

## Linkage Disequilibrium in Domestic Sheep

A. F. McRae,\* J. C. McEwan,\* K. G. Dodds,\* T. Wilson,<sup>†</sup> A. M. Crawford<sup>†</sup> and J. Slate\*<sup>1</sup>

\*AgResearch, Invermay Agricultural Centre, Mosgiel, New Zealand and <sup>†</sup>AgResearch MBU, Department of Biochemistry, University of Otago, Dunedin, New Zealand

Manuscript received June 11, 2001  
Accepted for publication December 20, 2001

### ABSTRACT

The last decade has seen a dramatic increase in the number of livestock QTL mapping studies. The next challenge awaiting livestock geneticists is to determine the actual genes responsible for variation of economically important traits. With the advent of high density single nucleotide polymorphism (SNP) maps, it may be possible to fine map genes by exploiting linkage disequilibrium between genes of interest and adjacent markers. However, the extent of linkage disequilibrium (LD) is generally unknown for livestock populations. In this article microsatellite genotype data are used to assess the extent of LD in two populations of domestic sheep. High levels of LD were found to extend for tens of centimorgans and declined as a function of marker distance. However, LD was also frequently observed between unlinked markers. The prospects for LD mapping in livestock appear encouraging provided that type I error can be minimized. Properties of the multiallelic LD coefficient  $D'$  were also explored.  $D'$  was found to be significantly related to marker heterozygosity, although the relationship did not appear to unduly influence the overall conclusions. Of potentially greater concern was the observation that  $D'$  may be skewed when rare alleles are present. It is recommended that the statistical significance of LD is used in conjunction with coefficients such as  $D'$  to determine the true extent of LD.

WITH the advent of molecular markers the last decade has witnessed a great many experiments to detect quantitative trait loci (QTL) for economically important traits in livestock (*e.g.*, ANDERSSON *et al.* 1994; GEORGES *et al.* 1995). The next challenge awaiting animal geneticists is to determine the actual genes underlying quantitative variation in these traits. To date there have been a few notable successes (GROBET *et al.* 1997; KAMBADUR *et al.* 1997; GALLOWAY *et al.* 2000; MILAN *et al.* 2000; WILSON *et al.* 2001), but positional cloning of QTL remains time consuming, elusive, and is still uncommon. The principal reason why relatively few genes have been discovered is that a typical genome scan maps a QTL to an ~20-cM interval. Thus it is likely that hundreds of genes are within the confidence limits of the QTL, making identification of the desired gene(s) difficult. Even when candidate genes from other populations or species have been identified it is probable that several will map to the identified region.

It has been suggested by both human and livestock geneticists that linkage disequilibrium can be exploited to help fine map QTL (TERWILLIGER and WEISS 1998; KRUGLYAK 1999; HALEY 1999; FARNIR *et al.* 2000). Rather than follow the segregation of marker alleles among related individuals of known phenotype, linkage disequilibrium mapping tests for associations between

marker alleles and trait value and can be applied to large samples of unrelated individuals. Associations will be observed between marker and trait only if the marker and the QTL are in linkage disequilibrium, and so it has been proposed that linkage disequilibrium (LD) mapping has the potential to position QTL to small chromosomal segments (perhaps of the order of <1 cM) provided that LD declines as a function of distance between the marker and the QTL.

Recent reviews have suggested that the efficacy of LD mapping will be dependent on the levels of LD in the study population, heterogeneity of LD across the genome, marker density, and perhaps most importantly the allelic heterogeneity of QTL (TERWILLIGER and WEISS 1998; HALEY 1999; KRUGLYAK 1999). To date these parameters have been explored more thoroughly in humans than any other species, although no firm consensus has been reached. For example, some simulation (KRUGLYAK 1999) and molecular (DUNNING *et al.* 2000) studies have concluded that useful levels of LD are unlikely to extend beyond 3–5 kb, while others have found strong LD extending beyond 1 Mb (TAILLON-MILLER *et al.* 2000; reviewed in BOEHNEKE 2000). Likewise it is unclear how population history has influenced LD in human populations. For example, TAILLON-MILLER *et al.* (2000) and EAVES *et al.* (2000) found similar levels of LD in isolated and mixed populations. However, WILSON and GOLDSTEIN (2000) demonstrated an excess of LD in an admixed population when compared to its founder populations, while REICH *et al.* (2001) showed

<sup>1</sup> Corresponding author: AgResearch, Invermay Agricultural Centre, Puddle Alley, Mosgiel, Private Bag 50034, New Zealand.  
E-mail: jon.slate@agresearch.co.nz

that LD extended considerably further in North Americans of European descent than in a Nigerian population. TAILLON-MILLER *et al.* (2000) showed dramatic heterogeneity in LD on the X chromosome of three different European populations, while both DUNNING *et al.* (2000) and ABECASIS *et al.* (2001) observed greater uniformity in LD within three regions on different chromosomes. It should be pointed out that maps of single nucleotide polymorphisms (SNPs) are likely to be sufficiently dense to allow LD mapping to be performed (SACHIDANANDAM *et al.* 2001), even under the pessimistic scenario of useful LD only extending 3 kb.

In contrast to the accumulating data in human populations, little is known about the extent of linkage disequilibrium in livestock species. It has been suggested that LD will be greater in livestock than humans, as the forces that can generate LD (genetic drift, admixture, selection, and small effective population sizes) are common features of many breeds (HALEY 1999). If LD extends for greater distances in livestock than man, then it is possible that LD mapping can be successfully applied to map QTL using relatively sparsely spaced markers, the caveat to this being that fine mapping of QTL may be achieved only with limited resolution. We are aware of just one study in which LD has been estimated in an agriculturally important breed—FARNIR *et al.*'s (2000) analysis of Dutch black and white dairy cattle. In that study a battery of 281 microsatellite markers was used to estimate genome-wide LD. The authors demonstrated that considerable LD extended for at least 20 cM and was very high for markers spaced <5 cM. However, as with analysis of human data sets, marker distance was not a particularly good predictor of LD, suggesting some heterogeneity in LD across the genome. Of greater concern was the relatively frequent detection (12%) of significant associations (at  $P < 0.05$ ) between pairs of unlinked markers. Thus, LD mapping in livestock may require techniques that simultaneously test for linkage and allelic associations (FARNIR *et al.* 2000). Clearly, more data on the extent of LD in livestock species are required before general LD mapping strategies are developed.

In this article we measure linkage disequilibrium in populations of Coopworth and Romney sheep. Coopworths were developed from the Border Leicester and Romney breeds during the 1960s in New Zealand. It is a dual-purpose breed (with equal emphasis on meat and wool) and is commonly used in the sheep industry. As the breed is only ~10 generations old, it is likely that substantial LD exists within present-day flocks. Furthermore, it is possible that alternate QTL alleles that were fixed in the founder breeds are segregating within today's flocks. Thus, Coopworths appear to be a good starting point when considering strategies for livestock linkage disequilibrium mapping (HALEY 1999). Romneys are the most widely farmed breed in New Zealand and are also dual purpose.

## MATERIALS AND METHODS

**Experimental design:** *Data set 1:* In a QTL mapping experiment described elsewhere (RAADSMAN *et al.* 1999; CRAWFORD 2001) two Texel × Coopworth sires were crossed to a total of 186 Coopworth dams resulting in 276 progeny. For the purposes of measuring LD we used genotypes at 90 microsatellites mapped to sheep chromosomes 1–10. Genotypes were available from progeny, sires, and grandsires, but not from the Coopworth dams.

*Data set 2:* First-pass genome-wide scans (such as that used in data set 1) typically use sets of polymorphic microsatellites spaced at regular 10- to 20-cM intervals. Thus data set 1 provided little information on LD between tightly linked markers. To circumvent this problem a second data set, originally generated to map the Booroola fecundity gene (MONTGOMERY *et al.* 1993; WILSON *et al.* 2001), was used. Fourteen sires were crossed with ~400 Romney, Perendale (another Romney-derived breed), and Coopworth dams, resulting in 482 offspring. The majority of dams were Romneys. Progeny and sires were typed at 26 markers [13 microsatellites and 13 restriction fragment length polymorphisms (RFLPs)] on chromosome 6. Again genotypes were available for sires and progeny but not from the dams. Unlike data set 1, genotypes were unavailable from the parents of the sires.

**Haplotype determination:** Linkage disequilibrium is most readily measured using haplotypes rather than multilocus genotypes (LYNCH and WALSH 1998, p. 97). Using genotype information in the sires and progeny, it was possible to determine the maternal haplotype transmitted to each progeny. Thus, while the microsatellite genotypes were obtained in composite breed animals, linkage disequilibrium was measured in the ungenotyped Coopworth dams (data set 1) and a mixed population of Coopworth, Romney, and Perendale dams (data set 2). The method for deriving maternally transmitted haplotypes is described below (see also Figure 1).

*Data set 1:* Initially the marker phase for each sire was determined by taking reference to the grandparents' genotypes. The sire allele at each marker was identified in all progeny. In cases where the inherited sire allele was ambiguous (when the sire and progeny had the same genotype), the sire-inherited alleles at adjacent markers were first used to identify the inherited sire haplotype. In cases where a recombination occurred between the adjacent markers, observed allele frequencies in haplotypes with the same adjacent sire alleles were used to select the most likely sire-inherited allele. Any remaining ambiguities were resolved by comparison of the frequency of the rival alleles among the dam population. The dam allele was inferred by the elimination of the sire allele. Thus, the haplotype inherited from the dam for every progeny was determined.

*Data set 2:* Although paternal grandparents were not genotyped the sire haplotypes could be reconstructed by identification of frequently cosegregating alleles at linked loci in the progeny. Ambiguous sire alleles and haplotypes inherited from the dams were then determined as in data set 1.

**Measuring linkage disequilibrium:** A variety of linkage disequilibrium measures have been discussed in some detail elsewhere (HEDRICK 1987; LEWONTIN 1988; DEVLIN and RISCH 1995; JORDE 2000), with the general conclusion that different measures are "best" depending on the question being asked. We measured LD in two ways. First we used HEDRICK's (1987) multiallelic extension of LEWONTIN's (1964) normalized coefficient  $D'$ . We chose this measure as it allows LD to be measured with highly polymorphic markers, is allegedly independent of allele frequencies (HEDRICK 1987; ZAPATA 2000; but see LEWONTIN 1988), and also allows comparison to the data presented in FARNIR *et al.* (2000).

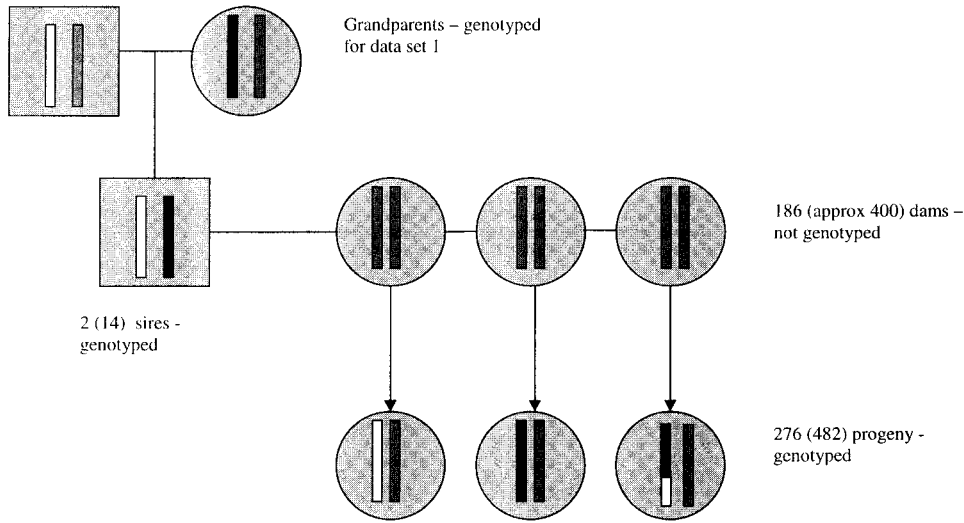


FIGURE 1.—Experimental design used to infer gametic haplotypes. For both data sets sires and progeny were genotyped while dams were not. Sample sizes refer to data set 1 with equivalent amounts for data set 2 in parentheses. Genotypes from paternal grandparents were additionally available for data set 1. Sire haplotypes are represented by black, white, and recombinant rectangles. By inferring the sire haplotype in each progeny, the dam gametic haplotype (gray rectangle) can be determined by a process of elimination.

Using the software package GOLD (ABECASIS and COOKSON 2000), HEDRICK'S (1987) measure of LD between two multiallelic markers was calculated as

$$D' = \frac{\sum_{i=1}^u \sum_{j=1}^v p_i q_j |D'_{ij}|}{D_{\max}}$$

where  $u$  and  $v$  are the number of alleles at each marker,  $p_i$  is the frequency of allele  $i$  at the first marker, and  $q_j$  is the frequency of allele  $j$  at the second marker.  $|D'_{ij}|$  is the absolute value of Lewontin's normalized LD measure (LEWONTIN 1964) calculated as

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

where

$$D_{ij} = x_{ij} - p_i q_j$$

and

$$D_{\max} = \begin{cases} \min[p_i q_j, (1 - p_i)(1 - q_j)]; & D_{ij} < 0 \\ \min[p_i(1 - q_j), (1 - p_i)q_j]; & D_{ij} > 0, \end{cases}$$

where  $x_{ij}$  is the frequency of gametes with alleles  $i$  at the first marker and  $j$  at the second marker and  $p_i$  and  $q_j$  are the frequencies of allele  $i$  at the first marker and allele  $j$  at the second marker.

LD was also measured by a test of independence between alleles at pairs of loci. This was implemented in Arlequin (SCHNEIDER *et al.* 2000), using a Markov chain extension to Fisher's exact test for  $R \times C$  contingency tables (SLATKIN 1994). The probability of finding a table with the same marginal totals that has a test statistic equal to or more extreme than the observed table was estimated using the Markov chain to efficiently explore the sample space. All probabilities were estimated with a standard error of  $<0.0025$ . No attempt was made to correct for multiple testing, as the large number of tests would result in very low statistical power.

**Examining the properties of  $D'$ :** As  $D'$  is a sum of absolute values it can take only a positive value (or zero). Simulation was used to determine the distribution of  $D'$  under the null scenario of no linkage disequilibrium. This was achieved by resampling data set 1 without replacement and randomizing genotypes at each locus across individuals.  $D'$  was then recalculated for each marker pair as for the real data set. The effect of sample size on  $D'$  was also examined. Populations of 100, 200, 400, 1000, and 2000 individuals were created by sampling

multilocus haplotypes from data set 1 with replacement. Each population was replicated 10 times and  $D'$  for each marker pair was recalculated. It was expected that mean  $D'$  across all marker pairs would decrease as a function of sample size, as the influence of rare alleles would diminish for larger populations. Thus, it was possible to examine whether estimates of  $D'$  from the real data were upwardly biased.

**Testing the independence of  $D'$  on allele frequency:** FARNIR *et al.* (2000) found that levels of LD varied significantly between chromosomes, but attributed interchromosomal differences to a partial dependence of  $D'$  on marker information content. Using nonsynthetic marker pairs the relationship between  $D'$  and marker heterozygosity was investigated, using a linear regression model with mean heterozygosity as the predictor and  $D'$  the response variable. Heterozygosity was tested as both a linear and a quadratic term.

The expected heterozygosity at a locus was calculated as

$$1 - \sum_i p_i^2,$$

where  $p_i$  is the estimated frequency of the  $i$ th allele at the locus.  $D'$  values for nonsynthetic marker pairs were transformed to the continuous variable  $C$ , where

$$C_{ij} = \ln\left(\frac{D'_{ij}}{1 - D'_{ij}}\right),$$

allowing a normal error structure to be assumed in regression models.

**Correcting  $D'$  for heterozygosity:** As mean heterozygosity was found to explain significant variation in  $D'$  for nonsynthetic marker pairs we adjusted  $D'$  values for all marker pairs using the following models.

Nonsynthetic markers were corrected by fitting

$$C_{ij} = \alpha + \beta_1 H_{ij} + \beta_2 H_{ij}^2 + \epsilon_{ij},$$

where  $H_{ij}$  is the average heterozygosity for markers  $i$  and  $j$ .

The residuals from this model,  $e_{ij}$ , were saved and  $C^*$  was formed as

$$C_{ij}^* = \hat{\alpha} + \hat{\beta}_1 \bar{H} + \hat{\beta}_2 \bar{H}^2 + e_{ij}$$

where  $\bar{H}$  is the average heterozygosity taken over all markers.  $C^*$  values were transformed back to the original scale by

$$D_{ij}^* = \frac{\exp(C_{ij}^*)}{1 + \exp(C_{ij}^*)}.$$

$D'$  values for syntenic marker pairs were corrected for heterozygosity by taking

$$C_{ij}^* = C_{ij} + \hat{\beta}_1(\bar{H} - H_{ij}) + \hat{\beta}_2(\bar{H}^2 - H_{ij}^2)$$

and transforming back to the original scale as for nonsyntenic marker pairs.

**Testing for interchromosomal variation in LD:** We performed one-way ANOVA (with chromosome as a factor) on  $D'$  for syntenic marker pairs to test for interchromosomal variation in LD. FARNIR *et al.* (2000) used nonsyntenic pairs to test for significant interaction terms between chromosomes—the presence of which may indicate chromosomal regions that have undergone selection for economically important traits. We chose not to perform a similar analysis for several reasons. First, the large number of cells in the interaction term make type I error a possibility. Second, analyzing only nonsyntenic marker pairs leads to problems of nonorthogonality, while the inclusion of syntenic marker pairs incorporates the potentially confounding effect of linkage. Third, the presence of a significant interaction term does not necessarily indicate previous selection of QTL—for example, cells with a relatively high proportion of markers with rare alleles may result in a significant interaction term due to  $D'$  being skewed in the presence of rare alleles. Finally we note that FARNIR *et al.* (2000) found no evidence for significant interaction terms, suggesting the power of the test is low—it seems probable that their population did contain chromosomal regions harboring previously coselected QTL.

**Measuring distance between markers:** All markers were previously mapped in sheep (DE GORTARI *et al.* 1998; MADDOX *et al.* 2001). To confirm marker order in our flocks we constructed linkage maps using the BUILD, CHROMPIC, and FLIPS options of the linkage mapping software CRIMAP (GREEN *et al.* 1990). No discrepancy between published order and CRIMAP order was detected. However, we chose to use published map distances for all subsequent analyses, as the additional markers and larger number of meioses in the international mapping flock (IMF) are likely to provide improved accuracy of interval length. Note that the two Texel-Coopworth sires in data set 1 were additionally members of the IMF.

## RESULTS

**Data set 1:** Linkage disequilibrium was estimated for 417 syntenic marker pairs and 3299 nonsyntenic marker pairs. LD could not be estimated for marker pairs when neither sire was heterozygous for both markers. Among the syntenic marker pairs 175 were separated by <60 cM. Of these, 12 pairs were separated by <10 cM and a further 28 pairs by 10–20 cM. Figure 2a shows the relationship between marker distance and  $D'$  for data set 1. Gametic disequilibrium between linked markers is expected to decline by  $(1 - c)^n$  over  $n$  generations, where  $c$  is the recombination fraction between markers. As expected,  $D'$  declined as a function of the distance between markers. Marker distance (log<sub>10</sub> transformed) was significantly and negatively correlated with  $D'$  for markers spaced <60 cM ( $r = -0.341$ ,  $P < 0.0001$ ). However, markers separated by <10 cM had lower  $D'$  values than markers separated by similar distances in the dairy cattle population of FARNIR *et al.* (2000). The maximum  $D'$  value observed between any two linked markers was 0.52 (the theoretical maximum is 1.0). Sam-

ple sizes were inadequate to determine whether  $D'$  declined appreciably with distance between 1 and 5 cM.

Gametic disequilibrium was also determined for the 3299 nonsyntenic marker pairs. Figure 3 shows the distribution of  $D'$  values for nonsyntenic marker pairs. The distributions for syntenic and nonsyntenic markers were strongly overlapping, although syntenic pairs did have a significantly higher mean  $D'$  [syntenic pairs mean  $D' = 0.211$  (SE = 0.004), *vs.* nonsyntenic pairs mean  $D' = 0.196$  (SE = 0.001);  $t_{3714} = 4.30$ ;  $P < 0.0001$ ].

Linkage disequilibrium was also assessed by measuring the significance of allelic associations. Figure 4 illustrates the cumulative frequency of the  $P$  values obtained with a Markov chain approach to determine the significance of LD. Significant LD was observed more frequently for syntenic markers separated by <60 cM than for nonsyntenic markers or linked markers separated by >60 cM. Among the 175 marker pairs separated by <60 cM, 60 (34.3%) were in significant (at  $P < 0.05$ ) linkage disequilibrium. In contrast, only 30/242 (12.4%) of marker pairs separated by >60 cM were in significant LD. Among nonsyntenic marker pairs, significant linkage disequilibrium was observed more than twice as often as expected under random segregation (380/3299 or 11.5%).

The dependence of  $D'$  on marker heterozygosity was examined using nonsyntenic marker pairs. Mean heterozygosity (of the two markers) was significantly associated with  $D'$  whether fitted as a linear or as a linear plus quadratic term (both  $P < 0.0001$ ), although slightly greater variance in  $D'$  was explained by the linear plus quadratic term. Highly variable marker pairs tended to have higher  $D'$  scores (see Figure 5), although a small number of data points with high  $D'$  involved the least variable marker (RM65 on sheep chromosome 1). As  $D'$  was partially dependent on marker variability we corrected syntenic and nonsyntenic  $D'$  values for heterozygosity (see MATERIALS AND METHODS). Figure 2b shows the relationship between corrected  $D'$  and distance for syntenic markers separated by <60 cM. Corrected  $D'$  (here termed  $D^*$ ) showed essentially the same relationship with distance as  $D'$  and declined as a function of log-transformed distance ( $r = -0.343$ ,  $P < 0.0001$ ). Thus, while we have demonstrated a significant relationship between  $D'$  and marker heterozygosity we do not believe that any such dependency has unduly influenced our overall conclusions.

A one-way ANOVA comparing mean  $D'$  across chromosomes 1–10 provided no evidence for interchromosomal heterogeneity in LD ( $F_{9,407} = 0.44$ ,  $P = 0.92$ ). Chromosomal effects were also tested by first fitting marker distance as an additional term in a general linear model, although there was still no evidence for interchromosomal variation in  $D'$  (data not shown). The distributions of  $D'$  values for each chromosome are shown in Figure 6. Linkage disequilibrium appears uni-

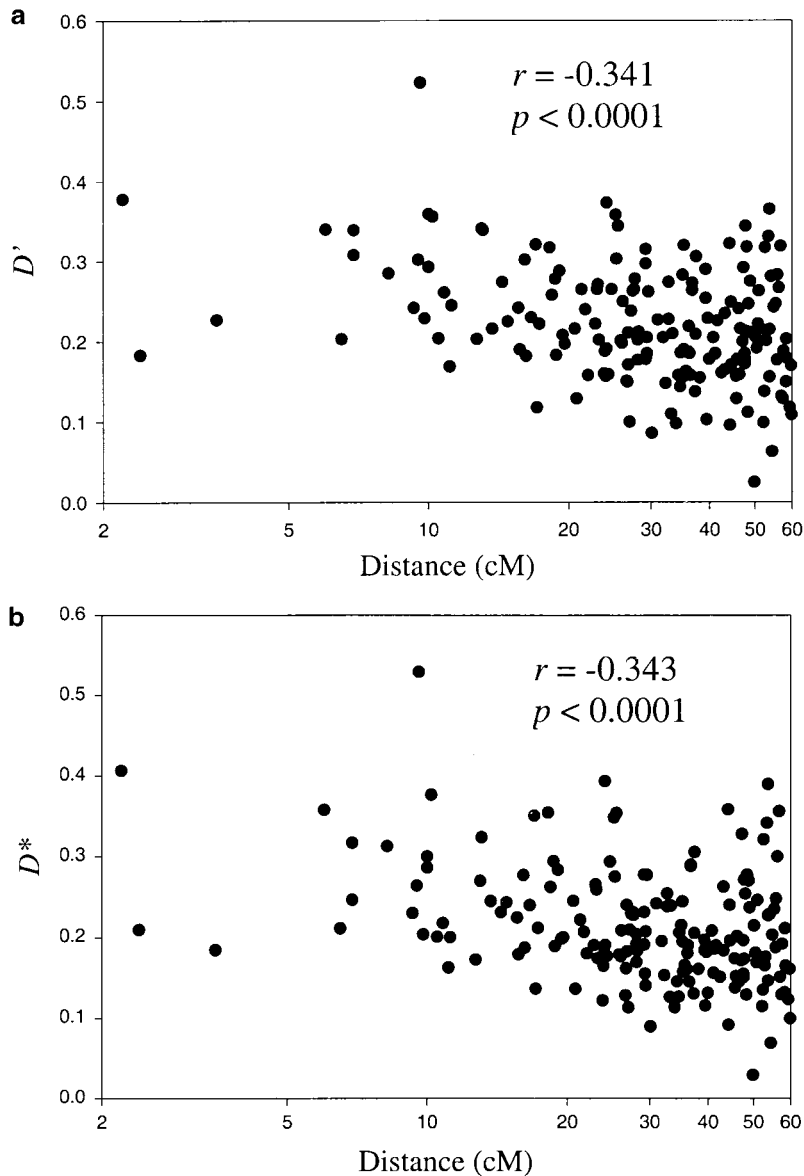


FIGURE 2.—(a) Linkage disequilibrium as a function of distance ( $\log^{10}$  transformed) for data set 1. LD was measured using the metric  $D'$  for syntenic markers separated by  $\leq 60$  cM. (b) Linkage disequilibrium as a function of distance for data set 1.  $D'$  scores were corrected for mean marker heterozygosity, giving the related metric  $D^*$ . Note the similarity to Figure 2a, where  $D'$  was not adjusted for heterozygosity.

form across chromosomes, although heterogeneity within individual chromosomes cannot be discounted.

One final potentially confounding factor in our analysis of LD concerns the pedigree structure used in our haplotype reconstruction. A number of progeny were full-sibs, such that maternal gametic haplotypes may have been reconstructed more than once. Theoretically, the duplication of maternal gametes may have artificially inflated our estimates of LD. We repeated our analyses, but considered only one randomly chosen progeny of each dam.  $D'$  was recalculated for pairs of nonsyntenic loci (to avoid the influence of linkage between markers) and compared to values obtained from the full data set. However, a paired  $t$ -test revealed that  $D'$  values were actually lower for the full data set than for the reduced data set [reduced data set, mean (SE)  $D' = 0.226$  (0.001); full data set, mean (SE)  $D' = 0.196$  (0.001);  $t_{3298} = 40.63$ ,  $P < 0.0001$ ]. Thus, there is no evidence

that the inclusion of multiple copies of dam haplotypes has led to an upward bias in estimates of LD.

**Data set 2:** Data set 1 provided limited information regarding the relationship between marker distance and linkage disequilibrium over short distances. Therefore we derived estimates of LD across 165 cM of sheep chromosome 6, using a panel of 26 markers (13 microsatellites and 13 RFLPs). These markers were used to fine map the Booroola fecundity locus (FecB; MONTGOMERY *et al.* 1993; WILSON *et al.* 2001) and the majority (17 markers) are concentrated within a 30-cM region. As  $D'$  can become skewed by rare alleles, we excluded any marker pairs where 50 or fewer dam haplotypes could be inferred.  $D'$  was estimated for a total of 169 syntenic marker pairs, of which 134 were separated by  $< 60$  cM. Of these, 34 pairs were separated by  $< 10$  cM while a further 26 pairs were 10–20 cM apart.  $D'$  was, on average, greater for syntenic markers  $< 60$  cM apart in data set

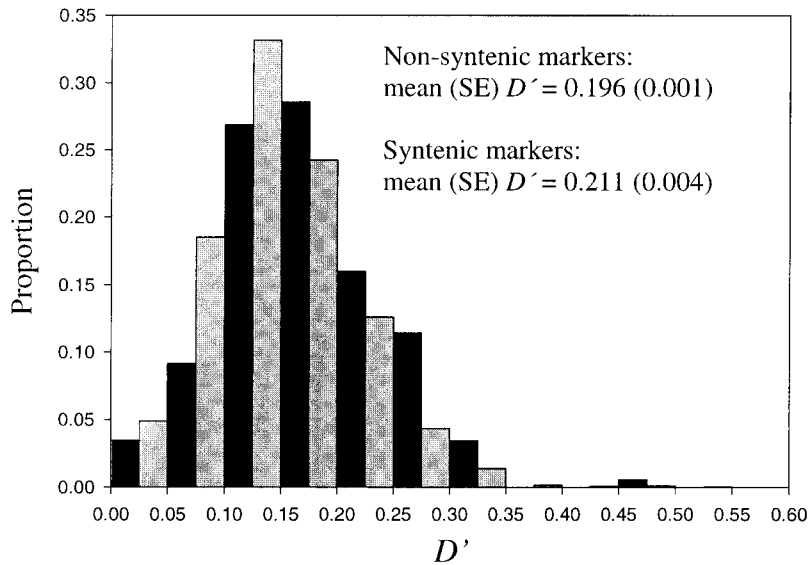


FIGURE 3.—Frequency distributions of  $D'$  for data set 1. Nonsyntenic marker pairs are represented by gray bars while syntenic marker pairs <60 cM apart are represented by black bars. The two distributions are strongly overlapping. For nonsyntenic markers the most numerous class is  $0.10 < D' < 0.15$ , while  $0.15 < D' < 0.20$  is most common for syntenic markers.

2 than in data set 1 [data set 2, mean (SE)  $D' = 0.274 (0.010)$ ]. The expected decay of LD with distance was more apparent than in data set 1 [correlation coefficient between distance ( $\log_{10}$  transformed) and  $D' = -0.536$ ,  $P < 0.0001$ ; see Figure 7). Among the 169 syntenic marker pairs linkage disequilibrium was significant (at  $P < 0.05$ ) for 47/134 (35.1%) of pairs separated by <60 cM but was not significant for any of the 35 pairs separated by >60 cM. A decline of LD with distance was apparent between 0 and 10 cM (Figure 7), suggesting that LD may be useful for fine-scale mapping in domestic sheep. However, there did appear to be considerable heterogeneity in LD over short distances. For example,  $D'$  varied between 0.44 and 0.70 for markers separated by 0–2 cM.

**Simulations:** Data set 1 was used to examine the distribution of  $D'$  under the null hypothesis of no LD. By randomizing genotype at each locus independently of other loci, any allelic associations can be attributed to chance sampling rather than population structure or admixture. The distribution of  $D'$  under this null scenario was remarkably similar to that obtained for nonsyntenic marker pairs for the real data. For the simulated data,  $D'$  took a mean of 0.189 (SE = 0.001), and had a variance of 0.004. The maximum value that  $D'$  took for a nonsyntenic pair was 0.52. Thus, the distribution of  $D'$  for nonsyntenic pairs is compatible with the null scenario of no gametic disequilibrium. Unlinked loci with high  $D'$  scores may result from chance sampling. Simulations were also performed to examine the ef-

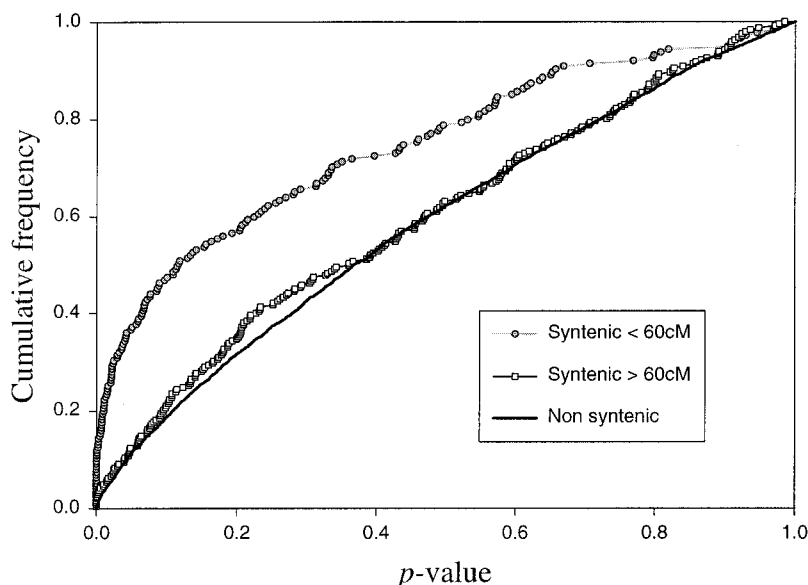


FIGURE 4.—Cumulative frequency plots of the significance of LD between markers in data set 1. Marker pairs are classified as nonsyntenic, syntenic but separated by >60 cM, or syntenic and separated by <60 cM.  $P$  values were obtained using Arlequin (see MATERIALS AND METHODS).

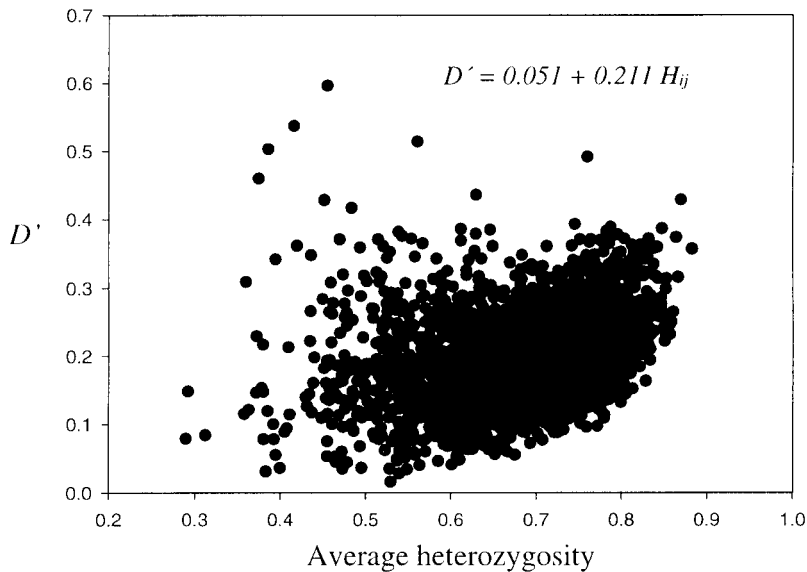


FIGURE 5.—The relationship between mean marker heterozygosity and  $D'$  for nonsyntenic marker pairs in data set 1.  $D'$  increases with mean marker heterozygosity, although high  $D'$  values can also be obtained when one marker contains rare alleles.

fect of sample size on  $D'$  (see Figure 8). When small samples ( $n = 100$ ) are analyzed,  $D'$  can be upwardly biased by as much as 0.10. Extrapolating from Figure 8, it is estimated that the average bias for data set 1 was 0.05 and for data set 2 was 0.025.

#### DISCUSSION

In this article, we used genotypes at microsatellite markers to estimate linkage disequilibrium in domestic sheep. As expected in a population that has undergone recent admixture, a small effective population size, and intense selection, LD was considerable across all 10 chromosomes considered. However, despite the Coopworth breed being perhaps only 8–10 generations old, there was an appreciable decline of LD with marker distance, suggesting that LD mapping may be feasible in this

population. Perhaps the most striking features of this study are the similarities to an earlier analysis of Dutch black and white dairy cattle (FARNIR *et al.* 2000)—the only other livestock breed for which LD has been measured. In both studies LD was considerable and declined as a function of distance. However, both populations exhibited substantial LD between pairs of unlinked markers. Thus, our data support the contention of FARNIR *et al.* (2000) that LD mapping approaches in livestock will need to test simultaneously for linkage and linkage disequilibrium to minimize the risk of type I error. While data on more populations are clearly required before general conclusions are reached, the available evidence suggests that LD mapping is a viable approach for livestock QTL discovery, although fine-mapping resolution may be limited. As recombination is expected to rapidly reduce LD between a QTL and all

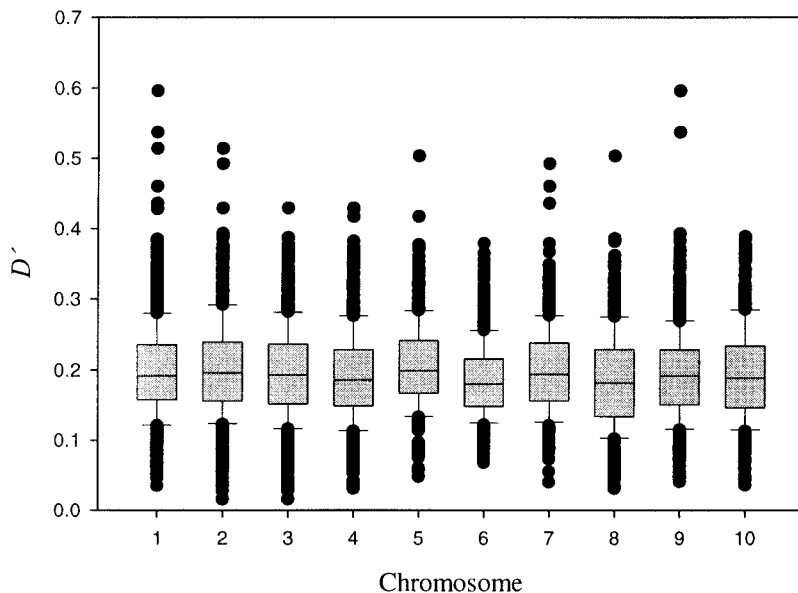


FIGURE 6.—Box plots showing the distributions of  $D'$  for chromosomes 1–10 for data set 1. Horizontal lines represent the mean, boxes range from the 25th to 75th percentile, and whiskers represent the 5th and 95th percentiles. Outliers are represented by additional data points.  $D'$  scores appear to be uniform across chromosomes.

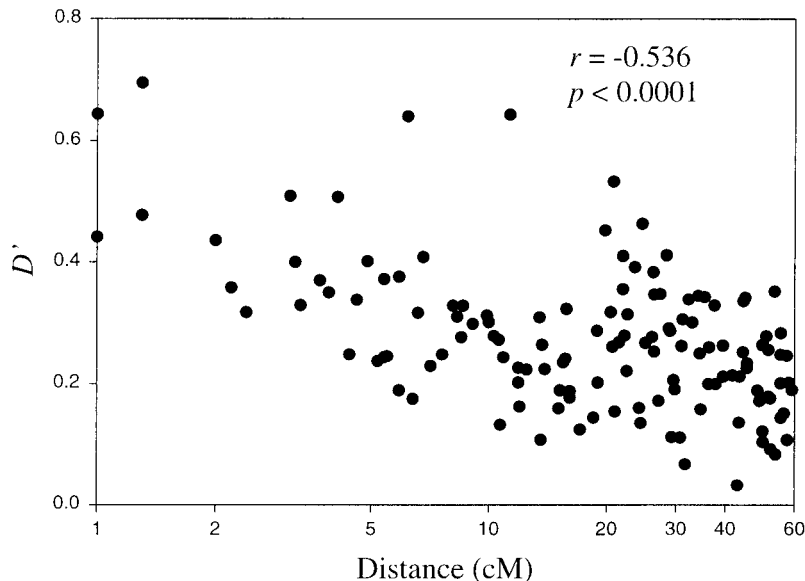


FIGURE 7.—LD (measured by  $D'$ ) as a function of  $\log_{10}$ -transformed distance for data set 2. Data set 2 includes more marker pairs separated by  $\leq 20$  cM than data set 1.

but its closest markers it should be possible to improve mapping resolution in a relatively small number of subsequent generations. Alternatively, LD mapping in slightly older breeds with a greater number of meioses since population admixture may permit more precise QTL resolution.

One potential source of bias in our data set concerns the haplotype reconstruction process. Where sire and progeny had the same heterozygous genotype, the genotype at flanking markers was used to infer which sire allele the progeny had inherited at the ambiguous locus. An underlying assumption of this process was that the progeny was not a double recombinant. For data set 1  $\sim 20\%$  of genotypes were inferred in this way. The mean marker interval was 20.0 cM. Thus it is anticipated that  $\sim 0.008$  ( $0.20 \times 0.20^2$ ) of genotypes are wrongly inferred due to undetected double recombinants. This equates

to an average of two genotypes per locus. Furthermore, double recombinants should be independent across loci and across individuals, so no systematic bias is expected. Alternatively, ambiguous genotypes could have been omitted, but this would have led to a 20% reduction in the sample size and would have caused an upward bias in estimates of  $D'$  (see Figure 8).

LD in both populations could be caused by two processes—population admixture and population structure attributable to recent coancestry. The Coopworth breed is 8–10 generations old and population genetics theory predicts that disequilibrium due to admixture should have declined to negligible levels for nonsynthetic markers, provided that the population was randomly mating and reasonably large (LYNCH and WALSH 1998, p. 96). Although relatively high  $D'$  scores were observed between nonsynthetic markers, the distribution of  $D'$  was

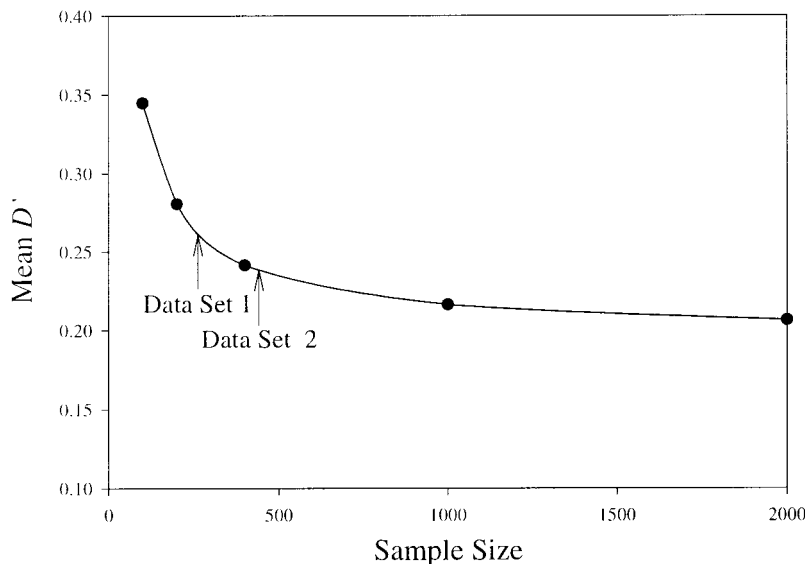


FIGURE 8.—The effect of sample size on  $D'$  as determined by simulation. Values of  $D'$  decline as a function of sample size. The sample sizes of data sets 1 and 2 are marked, suggesting upward biases of 0.05 and 0.025, respectively.



virtually identical to that obtained under the null hypothesis of no LD. However, the dam population used in data set 1 came from a single flock and some individuals are likely to have a history of recent coancestry. A similar argument has been made to explain nonsyntenic associations in dairy cattle (FARNIR *et al.* 2000). Data set 2 contained ewes from several different breeds. Although all of the breeds were Romney derived, it might be anticipated that differences in allele frequency between breeds would contribute to LD. However, a closer examination of data set 2 suggests that admixture has not contributed to LD in this population. It is known that some of the sires in data set 2 were mated to groups containing only Romney ewes, while other sires were mated to groups containing all three breeds. The groups containing all three breeds did not contain any alleles that were absent in the Romney-only groups. Furthermore, when we calculated  $D'$  from the Romney-only groups, the magnitude of LD and its relationship with distance were almost identical to the patterns observed in Figure 7 (data not shown). Thus it seems that admixture contributed negligibly to LD in this population.

Although a considerable number of linkage disequilibrium coefficients have been developed (for reviews see HEDRICK 1987; DEVLIN and RISCH 1995; JORDE 2000) the majority are suitable only for biallelic markers. While biallelic SNP markers are likely to predominate in LD mapping studies in human populations, it has been suggested that multiallelic microsatellites will continue to prove useful in livestock mapping as a lower marker density may be sufficient (FARNIR *et al.* 2000). Regardless of which marker type is used in future studies there are many available livestock microsatellite data sets with which to estimate LD. However, the properties of the most widely used multiallelic LD estimator—HEDRICK'S (1987) extension of LEWONTIN'S (1964)  $D'$ —are not yet well understood. Previous analyses have suggested that  $D'$  is relatively robust to variation in allele frequency (HEDRICK 1987; ZAPATA 2000), although no measure of LD is entirely independent of allele frequency (LEWONTIN 1988). We detected a significant relationship between  $D'$  and mean marker heterozygosity, with LD generally being greater for more variable marker pairs. Despite this relationship, it seems that variation in marker heterozygosity did not unduly influence our results. For example, the plots of  $D'$  (Figure 2a) and heterozygosity-adjusted  $D'$  (Figure 2b) against marker distance were almost identical for syntenic markers in data set 1. Perhaps of greater concern is the possibility that  $D'$  can be skewed when one or both markers contain rare alleles (EYRE-WALKER 2000). Data set 1 provided some examples of  $D'$  apparently being distorted by rare alleles. In the top left quadrant of Figure 5 are a number of data points that represent nonsyntenic marker pairs with low mean heterozygosity yet high  $D'$ . The majority of these points represent marker pairs that contain the microsatellite locus

RM65—the least variable marker in data set 1. Despite the high  $D'$  values obtained with these marker pairs, LD was not statistically significant, suggesting that the high  $D'$  scores are attributable to rare alleles at RM65. Clearly no measure is ideal for determining the extent of LD between multiallelic markers. One solution is to consider both an LD coefficient (such as  $D'$ ) and the statistical significance (as determined by a Markov chain approximation of Fisher's exact test) of any association between markers.

A further complication when using  $D'$  to measure LD is the influence of sample size. For small sample sizes  $D'$  is upwardly biased, leading to overestimates in LD. Simulations suggest that the sample sizes used in data sets 1 and 2 may have led to an overestimate of  $D'$  by 0.05 and 0.025, respectively (Figure 8). This bias is consistent for syntenic and nonsyntenic marker pairs. Thus the overall conclusions are not affected by limitations of our sample size. It is notable that several studies of LD in human populations have relied on considerably smaller samples than are described here (*e.g.*, TAILLON-MILLER *et al.* 2000; WILSON and GOLDSTEIN 2000; REICH *et al.* 2001).

There is one additional plausible explanation for the observation of nonsyntenic marker pairs with low heterozygosity and high  $D'$  scores. It is highly likely that QTL for favorable wool or meat characteristics have been under strong selection during recent domestication events. Regions harboring QTL may be selected in tandem, leaving "signatures" of low variability at adjacent markers and strong linkage disequilibrium between the two regions. However, if this were the case then one would also expect the LD between the regions to be statistically significant—an observation that we did not make for high  $D'$ -low variability marker pairs in data set 1. Thus we believe that in this study nonsyntenic marker pairs with low variability and high  $D'$  are attributable to rare alleles rather than the presence of QTL. This is probably an area worthy of further investigation, both with simulation studies and perhaps by measuring LD in breeds with large numbers of identified QTL.

Finally, we note that there has been some concern in the human genetics literature that allelic heterogeneity may be a common phenomenon for many complex traits (TERWILLIGER and WEISS 1998; KRUGLYAK 1999; JORDE 2000). Thus, associations between particular marker alleles and phenotypes can be weak even when the marker and QTL are very closely linked, making LD mapping problematic. In contrast, livestock have undergone recent admixture, selection, and drift—all of which are forces that are likely to reduce allelic heterogeneity (HALEY 1999). In summary we cautiously believe that LD mapping is likely to become an increasingly used tool by livestock geneticists in their attempts to determine the actual loci responsible for variation of economically important traits. A first step is to establish the extent of LD in more populations of interest.

We thank A. Beattie, G. Greer, K. Knowler, E. Lord, J. Lumsden, P. McDonald, and G. Montgomery for management, data recording, and genotyping of the animals used in this study. P. Visscher, M. Tate, A. Campbell, and three anonymous referees made helpful comments on an earlier draft of the manuscript. This work was funded by an AgResearch summer student bursary (A.F.M.) and by the Royal Society (J.S.). The genotypes of animals used in this study were generated as part of projects funded by the New Zealand Foundation of Research, Science and Technology.

## LITERATURE CITED

- ABECASIS, G. R., and W. O. C. COOKSON, 2000 GOLD—graphical overview of linkage disequilibrium. *Bioinformatics* **16**: 182–183.
- ABECASIS, G. R., E. NOGUCHI, A. HEINZMANN, J. A. TRAHERNE, S. BHATTACHARYYA *et al.*, 2001 Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**: 191–197.
- ANDERSSON, L., C. S. HALEY, H. ELLEGREN, S. A. KNOTT, M. JOHANSSON *et al.*, 1994 Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* **263**: 1771–1774.
- BOEHNKE, M., 2000 A look at linkage disequilibrium. *Nat. Genet.* **25**: 246–247.
- CRAWFORD, A. M., 2001 A review of QTL experiments in sheep. *Proc. Assoc. Adv. Anim. Breed. Genet.* **14**: 33–38.
- DE GORTARI, M. J., B. A. FREKING, R. P. CUTHBERTSON, S. M. KAPPES, J. W. KEELE *et al.*, 1998 A second-generation linkage map of the sheep genome. *Mamm. Genome* **9**: 204–209.
- DEVLIN, B., and N. RISCH, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311–322.
- DUNNING, A. M., F. DUROCHER, C. S. HEALEY, M. D. TEARE, S. E. MCBRIDE *et al.*, 2000 The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67**: 1544–1554.
- EAVES, I. A., T. R. MERRIMAN, R. A. BARBER, S. NUTLAND, E. TUOMILEHTO-WOLF *et al.*, 2000 The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **25**: 320–323.
- EYRE-WALKER, A., 2000 Do mitochondria recombine in humans? *Philos. Trans. R. Soc. Lond. Ser. B* **355**: 1573–1580.
- FARNIR, F., W. COPPETIERS, J.-J. ARRANZ, P. BERZI, N. CAMBISANO *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.* **10**: 220–227.
- GALLOWAY, S. M., K. P. McNATTY, L. M. CAMBRIDGE, M. P. E. LAITINEN, J. L. JUENGEL *et al.*, 2000 Mutations in an oocyte-derived growth factor gene (BMP15) cause increased ovulation rate and infertility in a dosage-sensitive manner. *Nat. Genet.* **25**: 279–283.
- GEORGES, M., D. NIELSEN, M. MACKINNON, A. MISHRA, R. OKIMOTO *et al.*, 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907–920.
- GREEN, P., K. FALLS and S. CROOKS, 1990 Documentation for CRI-MAP. Washington University, St Louis.
- GROBET, L., L. J. R. MARTIN, D. PONCELET, D. PIRROTIN, B. BROWERS *et al.*, 1997 A deletion in the bovine myostatin gene causes the double muscled phenotype in cattle. *Nat. Genet.* **17**: 71–74.
- HALEY, C., 1999 Advances in quantitative trait locus mapping, pp. 47–59 in *From J. L. Lush to Genomics: Visions for Animal Breeding and Genetics*, edited by J. C. M. DEKKERS, S. J. LAMONT and M. F. ROTHSCHILD. Iowa State University, Ames, IA.
- HEDRICK, P., 1987 Gametic disequilibrium measures: proceed with caution. *Genetics* **117**: 331–341.
- JORDE, L. B., 2000 Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**: 1435–1444.
- KAMBADUR, R., M. SHARMA, T. P. L. SMITH and J. J. BASS, 1997 Mutations in myostatin (GDF8) in double-muscled Belgian blue and Piedmontese cattle. *Genome Res.* **7**: 910–916.
- KRUGLYAK, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**: 139–144.
- LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**: 49–67.
- LEWONTIN, R. C., 1988 On measures of gametic disequilibrium. *Genetics* **120**: 849–852.
- LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.
- MADDOX, J. F., K. P. DAVIES, A. M. CRAWFORD, D. J. HULME, D. VAIMAN *et al.*, 2001 An enhanced linkage map of the sheep genome comprising more than 1000 loci. *Genome Res.* **11**: 1275–1289.
- MILAN, D., J. T. JEON, C. LOOFT, V. ARMARGER, A. ROBIC *et al.*, 2000 A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* **288**: 1248–1251.
- MONTGOMERY, G. W., A. M. CRAWFORD, J. M. PENY, K. G. DODDS, A. E. EDE *et al.*, 1993 The ovine Booroola fecundity gene (FecB) is linked to markers from a region of human chromosome 4q. *Nat. Genet.* **4**: 410–414.
- RAADSMAN, H. W., J. C. McEWAN, M. J. STEAR and A. M. CRAWFORD, 1999 Genetic characterisation of protective vaccine responses in sheep using multi-valent *Dichelobacter nodosus* vaccines. *Vet. Immunol. Immunopathol.* **72**: 219–229.
- REICH, D. E., M. CARGILL, S. BOLK, J. IRELAND, P. C. SABETI *et al.*, 2001 Linkage disequilibrium in the human genome. *Nature* **411**: 199–204.
- SACHIDANANDAM, R., D. WEISSMAN, S. C. SCHMIDT, J. M. KAKOL, L. D. STEIN *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928–933.
- SCHNEIDER, S., D. ROESSLI and L. EXCOFFIER, 2000 *Arlequin Ver. 2.000: A Software for Genetic Data Analysis*. Genetics and Biometry Laboratory, University of Geneva, Geneva.
- SLATKIN, M., 1994 Linkage disequilibrium in growing and stable populations. *Genetics* **137**: 331–336.
- TAILLON-MILLER, P., I. BAUER-SARDINA, N. L. SACCONI, J. PUTZEL, T. LAITINEN *et al.*, 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.* **25**: 324–328.
- TERWILLIGER, J. D., and K. M. WEISS, 1998 Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* **9**: 578–594.
- WILSON, J. F., and D. B. GOLDSTEIN, 2000 Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. *Am. J. Hum. Genet.* **67**: 926–935.
- WILSON, T., X.-Y. WU, J. L. JUENGEL, I. K. ROSS, J. M. LUMSDEN *et al.*, 2001 Highly prolific Booroola sheep have a mutation in the intracellular kinase domain of bone morphogenetic protein IB receptor (ALK-6) that is expressed in both oocytes and granulosa cells. *Biol. Reprod.* **64**: 1225–1235.
- ZAPATA, C., 2000 The *D'* measure of overall gametic disequilibrium between pairs of multiallelic loci. *Evolution* **54**: 1809–1812.

Communicating editor: C. HALEY