. To evaluate whether this relationship was consistent across population sizes and number of marker alleles, estimates $\hat{\beta}_{MQ}$ (i, j, k) obtained for population size i ($i = 50$, 100, 150 or 200), number of marker alleles j ($j = 2$, 4, 6, 8 or 10) and replicate k ($k = 1, 2, \ldots, 100$) were analysed using a model that included $\hat{\beta}_{MM}$ (i, j, k) as a covariate, population size i and number of marker alleles j as class variables, and all interactions among these three variables. The second criterion used to compare estimated LD curves was the mean of the squared difference between LD predicted based on marker–QTL LD and LD predicted using marker–marker LD over distances of 1, 2, …, 20 cM:

$$MSE = \frac{\sum_{i=1}^{20} \left(\hat{LD}_{MQ(i)} - \hat{LD}_{MM(i)} \right)^2}{20}, \quad (12)$$

where $\hat{LD}_{MQ(i)}$ and $\hat{LD}_{MM(i)}$ are LD predicted at i cM ($i = 1, 2, \ldots, 20$) using $\hat{\beta}_{MQ}$ and $\hat{\beta}_{MM}$, respectively, in equation (11).

(v) *Relationship of marker–QTL LD with local marker–marker LD*

The previous comparisons quantify the extent of LD in a population, as measured by marker–marker LD in relation to marker–QTL LD, as a function of distance. This quantifies the general magnitude and extent of LD within a population. It is, however, well known that the extent of LD within a population can differ from region to region, even if variability of LD is quantified against map distance (Heifetz *et al.*, 2005) rather than physical distance (Taillon-Miller *et al.*, 2000; Nordborg & Tavar, 2002). It is, therefore, of interest to determine whether local marker–marker LD can be used to identify genomic regions with high marker–QTL LD. To assess this, LD between two linked markers was compared with the LD of these same markers with a QTL that is bracketed by these markers. For this purpose, usable LD between a pair of markers and a bracketed QTL was quantified by

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

regressing each QTL allele on the haplotype of its two flanking markers. The R^2 of regression of Q on flanking marker haplotype A_iB_j was calculated as:

$$R^2_{hap} = \sum_{i=1}^{k}\sum_{j=1}^{m} p(A_iB_j)\frac{[p(Q|A_iB_j)-p(Q)]^2}{p(Q)(1-p(Q))}, \qquad (13)$$

where $p(Q|A_iB_j)$ is the frequency of Q given A_iB_j. The correlation of marker–marker LD measures with marker–QTL LD was used to indicate whether marker–QTL LD was greater in marker intervals that showed strong marker–marker LD. This was done for various levels of effective population size.

3. Results

(i) *Decline of LD with distance*

Fig. 1 illustrates observed relationships of several LD measures with distance for a representative replicate with a population size of 100 and 4 alleles per marker. Extensive LD between markers and QTL existed at short distances but declined rapidly with distance (Fig. 1A). Similar declines were observed when using r^2, r^2_{hap}, D^*, χ^2, χ^2_{df}, $\chi^{2\prime}$ (Fig. 1C) and χ^2_{tr}. Marker–marker LD measured by D' (Fig. 1B) and D_{hap} was strongly inflated relative to marker–QTL LD (Fig. 1A), and high values were obtained even for markers in near equilibrium.

To assess the decline of LD with distance, equation (11) was fitted to the sample data for the replicate pictured in Fig. 1. Estimates were $\hat{\beta}=53{\cdot}3$ for marker–QTL LD, and 5·4, 5·4, 92·0, 89·8, 93·5, 110·4, 42·6 and 24·1 for D', D_{hap}, r^2, r^2_{hap}, D^*, χ^2_{df}, $\chi^{2\prime}$ and χ^2_{tr}, respectively. Measure χ^2 was not used to estimate β because of its non-standardized scale. Estimate $\hat{\beta}$ obtained from $\chi^{2\prime}$ was most similar to $\hat{\beta}$ obtained from marker–QTL LD (42·6 vs 53·3) and resulted in very similar estimated LD curves (Fig. 1C). Based on mean LD at a given distance, the estimated curves appeared to provide a good fit to the data for marker–QTL LD (Fig. 1A) and for all marker–marker LD measures except for D' (Fig. 1B) and D_{hap} due to their inflated values at larger distances.

(ii) *Comparison of LD curves predicted from marker–QTL and marker–marker LD*

Results in this section are based on analysing 100 replicates for each of the 20 combinations of population size and number of marker alleles. All LD measures were evaluated except χ^2.

Table 1 shows the mean $\hat{\beta}$ across 100 replicates obtained from marker–QTL and marker–marker LD for each simulated scenario. Comparing simulations with 2 and 10 alleles per marker in generation 0, the average number of marker alleles still segregating in

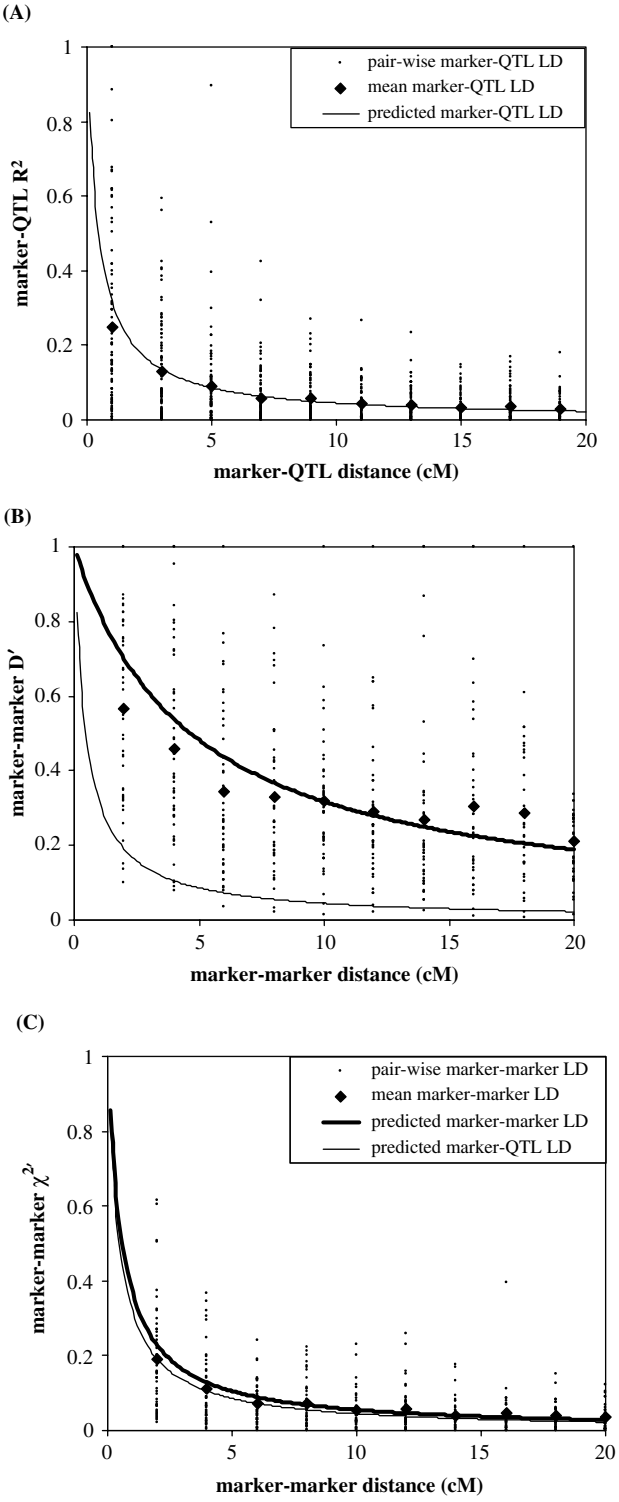

Fig. 1. Observed relationships of marker–QTL LD (*A*) and marker–marker LD measured by D' (*B*) and $\chi^{2\prime}$ (*C*) against map distance for a representative replicate with a population size of 100 and four alleles per marker. The legend for (*B*) is the same as that for (*C*). LD at distance d morgans was predicted from $\hat{LD}_d = 1/(1+4\hat{\beta}d)$, where $\hat{\beta}$ was obtained from the simulated data for each LD measure.

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

Table 1. *Mean estimates of the decline in LD with distance (β) over 100 replicates based on a measure of marker–QTL LD (R^2) and eight measures of marker–marker LD for simulated data based on different combinations of number of marker alleles in generation 0 (g_0) and population size*

No. of marker alleles (g_0)	Population size	No. of marker alleles (g_{100})	R^2	D'	D_{hap}	r^2	r^2_{hap}	D^*	χ^2_{df}	$\chi^{2\prime}$	χ^2_{tr}
2	50	2	52·2	2·5	2·5	55·7	55·7	55·7	55·7	55·7	13·1
	100	2	102·0	4·2	4·2	101·5	101·5	101·5	101·5	101·5	24·6
	150	2	152·4	6·6	6·6	148·6	148·6	148·6	148·6	148·6	40·5
	200	2	199·4	9·9	9·9	193·0	193·0	193·0	193·0	193·0	62·9
4	50	2·2	39·9	2·7	2·7	49·1	48·8	49·3	51·1	35·5	12·8
	100	2·8	55·0	5·5	5·6	92·6	89·7	93·8	104·6	46·2	25·6
	150	3·3	64·2	8·7	9·1	130·3	123·4	133·0	156·1	56·0	37·0
	200	3·6	75·6	11·3	12·2	176·4	164·6	178·9	207·0	69·0	48·5
6	50	2·3	36·8	2·8	2·8	47·8	47·2	48·0	50·5	31·6	12·8
	100	3·2	44·4	5·7	5·9	89·4	83·9	91·3	106·3	37·4	23·7
	150	4·0	48·1	8·0	8·7	130·1	115·4	133·4	161·4	42·5	31·0
	200	4·6	53·7	9·3	10·7	174·1	148·2	178·5	218·1	49·0	37·3
8	50	2·4	35·8	2·7	2·7	46·9	46·3	47·2	49·8	29·8	12·4
	100	3·5	39·5	5·7	6·0	87·3	79·6	89·6	107·5	33·1	22·5
	150	4·4	42·2	7·4	8·3	128·3	109·1	131·7	162·6	37·3	28·0
	200	5·2	45·1	8·3	10·0	170·8	137·9	175·5	219·0	41·5	32·4
10	50	2·4	34·5	2·9	2·9	47·1	46·3	47·7	51·4	29·4	13·0
	100	3·6	37·4	5·7	6·0	87·9	79·1	90·4	107·8	31·5	21·8
	150	4·7	38·4	7·0	8·1	127·1	105·0	131·0	165·5	34·3	26·2
	200	5·7	40·2	7·4	9·3	170·4	130·8	175·3	222·7	37·3	29·4

The number of marker alleles in generation 100 (g_{100}) is the average of the mean number of alleles across markers still segregating in g_{100} over 100 replicates.

generation 100 for $N=50$ were 2 and 2·4, respectively, and corresponding mean estimates of $\hat{\beta}$ for usable marker–QTL LD (R^2) decreased from 52·2 to 34·5 (Table 1). As population size increased, LD due to drift decreased. However, less drift also increased the number of alleles per marker that remained at segregating loci, e.g. to 2·4 for $N=50$ and to 5·7 for $N=200$ when starting with 10 alleles (Table 1), which increased LD by providing more information about the amount of association between alleles at different loci. The combination of these two processes resulted in a decline in mean estimates of $\hat{\beta}$ for R^2 for a given population size with an increase in the number of marker alleles that remained (Table 1). This phenomenon was more obvious for larger population sizes (Table 1). These changes were best captured by mean estimates of $\hat{\beta}$ obtained from marker–marker $\chi^{2\prime}$, which was very close to the mean $\hat{\beta}$ for R^2 (Table 1).

For biallelic markers, r^2 provided good estimates of N_e (recall that the population size is equal to N_e in our simulation) (Table 1). For multi-allelic markers, neither R^2 nor $\chi^{2\prime}$ provided good estimates of N_e (Table 1). Instead, mean $\hat{\beta}$ for χ^2_{df} was closest to the true N_e for most cases, with an upward bias of less than 12% from the true N_e (Table 1). Although slightly worse than χ^2_{df}, mean $\hat{\beta}$ for r^2 and D^* were also good estimates of N_e, but were biased downward (Table 1). Mean $\hat{\beta}$ for D' and D_{hap} were very low and did not reflect N_e (Table 1).

To obtain a better understanding of the relationship of marker–marker LD with marker–QTL LD for a given population, estimates $\hat{\beta}$ obtained from each replicate were analysed and results are shown in Fig. 2 and Table 2. Fig. 2 illustrates the relationship of $\hat{\beta}$ for marker–QTL LD ($\hat{\beta}_{MQ}$) with $\hat{\beta}$ for marker–marker LD ($\hat{\beta}_{MM}$) across the 20 simulated cases with varying population size and number of marker alleles. Results for biallelic markers were distinctly different from those for multi-allelic markers for D' (Fig. 2A). The same was true for D_{hap}, r^2, r^2_{hap}, D^*, χ^2_{df} and χ^2_{tr} (results not shown). Fig. 2B shows a good linear relationship of $\hat{\beta}_{MM}$ for $\chi^{2\prime}$ with $\hat{\beta}_{MQ}$, and the regression lines for biallelic and multi-allelic markers were almost overlapping.

Table 2 shows the correlation and slope of the regression of $\hat{\beta}_{MQ}$ on $\hat{\beta}_{MM}$ pictured in Fig. 2 for biallelic, multi-allelic and all markers. Correlations and slopes differed greatly between biallelic and multi-allelic markers for all LD measures except for $\chi^{2\prime}$ (Table 2). For $\chi^{2\prime}$, the correlation of $\hat{\beta}_{MM}$ with $\hat{\beta}_{MQ}$ was consistently high ($\geqslant 0{\cdot}95$) and the slope was close to 1, regardless of the number of marker alleles (Table 2). Using all markers, the regression line for $\chi^{2\prime}$ in Fig. 2B was $\hat{\beta}_{MQ}=6{\cdot}22+0{\cdot}98\hat{\beta}_{MM}$, with a correlation of 0·98, showing good correspondence of this

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

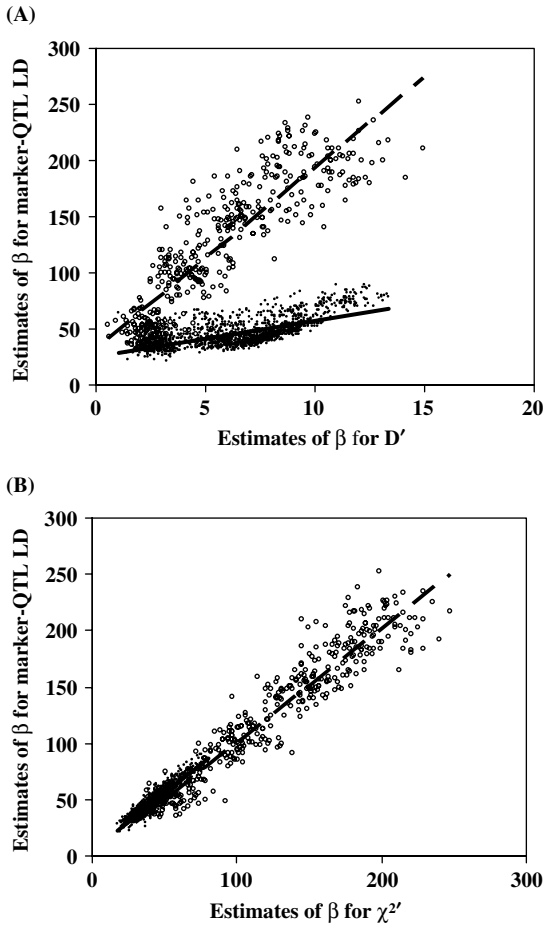

Fig. 2. Regression of estimates of the decline of LD with distance (β) obtained from each replicate for marker–QTL LD on estimates of β for marker–marker LD measured by D' (*A*) and $\chi^{2\prime}$ (*B*) for biallelic (open circles) and multi-allelic (filled circles) markers. Dashed and unbroken lines indicate the regression lines for biallelic and multi-allelic markers, respectively. Data are based on 100 replicates simulated for each of the 20 combinations of population size (50, 100, 150 or 200) and number of marker alleles (2, 4, 6, 8 or 10).

measure of marker–marker LD with marker–QTL LD.

The effects of population size and number of marker alleles on the relationship between $\hat{\beta}_{MQ}$ and $\hat{\beta}_{MM}$ pictured in Fig. 2 were tested using analysis of variance. The proportion of variance in $\hat{\beta}_{MQ}$ that was explained by simple regression on $\hat{\beta}_{MM}$ across the 20 simulated cases was 0·96 for $\chi^{2\prime}$ and ranged from 0·07 to 0·49 for the other LD measures. After including effects of population size, number of marker alleles, and all interactions among them and $\hat{\beta}_{MM}$, these proportions increased slightly for $\chi^{2\prime}$ (from 0·96 to 0·98) but greatly (from as low as 0·07 to 0·97) for other measures. Although population size and number of marker alleles explained significant ($P<0{\cdot}001$) amounts of variance in $\hat{\beta}_{MQ}$ for all LD measures (including $\chi^{2\prime}$), the relationship between $\hat{\beta}_{MQ}$ and $\hat{\beta}_{MM}$ was relatively independent of the effects of population size and number of marker alleles for $\chi^{2\prime}$.

Table 3 shows the average mean square error (MSE) (*1000) over 100 replicates for various marker–marker LD measures. The MSE was largest for D' and smallest for $\chi^{2\prime}$ for all 20 simulated cases (Table 3). This implies that, regardless of population size and number of marker alleles in the ranges we considered, LD curves predicted from $\chi^{2\prime}$ were very close to LD curves predicted from marker–QTL LD.

(iii) *Relationship of marker–QTL LD with local marker–marker LD*

Relationship of marker–QTL LD with local marker–marker LD was tested for different combinations of population size (25, 50, 75 or 100) and marker–QTL distance (0·5, 1 or 2 cM). The correlation of $\chi^{2\prime}$ between two biallelic markers with LD of these same markers with a bracketed QTL increased as population size decreased (Table 4). For a population size of 50, correlations were 0·06, 0·11 and 0·10 for marker–QTL distances of 0·5, 1 and 2 cM, respectively (Table 4). The low correlation implies that, in a population with LD generated by drift alone, LD between markers and QTL will be determined by the overall degree of LD in the population, but will not necessarily be greater in marker intervals that show strong LD between markers.

4. Discussion and conclusions

Various measures of LD between multi-allelic markers were evaluated as predictors of usable LD of multi-allelic markers with QTL for the purpose of QTL detection and MAS. The R^2 of the regression of QTL allele on alleles at a single marker was used as the standard for evaluation of the various LD measures between markers, because it quantifies the ability to predict the allele at a linked biallelic QTL based on the observed marker allele. Although biallelic QTL were simulated in this study, the results are expected to hold for multi-allelic QTL as well, because QTL alleles can always be grouped into favourable and unfavourable alleles. Although the focus was on predicting marker–QTL LD, our conclusions also apply to relating multi-allelic marker LD to LD of multi-allelic markers with SNPs. However, results do not apply to predicting LD among SNPs or between SNPs and biallelic QTL, which will be addressed in a subsequent paper.

Our study showed that $\chi^{2\prime}$ is the best measure of LD among multi-allelic markers to predict the extent of LD of those markers with QTL across population sizes and number of marker alleles. Estimates of the decline of LD with distance (β) based on $\chi^{2\prime}$ were highly and linearly related to those obtained for

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

Table 2. *Correlation and slope of the regression of the decline of LD with distance (β) estimated from marker–QTL LD on β estimated from different measures of marker–marker LD for biallelic, multi-allelic (4, 6, 8 or 10) and all markers across four population sizes (50, 100, 150 or 200), with 100 replicates for each combination of population size and number of marker alleles*

		D'	D_{hap}	r^2	r^2_{hap}	D^*	χ^2_{df}	$\chi^{2'}$	χ^2_{tr}
Biallelic markers	Correlation	0·87	0·87	0·95	0·95	0·95	0·95	0·95	0·87
	Slope	16·05	16·05	1·01	1·01	1·01	1·01	1·01	2·37
Multi-allelic markers	Correlation	0·69	0·61	0·57	0·69	0·56	0·49	0·95	0·79
	Slope	3·15	2·40	0·14	0·22	0·14	0·09	1·02	0·93
All markers	Correlation	0·36	0·26	0·50	0·65	0·47	0·29	0·98	0·70
	Slope	5·55	3·53	0·43	0·64	0·40	0·20	0·98	2·19

Table 3. *The mean of the squared difference (MSE) between LD predicted based on marker–marker and marker–QTL LD at 1, 2, …, 20 cM for simulated data generated from different combinations of population size and number of marker alleles in generation 0 (g_0)*

No. of marker alleles (g_0)	Population size	D'	r^2	$\chi^{2'}$	χ^2_{tr}
2	50	242·4	0·3	0·3	30·5
	100	168·6	0·1	0·1	15·9
	150	115·4	0·0	0·0	7·9
	200	75·5	0·0	0·0	3·6
4	50	197·7	0·5	0·3	22·0
	100	103·0	1·2	0·3	5·9
	150	59·0	1·6	0·1	2·2
	200	43·5	1·6	0·0	1·1
6	50	183·1	0·8	0·4	19·4
	100	86·2	2·5	0·3	4·5
	150	55·6	3·8	0·1	1·8
	200	47·1	4·0	0·1	1·0
8	50	186·2	0·8	0·6	20·0
	100	81·7	3·6	0·3	3·9
	150	56·6	5·2	0·2	1·7
	200	50·1	6·0	0·1	1·0
10	50	173·5	1·0	0·4	17·2
	100	78·0	4·3	0·3	3·7
	150	57·6	6·5	0·1	1·6
	200	54·6	7·6	0·1	1·0

Values are the average MSE over 100 replicates multiplied by 1000 for each combination. Results for D_{hap} (not shown) were similar to those for D', and results for r^2_{hap}, D^* and χ^2_{df} were similar to those for r^2.

Table 4. *The correlation of observable LD between two markers using $\chi^{2'}$ with LD of these same markers with a bracketed QTL*

Population size	Marker–QTL distance (cM)		
	0·5	1	2
25	0·19	0·21	0·19
50	0·06	0·11	0·10
75	0·02	0·07	0·07
100	0·02	0·06	0·06

Markers and QTL were biallelic and segregating in generation 100. Population size was 25, 50, 75 or 100. Marker–QTL distance was 0·5, 1 or 2 cM. Results are based on 10 000 replicates.

marker–QTL LD across population structures and number of marker alleles, and resulted in very similar LD curves. Corresponding relationships were poorer for the other marker–marker LD measures.

In the simulated populations, extensive marker–QTL LD existed at short distances but declined rapidly with distance. Similar declines were observed for all LD measures between markers, except for D' and D_{hap}. Due to haplotypes of low or zero frequencies in small samples, these measures gave rise to LD estimates that were strongly inflated relative to marker–QTL LD and could be high for markers that were in near equilibrium. Therefore, D' and D_{hap} are not good for high-resolution LD mapping of QTL. Measure D' was used to study the extent of LD in the Dutch black-and-white dairy cattle population by Farnir *et al.* (2000), in Coopworth and Romney sheep populations by McRae *et al.* (2002), in the UK dairy cattle population by Tenesa *et al.* (2003), and in five populations of commercial pigs by Nsengimana *et al.* (2004). Using this measure, substantial LD was observed over a long range in all four studies, but it is not clear to what extent this may be a result of the above artefact.

Although β is related to N_e, estimates of β obtained from $\chi^{2'}$ and marker–QTL LD (R^2) were not useful estimates of N_e, because they reflect not only N_e but also the number of marker alleles that remained in the generation under consideration. Sved (1971) showed that for biallelic markers, the decline in LD measured by r^2 estimates N_e, which was also observed in our study (Table 1). For multi-allelic markers, r^2, D^* and χ^2_{df} all provided good estimates of N_e.

The upper bound for the maximum of χ^2 used in our new measure χ^2_{tr} is sharper than the upper bound

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

used in $\chi^{2\prime}$. Nevertheless, χ^2_{tr} was a poorer predictor of usable marker–QTL LD than $\chi^{2\prime}$. The reason for this is that marker–QTL LD measured by R^2 attains 1·0 if and only if there is a perfect dependence of QTL alleles on marker alleles. This can only occur when each QTL allele frequency is equal to the sum of the frequencies of one or more alleles at the marker. When this condition is not satisfied, the maximum possible R^2 is less than 1·0, yet the maximum of χ^2_{tr} will be close to 1·0 because χ^2 is standardized by a relatively sharp upper bound to χ^2, conditional on marker allele frequencies. Thus, χ^2_{tr} over-standardizes χ^2 in predicting marker–QTL LD. Nevertheless, χ^2_{tr} might be of interest for other circumstances where the χ^2-metric is used.

In summary, $\chi^{2\prime}$ is recommended to quantify the amount and extent of usable LD in a population for QTL mapping and MAS for a group of multi-allelic markers when LD is generated by drift alone. However, it must be noted that, while marker–marker LD enables assessment of the general extent of usable LD in populations, high marker–marker LD in specific regions may not necessarily identify regions with high marker–QTL LD; in the simulated data, with LD generated by drift alone, observed LD between two markers was not correlated with LD of these same markers with a bracketed QTL. This implies that, for a given population and when quantified against map distance rather than physical distance, LD between markers and QTL will not necessarily be greater in marker intervals that show strong LD between markers.

The populations under study were simulated with maximum QTL segregation in the founder generation and LD generated by drift alone. Under these circumstances, the effect of mutation on marker–QTL LD should not change our conclusions because mutation rates are generally very low (Falconer & Mackay, 1996). Although selection also causes LD (Bulmer, 1971), it preferentially generates LD between QTL affecting the selected trait rather than between markers and QTL (Farnir *et al.*, 2000). Selection decreases N_e, which accordingly increases LD through the effect of drift. Therefore, our conclusions are expected to hold for populations that are under selection or mutation. Selection can, however, result in differences in LD between genomic regions on the linkage map scale because of selective sweeps (Kim & Nielsen, 2004). This would result in some ability of local marker–marker LD to predict the extent of marker–QTL LD relative to other regions in the genome, unlike what was observed here for LD generated by drift alone.

We are very grateful to Rohan Fernando for providing us with programs used in our simulations. We also thank Laura Grapes, Radu Totir, Eli Heifetz and Janet Fulton for their help and valuable discussion. The editor and reviewers are acknowledged for suggestions that resulted in substantial improvements in the final manuscript. This research was supported by State of Iowa Hatch and Multi-state Research Funds.

Appendix 1. Derivation of sharp bounds for the maximum of χ^2

Consider $2N$ haplotypes with two loci: locus A with k alleles and locus B with m alleles. The frequency of allele A_i at locus A is a_i ($i=1, \ldots, k$), the frequency of allele B_j at locus B is b_j ($j=1, \ldots, m$), and $\sum_{i=1}^{k} a_i = \sum_{j=1}^{m} b_j = 2N$. The frequency of haplotype A_iB_j is x_{ij} such that $\sum_{j=1}^{m} x_{ij} = a_i$, $\sum_{i=1}^{k} x_{ij} = b_j$, $x_{ij} \geqslant 0$. The classical χ^2 contingency table is:

	B_1	…	…	B_m	
A_1	x_{11}			x_{1m}	a_1
…	…	…	…	…	…
…	…	…	…	…	…
A_k	x_{k1}	…	…	x_{km}	a_k
	b_1	…	…	b_m	$2N$

The chi-square statistic for testing for association between alleles is:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(x_{ij} - a_i b_j/(2N))^2}{a_i b_j/(2N)}$$
$$= 2N\left(\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{x_{ij}^2}{a_i b_j} - 1\right) = 2N(g(x) - 1),$$

where

$$g(x) = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{x_{ij}^2}{a_i b_j}.$$

In order to standardize χ^2, we want to find the set of $x = (x_{11}, \ldots, x_{km})$ that can maximize $g(x)$ under the constraints:

$$\sum_{j=1}^{m} x_{ij} = a_i, \quad \sum_{i=1}^{k} x_{ij} = b_j \text{ and } x_{ij} \geqslant 0 \quad (i=1, \ldots, k; j=1, \ldots, m). \tag{A1}$$

However, this is computationally hard (Kalantari *et al.*, 1993). The idea that Kalantari *et al.* (1993) introduced is to replace $g(x)$ by an upper plane $h(x)$ such that $h(x) \geqslant g(x)$:

$$h(x) = \sum_{i=1}^{k}\sum_{j=1}^{m} h_{ij}(x_{ij}) = \sum_{i=1}^{k}\sum_{j=1}^{m} \times \left\{ g_{ij}(l_{ij}) + \frac{g_{ij}(u_{ij}) - g_{ij}(l_{ij})}{u_{ij} - l_{ij}}(x_{ij} - l_{ij}) \right\},$$

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

where $g_{ij}(x_{ij})=\frac{x_{ij}^2}{a_i b_j}$, $l_{ij}=\max(0,\ a_i+b_j-2N)$, $u_{ij}=\min(a_i, b_j)$ and $l_{ij}\leqslant x_{ij}\leqslant u_{ij}$.

Now the question is how to find the set of $x=(x_{11}, \ldots, x_{km})$ that can maximize $h(x)$ under the constraints in (A1). Maximizing $h(x)$ is equivalent to maximizing

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\left\{\frac{g_{ij}(u_{ij})-g_{ij}(l_{ij})}{u_{ij}-l_{ij}}x_{ij}\right\}=\sum_{i=1}^{k}\sum_{j=1}^{m}\{c_{ij}x_{ij}\},$$

where $c_{ij}=\dfrac{g_{ij}(u_{ij})-g_{ij}(l_{ij})}{u_{ij}-l_{ij}}$.

Maximizing $\sum_{i=1}^{k}\sum_{j=1}^{m}\{c_{ij}x_{ij}\}$ under the constraints in (A1) is an ordinary linear transportation problem (Winston, 1991) where c_{ij} can be considered as the 'cost' for cell (i, j) in the χ^2 contingency table. It can be solved by the transportation simplex method (Winston, 1991).

If $\hat{x}$ is the optimal solution to this transportation problem, then $\chi^2_{\max}=2N(h(\hat{x})-1)$ is an upper bound for the maximum of χ^2. Kalantari *et al.* (1993) proved that this upper bound is never worse than the upper bound used in $\chi^{2\prime}$, that is $\chi^2_{\max}\leqslant(2N)\min(k-1, m-1)$.

A C++ program was developed to solve the transportation problem and to obtain $\chi^2_{\max}$ given the allele frequencies at two loci.

Appendix 2. Derivation of regression R^2 for marker–QTL LD

Consider $2N$ haplotypes with two loci: marker A with k alleles and a QTL with two alleles (Q and q). The estimated frequency of allele A_i $(i=1, 2, \ldots, k)$ is $p(A_i)$, and the estimated frequency of allele Q is $p(Q)$.

Let $Y_j=1$ $(j=1, 2, \ldots, 2N)$ if the allele Q is present in the jth haplotype and $Y_j=0$ otherwise.

Let $X_{ij}=1$ $(i=1,\ 2, \ldots, k;\ j=1,\ 2, \ldots, 2N)$ if the allele A_i is present in the jth haplotype and $X_{ij}=0$ otherwise.

Let $Y=[Y_1, \ldots, Y_{2N}]'$, $X_i=[X_{i1},\ \ldots,\ X_{i(2N)}]'$ and $X=[X_1, \ldots, X_k]$.

Then the R^2 for the regression of Y on X (i.e. the proportion of QTL variance explained by marker A) is:

$$R^2=\frac{(\hat{Y}-1\overline{Y})'(\hat{Y}-1\overline{Y})}{(Y-1\overline{Y})'(Y-1\overline{Y})}, \tag{A2}$$

where 1 denotes a vector of $2N$ ones, $\hat{Y}=X(X'X)^{-1}X'Y$, and $\overline{Y}=\frac{1}{2N}\sum_{j=1}^{2N}Y_j$.

First, we calculate the numerator in (A2). Because $X'X=(2N)diag[p(A_1), \ldots, p(A_k)]$ and $X'Y=2N[p(QA_1), \ldots, p(QA_k)]'$ where $p(QA_i)$ is the estimated frequency of haplotype QA_i $(i=1, 2, \ldots, k)$, we obtain $(X'X)^{-1}X'Y=[p(Q|A_1), \ldots, p(Q|A_k)]'$ and $\hat{Y}_j=\sum_{i=1}^{k}[X_{ij}p(Q|A_i)]$ where $p(Q|A_i)$ is the estimated conditional probability of allele Q given A_i $(i=1, 2, \ldots, k)$. Therefore,

$$\begin{aligned}(\hat{Y}-1\overline{Y})'(\hat{Y}-1\overline{Y})&=\sum_{j=1}^{2N}\left[\sum_{i=1}^{k}[X_{ij}p(Q|A_i)]-p(Q)\right]^2\\&=2N\sum_{i=1}^{k}p(A_i)[p(Q|A_i)-p(Q)]^2.\end{aligned} \tag{A3}$$

Second, we calculate the denominator in (A2):

$$\begin{aligned}(Y-1\overline{Y})'(Y-1\overline{Y})&=\sum_{j=1}^{2N}[Y_j-p(Q)]^2\\&=(2N)p(Q)[1-p(Q)].\end{aligned} \tag{A4}$$

From equations (A2), (A3) and (A4), we obtain

$$R^2=\frac{\sum_{i=1}^{k}p(A_i)[p(Q|A_i)-p(Q)]^2}{p(Q)[1-p(Q)]}.$$

References

Ardlie, K. G., Kruglyak, L. & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews: Genetics* **3**, 299–309.

Bulmer, M. G. (1971). The effect of selection on genetic variability. *The American Naturalist* **105**, 201–211.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Dekkers, J. C. M. & Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews: Genetics* **3**, 22–32.

Devlin, B. & Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. Harlow, UK: Addison-Wesley Longman.

Farnir, F., Coppieters, W., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D. & Georges, M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10**, 220–227.

Flint-Garcia, S. A., Thornsberry, J. M. & Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* **54**, 357–374.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 299–309.

Hayes, B. J., Visscher, P. M., McPartlan, H. C. & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research* **13**, 635–643.

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.

Heifetz, E. M., Fulton, J. E., O'Sullivan, N., Zhao, H., Dekkers, J. C. M. & Soller, M. (2005). Extent and

https://doi.org/10.1017/S001667230500769X Published online by Cambridge University Press

consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* (accepted).

Hill, W. G. (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**, 117–126.

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

Kalantari, B., Lari, I., Rizzi, A. & Simeone, B. (1993). Sharp bounds for the maximum of the chi-square index in a class of contingency tables with given marginals. *Computational Statistics & Data Analysis* **16**, 19–34.

Karlin, S. & Piazza, A. (1981). Statistical methods for assessing linkage disequilibrium at the HLA-A, B, C loci. *Annals of Human Genetics* **45**, 79–94.

Kim, Y. & Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513–1524.

Lande, R. & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743–756.

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

Maruyama, T. (1982). Stochastic integrals and their application to population genetics. In *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, pp. 151–166. Tokyo: Japan Scientific Societies Press.

McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Crawford, A. M. & Slate, J. (2002). Linkage disequilibrium in domestic sheep. *Genetics* **160**, 1113–1122.

Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.-Y. & Collins, A. (2001). The optimal measure of allelic association. *Proceedings of the National Academy of Sciences of the USA* **98**, 5217–5221.

Nordborg, M. & Tavar, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**, 83–90.

Nsengimana, J., Baret, P., Haley, C. S. & Visscher, P. M. (2004). Linkage disequilibrium in the domesticated pig. *Genetics* **166**, 1395–1404.

Pritchard, J. K. & Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics* **69**, 1–14.

Sabatti, C. & Risch, N. (2002). Homozygosity and linkage disequilibrium. *Genetics* **160**, 1707–1719.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology* **2**, 125–141.

Taillon-Miller, P., Bauer-Sardiña, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. & Kwok, P.-Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genetics* **25**, 324–328.

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L. & Visscher, P. M. (2003). Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *Journal of Animal Science* **81**, 617–623.

Terwilliger, J. D. & Weiss, K. M. (1998). Linkage disequilibrium mapping of complex disease: fantasy or reality? *Current Opinion in Biotechnology* **9**, 578–594.

Winston, W. L. (1991). *Operations Research: Applications and Algorithms*, 2nd edn. Boston: PWS-Kent.

Yamazaki, T. (1977). The effects of overdominance on linkage in a multilocus system. *Genetics* **86**, 227–236.