Maximum Likelihood Estimation of Linkage Disequilibrium in Half-Sib Families

L. Gomez-Raya[1]
Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología
Agraria y Alimentaria (INIA), 28040 Madrid, Spain

ABSTRACT Maximum likelihood methods for the estimation of linkage disequilibrium between biallelic DNA-markers in half-sib families (half-sib method) are developed for single and multifamily situations. Monte Carlo computer simulations were carried out for a variety of scenarios regarding sire genotypes, linkage disequilibrium, recombination fraction, family size, and number of families. A double heterozygote sire was simulated with recombination fraction of 0.00, linkage disequilibrium among dams of δ= 0.10, and alleles at both markers segregating at intermediate frequencies for a family size of 500. The average estimates of δ were 0.17, 0.25, and 0.10 for Excoffier and Slatkin (1995), maternal informative haplotypes, and the half-sib method, respectively. A multifamily EM algorithm was tested at intermediate frequencies by computer simulation. The range of the absolute difference between estimated and simulated δ was between 0.000 and 0.008. A cattle half-sib family was genotyped with the Illumina 50K BeadChip. There were 314,730 SNP pairs for which the sire was a homo-heterozygote with average estimates of r^2 of 0.115, 0.067, and 0.111 for half-sib, Excoffier and Slatkin (1995), and maternal informative haplotypes methods, respectively. There were 208,872 SNP pairs for which the sire was double heterozygote with average estimates of r^2 across the genome of 0.100, 0.267, and 0.925 for half-sib, Excoffier and Slatkin (1995), and maternal informative haplotypes methods, respectively. Genome analyses for all possible sire genotypes with 829,042 tests showed that ignoring half-sib family structure leads to upward biased estimates of linkage disequilibrium. Published inferences on population structure and evolution of cattle should be revisited after accommodating existing half-sib family structure in the estimation of linkage disequilibrium.

TRADITIONAL methods for gene mapping are based on linkage, which requires a family structure because loci are mapped by tracing inheritance of marker alleles in progeny from at least one ancestor. The DNA markers of choice were microsatellites because they were abundant and very informative. Linkage maps of microsatellites were developed for farm animal species with a half-sib structure such as cattle (Da and Lewin, 1995; Ma *et al.* 1996; Kappes *et al.* 1997; Barendse *et al.* 1997; Våge *et al.* 2000). In recent years, a revolution has been initiated in human genetics with the large-scale DNA sequencing of the HAP MAP project (2007), which allowed the discovery of vast amounts of single nucleotide polymorphism (SNP). SNP sequences were used in arrays allowing interrogation of the human genome from thousands to over a million SNPs. The biggest interest in humans is the application of this technology for identification of variants that are associated with genetic diseases in the so-called case-control studies.

The development of SNP arrays in human genetics was followed by animal geneticists. There are commercially available arrays for over 50 or 60 thousands SNPs for the cow, sheep, and swine. The statistical treatment of SNP arrays in animal populations is carried out without consideration for the breeding structure currently present in farm animals but lacking in experiments for case-control studies in human populations. Contrary to human populations, animals at the farm are highly related because of the use of artificial insemination (AI) and intensive breeding. For example, dairy bulls with high estimated breeding values might have over a million daughters (http://www.crv4all.com/eng/halloffame/Sunny_Boy_HOFame.pdf). Analysis of linkage disequilibrium (LD) between pairs of SNPs in cattle populations has been carried out either using the expectation maximization (EM) algorithm for unrelated individuals

doi: 10.1534/genetics.111.137521
Manuscript received December 7, 2011; accepted for publication February 19, 2012

[1] Address for correspondence: Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Ctra. de La Coruña km 7, 28040 Madrid, Spain. E-mail: gomez.luis@inia.es

(*e.g.*, Sargolzaei *et al.* 2008) or using the most likely of phased haplotypes (*e.g.*, McKay *et al.* 2007; Qanbari *et al.* 2010). The first method ignores that the contribution of haplotypes from sires to progeny exceeds its true counts in the population because each offspring receives one haplotype from their sire. The second method ignores that marker informativity might cause a systematic increase or decrease of informative sire haplotype counts (and consequently informative haplotype counts from dams) depending on the genetic distance between markers. Consequently, bias in the estimation of LD using half-sib data might occur.

The objective of this article is to develop maximum likelihood methods for the estimation of linkage disequilibrium between codominant DNA markers in half-sib families. It is shown that severe biased estimation may occur after ignoring half-sib relationships. The methods are tested via Monte Carlo computer simulation. Comparison of alternative methods of estimation of linkage disequilibrium is carried out after genotyping a half-sib family with 36 calves with the Illumina 50K BeadChip.

Theory and Methods

AI is in widespread use in cattle with the most common situation being a sire having a single progeny from a number of dams. Three situations are possible when estimating second-order linkage disequilibrium (disequilibrium considering two loci) between two DNA markers: (a) the sire is a homozygote at the two loci, (b) the sire is a homozygote at one locus and a heterozygote at the other, and (c) the sire is a heterozygote at the two loci. For the following derivations, assumptions are: (1) recombination fraction is known without error, and (2) the linkage phase (combination of alleles at two loci on the two homologous chromosomes in diploid individuals) in the sire is known. The impact of departures from these assumptions is addressed in the *Discussion*.

Double homozygote sire

Let the sire have genotype *TTMM* at two SNPs, *T/t*, and *M/m*. Offspring might have genotypes *TTMM*, *TTMm*, *TtMM*, *TtMm* indicating that haplotypes *TM*, *Tm*, *tM*, and *tm* were inherited from dams, respectively. Therefore, the haplotypes in half-sibs are fully informative and linkage disequilibrium can be estimated directly from haplotype counts. Thus, for alleles *T* and *M* at two loci, the disequilibrium can be estimated by substituting haplotype and allelic frequencies into $D_{TM}=f_{TM}-f_Tf_M$; $D_{Tm}=f_{Tm}-f_Tf_m$; $D_{tM}=f_{tM}-f_tf_M$; and $D_{tm}=f_{tm}-f_tf_m$, where f_k is the frequency of the *k*th allele, D_{kt} and f_{kt} are the linkage disequilibrium and haplotype frequencies between the *k*th and *t*th alleles at the two loci, respectively. In addition to allele frequencies, only one parameter for the linkage disequilibrium, δ, needs to be estimated since $D_{TM}=\delta$; $D_{Tm}=-\delta$; $D_{tM}=-\delta$; and $D_{tm}=\delta$. Estimating disequilibrium by direct counts of haplotype and allele frequencies is also the maximum likelihood estimate of linkage disequilibrium. The sampling variance of the estimates of the disequilibrium parameter for the *i*th family is derived in *Appendix A*,

$$Var(\hat{\delta}) \approx \frac{1}{\left[-\left(\partial^2 \ln L_i(\delta|nG)/\partial\delta^2\right)\right]_{\delta=\hat{\delta}}},$$

where $L_i(\delta|nG)$ is the maximum likelihood function of the disequilibrium parameter, δ, conditional to the haplotype counts, *nG*.

The value of the second derivative with respect to the disequilibrium parameter is

$$\frac{\partial^2 \ln L_i(\hat{\delta}|nG)}{\partial\delta^2}=-\frac{n_{TM,i}}{(\delta+f_Tf_M)^2}-\frac{n_{Tm,i}}{(-\delta+f_Tf_m)^2}-\frac{n_{tM,i}}{(-\delta+f_tf_M)^2}-\frac{n_{tm,i}}{(\delta+f_tf_m)^2}.$$

Sire is homozygote at one locus and heterozygote at the other

A full and a reduced model are developed in this section. A full model estimates all unknowns (linkage disequilibrium, and allele frequencies for the marker for which the sire is heterozygote) simultaneously. The reduced model estimate only linkage disequilibrium assuming that allele frequencies are known without error (or estimated in a previous step).

Full model for estimating LD in a homo-heterozygote sire: Let the sire have genotype *TTMm* at two SNPs, *T/t*, and *M/m*. The likelihood equation for the *i*th family is

$$\begin{aligned} L_i(\hat{\delta},\hat{f}_M|nG) = {} & K(\phi_{TTMM})^{n_{TTMM,i}}(\phi_{TTMm})^{n_{TTMm,i}}(\phi_{TTmm})^{n_{TTmm,i}} \\ & \times(\phi_{TtMM})^{n_{TtMM,i}}(\phi_{TtMm})^{n_{TtMm,i}}(\phi_{Ttmm})^{n_{Ttmm,i}}, \end{aligned} \qquad (1)$$

where $n_{j,i}$ are the genotype counts (*nG*) from offspring from the *i*th sire family (*j* = *TTMM*, *TTMm*, *TTmm*, *TtMM*, *TtMm*, and *Ttmm*), and ϕ_j is the probability of the *j*th genotype among progeny. These probabilities can be obtained after adding the corresponding frequencies for all possible matings (Table 1): $\phi_{TTMM}=\frac{1}{2}f_{TM}$; $\phi_{TTMm}=\frac{1}{2}f_{Tm}+\frac{1}{2}f_{TM}$; $\phi_{TTmm}=\frac{1}{2}f_{Tm}$; $\phi_{TtMM}=\frac{1}{2}f_{tM}$; $\phi_{TtMm}=\frac{1}{2}f_{tM}+\frac{1}{2}f_{tm}$; $\phi_{Ttmm}=\frac{1}{2}f_{tm}$. Equation 1 can be solved by the EM algorithm after making haplotype frequencies equal to their expected values,

$$f^i_{TM}=\frac{1}{N_i}\left(n_{TTMM,i}+\frac{\hat{f}^i_{TM}}{\hat{f}_T}n_{TTMm,i}\right)$$

$$\hat{f}^i_{Tm}=\frac{1}{N_i}\left(\left(1-\frac{\hat{f}^i_{TM}}{\hat{f}_T}\right)n_{TTMm,i}+n_{TTmm,i}\right)$$

$$\hat{f}^i_{tM}=\frac{1}{N_i}\left(n_{TtMM,i}+\left(\frac{\hat{f}^i_{tM}}{1-\hat{f}_T}\right)n_{TtMm,i}\right)$$

Table 1 Genotypes in the half-sib offspring from all possible gamete combinations produced from a heterozygote sire at one SNP, *M/m*, and homozygote at the other SNP, *T/t*.

		Sire(*TM/Tm*)	
Dam		*TM*	*Tm*
G	freq	1/2	1/2
TM	f_{TM}	TTMM	TTMm
		$\frac{1}{2}f_{TM}$	$\frac{1}{2}f_{TM}$
Tm	f_{Tm}	TTMm	TTmm
		$\frac{1}{2}f_{Tm}$	$\frac{1}{2}f_{Tm}$
tM	f_{tM}	TtMM	TtMm
		$\frac{1}{2}f_{tM}$	$\frac{1}{2}f_{tM}$
Tm	f_{tm}	TtMm	Ttmm
		$\frac{1}{2}f_{tm}$	$\frac{1}{2}f_{tm}$

G, gametes; freq, frequency.

$$\hat{f}^{i}_{tm} = \frac{1}{N_i}\left(\left(1 - \frac{\hat{f}^{i}_{tM}}{1 - \hat{f}_T}\right) n_{TtMm,i} + n_{Ttmm,i}\right), \tag{2}$$

where N_i is the size of the ith half-sib family. Equations 2 can be solved iteratively after giving a starting value to the haplotype frequencies and by estimating in each iteration $\hat{f}_T = \hat{f}^{i}_{Tm} + \hat{f}^{i}_{TM}$. The starting values used in this study were the product of allele frequencies, so disequilibrium was null ($\delta = 0$).

Reduced model for estimating LD in a homo-heterozygote sire family: In a reduced model, allele frequencies are not estimated simultaneously with haplotype frequencies but are assumed to be known. The estimate of linkage disequilibrium is

$$\hat{\delta} = \hat{f}^{i}_{TM} - \hat{f}_T \hat{f}_M,$$

where $\hat{f}_T = (1/N_i)(n_{TTMM,i} + n_{TTMm,i} + n_{TTmm,i})$,

$$\hat{f}_M = \frac{n_{TTMM,i} + n_{TtMM,i}}{n_{TTMM,i} + n_{TtMM,i} + n_{TTmm,i} + n_{Ttmm,i}} \quad \text{and}$$

$$\hat{f}^{i}_{TM} = \frac{\hat{f}_T \ n_{TTMM,i}}{N_i \hat{f}_T - n_{TTMm,i}}.$$

The derivation is given in *Appendix B*.

The disequilibrium estimated in the reduced model gives slightly different estimates than the disequilibrium estimated using a full model but has the advantage of faster computation when a large number of SNPs are tested. The approximated sampling variance of the estimates of the disequilibrium parameter for the ith family is

$$Var(\hat{\delta}) \approx \frac{1}{\left[-\left(\partial^2 \ln L_i(\delta|nG)/d\delta^2\right)\right]_{\delta=\hat{\delta}}},$$

where

$$\frac{\partial^2 \ln L_i(\hat{\delta}|nG)}{\partial \delta^2} = -\frac{n_{TTMM,i}}{(\delta + \hat{f}_T \hat{f}_M)^2} - \frac{n_{TTmm,i}}{(-\delta + \hat{f}_T \hat{f}_m)^2} - \frac{n_{TtMM,i}}{(-\delta + \hat{f}_t \hat{f}_M)^2} - \frac{n_{Ttmm,i}}{(\delta + \hat{f}_t \hat{f}_m)^2}$$

as derived in *Appendix A*.

Sire is heterozygote at two SNPs

Equations for a full and a reduced model follow. A full model estimates allele and haplotype frequencies simultaneously whereas a reduced model works first estimating allele frequencies and then haplotype frequencies. The full model has better statistical properties but the reduced model has faster computation and, therefore, is practical for large-scale testing of disequilibria among SNPs.

Full model for estimating LD in a double-heterozygote sire family: Let the sire have genotype *TtMm* at two SNPs, *T/t*, and *M/m* and linkage phase (*TM/tm*). As before, $n_{j,i}$ are the genotype counts from offspring from the ith sire family (j = *TTMM*, *TTMm*, *TTmm*, *TtMM*, *TtMm*, *Ttmm*, *ttMM*, *ttMm*, and *ttmm*). The recombination fraction is c, which is assumed to be known without error. The likelihood equation for data of the ith half-sib family is

$$\begin{aligned} L_i(\delta, f_T, f_M|nG) = K&(\phi_{TTMM})^{n_{TTMM,i}}(\phi_{TTMm})^{n_{TTMm,i}}(\phi_{TTmm})^{n_{TTmm,i}} \\ &\times (\phi_{TtMM})^{n_{TtMM,i}}(\phi_{TtMm})^{n_{TtMm,i}}(\phi_{Ttmm})^{n_{Ttmm,i}} \\ &\times (\phi_{ttMM})^{n_{ttMM,i}}(\phi_{ttMm})^{n_{ttMm,i}}(\phi_{ttmm})^{n_{ttmm,i}}, \end{aligned} \tag{3}$$

where the probabilities of offspring genotypes among half-sib offspring are obtained from Table 2: $\phi_{TTMM} = \frac{1}{2}(1-c)f_{TM}$; $\phi_{TTMm} = \frac{1}{2}(1-c)f_{Tm} + \frac{1}{2}c\ f_{TM}$; $\phi_{TTmm} = \frac{1}{2}c\ f_{Tm}$; $\phi_{TtMM} = \frac{1}{2}(1-c)f_{tM} + \frac{1}{2}c\ f_{TM}$; $\phi_{TtMm} = \frac{1}{2}(1-c)(f_{tm} + f_{TM}) + \frac{1}{2}c\ (f_{tM} + f_{Tm})$; $\phi_{Ttmm} = \frac{1}{2}(1-c)f_{Tm} + \frac{1}{2}c\ f_{tm}$; $\phi_{ttMM} = \frac{1}{2}c\ f_{tM}$; $\phi_{ttMm} = \frac{1}{2}(1-c)f_{tM} + \frac{1}{2}c\ f_{tm}$; and $\phi_{ttmm} = \frac{1}{2}(1-c)f_{tm}$.

Likelihood Equation 3 can be solved by applying the EM algorithm,

$$\hat{f}^{\,i}_{TM} = \frac{1}{N_i}\left(n_{TTMM,i} + \frac{c\hat{f}^{i}_{TM} n_{TTMm,i}}{c\hat{f}^{i}_{TM} + (1-c)\hat{f}^{i}_{Tm}} + \frac{c\hat{f}^{i}_{TM} n_{TtMM,i}}{c\hat{f}^{i}_{TM} + (1-c)\hat{f}^{i}_{tM}} + \frac{(1-c)\hat{f}^{i}_{TM} n_{TtMm,i}}{\left[c\left(\hat{f}^{i}_{\mathrm{TM}} + \hat{f}^{i}_{tM}\right) + (1-c)\left(\hat{f}^{i}_{TM} + \hat{f}^{i}_{tm}\right)\right]}\right)$$

$$\hat{f}^{i}_{Tm} = \frac{1}{N_i}\left(n_{TTmm,i} + \frac{(1-c)\hat{f}^{i}_{Tm} n_{TTMm,i}}{c\hat{f}^{i}_{TM} + (1-c)\hat{f}^{i}_{Tm}} + \frac{(1-c)\hat{f}^{i}_{Tm} n_{Ttmm,i}}{c\hat{f}^{i}_{tm} + (1-c)\hat{f}^{i}_{Tm}} + \frac{c\hat{f}^{i}_{Tm} n_{TtMm,i}}{\left[c\left(\hat{f}^{i}_{Tm} + \hat{f}^{i}_{tM}\right) + (1-c)\left(\hat{f}^{i}_{TM} + \hat{f}^{i}_{tm}\right)\right]}\right)$$

Table 2 Genotypes and their frequencies among half-sib progeny from a double heterozygote sire

		Sire (phase TM/tm)			
Dam		TM	Tm	tM	tm
G	freq	$\frac{1}{2}$ *(1 − c)*	$\frac{1}{2}$*c*	$\frac{1}{2}$*c*	$\frac{1}{2}$ *(1-c)*
TM	f$_{TM}$	TTMM	TTMm	TtMM	TtMm
		$\frac{1}{2}$ (1 − c) f$_{TM}$	$\frac{1}{2}$c f$_{TM}$	$\frac{1}{2}$c f$_{TM}$	$\frac{1}{2}$ (1 − c) f$_{TM}$
Tm	f$_{Tm}$	TTMm	TTmm	TtMm	Ttmm
		$\frac{1}{2}$ (1 − c) f$_{Tm}$	$\frac{1}{2}$c f$_{Tm}$	$\frac{1}{2}$c f$_{Tm}$	$\frac{1}{2}$ (1 − c) f$_{Tm}$
tM	f$_{tM}$	TtMM	TtMm	ttMM	ttMm
		$\frac{1}{2}$ (1 − c) f$_{tM}$	$\frac{1}{2}$c f$_{tM}$	$\frac{1}{2}$c f$_{tM}$	$\frac{1}{2}$ (1 − c) f$_{tM}$
tm	f$_{tm}$	TtMm	Ttmm	ttMm	ttmm
		$\frac{1}{2}$ (1 − c) f$_{tm}$	$\frac{1}{2}$c f$_{tm}$	$\frac{1}{2}$c f$_{tm}$	$\frac{1}{2}$ (1 − c) f$_{tm}$
		Sire (phase Tm/tM)			
		TM	Tm	tM	tm
G	freq	$\frac{1}{2}$c	$\frac{1}{2}$ (1-c)	$\frac{1}{2}$ (1-c)	$\frac{1}{2}$c
TM	f$_{TM}$	TTMM	TTMm	TtMM	TtMm
		$\frac{1}{2}$c f$_{TM}$	$\frac{1}{2}$ (1-c)f$_{TM}$	$\frac{1}{2}$ (1 − c)f$_{TM}$	$\frac{1}{2}$c f$_{TM}$
Tm	f$_{Tm}$	TTMm	TTmm	TtMm	Ttmm
		$\frac{1}{2}$f$_{Tm}$	$\frac{1}{2}$ (1 − c)f$_{Tm}$	$\frac{1}{2}$ (1 − c)f$_{Tm}$	$\frac{1}{2}$c f$_{Tm}$
tM	f$_{tM}$	TtMM	TtMm	ttMM	ttMm
		$\frac{1}{2}$c f$_{tM}$	$\frac{1}{2}$ (1 − c)f$_{tM}$	$\frac{1}{2}$ (1 − c)f$_{tM}$	$\frac{1}{2}$c f$_{tM}$
tm	f$_{tm}$	TtMm	Ttmm	ttMm	ttmm
		$\frac{1}{2}$c f$_{tm}$	$\frac{1}{2}$ (1 − c)f$_{tm}$	$\frac{1}{2}$ (1 − c)f$_{tm}$	$\frac{1}{2}$c f$_{tm}$

G, gametes; freq, frequency.

$$\hat{f}^i_{tM} = \frac{1}{N_i}\left(n_{ttMM,i} + \frac{(1-c)\hat{f}^i_{tM} n_{TtMM,i}}{c\hat{f}^i_{TM} + (1-c)\hat{f}^i_{tM}} + \frac{(1-c)\hat{f}^i_{tM} n_{ttMm,i}}{c\hat{f}^i_{tm} + (1-c)\hat{f}^i_{tM}} + \frac{c\hat{f}^i_{tM} n_{TtMm,i}}{\left[c\left(\hat{f}^i_{Tm} + \hat{f}^i_{tM}\right) + (1-c)\left(\hat{f}^i_{TM} + c\hat{f}^i_{tm}\right)\right]}\right)$$

$$\hat{f}^i_{tm} = \frac{1}{N_i}\left(n_{ttmm,i} + \frac{c\hat{f}^i_{tm} n_{Ttmm,i}}{c\hat{f}^i_{tm} + (1-c)\hat{f}^i_{Tm}} + \frac{c\hat{f}^i_{tm} n_{ttMm,i}}{\hat{f}^i_{tm} + (1-c)\hat{f}_{tM}} + \frac{(1-c)\hat{f}^i_{tm} n_{TtMm,i}}{\left[c\left(\hat{f}^i_{Tm} + \hat{f}^i_{tM}\right) + (1-c)\left(\hat{f}^i_{TM} + \hat{f}^i_{tm}\right)\right]}\right), \quad (4)$$

where, as before, N_i is the size of the ith half-sib family. Using initial values of the haplotype frequencies and iterating over Equation 4 will converge to ML estimates of haplotype frequencies. Linkage disequilibrium is estimated by $\hat{\delta} = \hat{f}^i_{TM}\hat{f}^i_{tm} - \hat{f}^i_{Tm}\hat{f}^i_{tM}$.

If the linkage phase of the sire is *Tm/tM* then the EM equations are

$$\hat{f}^i_{TM} = \frac{1}{N_i}\left(n_{TTMM,i} + \frac{(1-c)\hat{f}^i_{TM} n_{TTMm,i}}{(1-c)\hat{f}^i_{TM} + c\hat{f}^i_{Tm}} + \frac{(1-c)\hat{f}^i_{TM} n_{TtMM,i}}{(1-c)\hat{f}^i_{TM} + c\hat{f}^i_{tM}} + \frac{c\hat{f}^i_{TM} n_{TtMm,i}}{\left[(1-c)\left(\hat{f}^i_{Tm} + \hat{f}^i_{tM}\right) + c\left(\hat{f}^i_{TM} + \hat{f}^i_{tm}\right)\right]}\right)$$

$$\hat{f}^i_{Tm} = \frac{1}{N_i}\left(n_{TTmm,i} + \frac{c\hat{f}^i_{Tm} n_{TTMm,i}}{(1-c)\hat{f}^i_{TM} + c\hat{f}^i_{Tm}} + \frac{c\hat{f}^i_{Tm} n_{Ttmm,i}}{(1-c)\hat{f}^i_{tm} + c\hat{f}^i_{Tm}} + \frac{(1-c)\hat{f}^i_{Tm} n_{TtMm,i}}{\left[(1-c)\left(\hat{f}^i_{Tm} + \hat{f}^i_{tM}\right) + c\left(\hat{f}^i_{TM} + \hat{f}^i_{tm}\right)\right]}\right)$$

$$\hat{f}^i_{tM} = \frac{1}{N_i}\left(n_{ttMM,i} + \frac{c\hat{f}^i_{tM} n_{TtMM,i}}{(1-c)\hat{f}^i_{TM} + c\hat{f}^i_{tM}} + \frac{c\hat{f}^i_{tM} n_{ttMm,i}}{(1-c)\hat{f}^i_{tm} + c\hat{f}^i_{tM}} + \frac{(1-c)\hat{f}^i_{tM} n_{TtMm,i}}{\left[(1-c)\left(\hat{f}^i_{Tm} + \hat{f}^i_{tM}\right) + c\left(\hat{f}^i_{TM} + \hat{f}^i_{tm}\right)\right]}\right)$$

$$\hat{f}^i_{tm} = \frac{1}{N_i}\left(n_{ttmm,i} + \frac{(1-c)\hat{f}^i_{tm} n_{Ttmm,i}}{(1-c)\hat{f}^i_{tm} + c\hat{f}^i_{Tm}} + \frac{(1-c)\hat{f}^i_{tm} n_{ttMm,i}}{(1-c)\hat{f}^i_{tm} + c\hat{f}^i_{tM}} + \frac{c\hat{f}^i_{tm} n_{TtMm,i}}{\left[(1-c)\left(\hat{f}^i_{Tm} + \hat{f}^i_{tM}\right) + c\left(\hat{f}^i_{TM} + \hat{f}^i_{tm}\right)\right]}\right).$$

However, the same results can be obtained by making the following substitutions in Equation 4: $n_{TTmm,i}$ by $n_{TTMM,i}$; $n_{TTMM,i}$ by $n_{TTmm,i}$; $n_{Ttmm,i}$ by $n_{TtMM,i}$; $n_{TtMM,i}$ by $n_{Ttmm,i}$; $n_{ttmm,i}$ by $n_{ttMM,i}$; and $n_{ttMM,i}$ by $n_{ttmm,i}$. Linkage phase can be estimated simultaneously to recombination fraction (Gomez-Raya 2001).

Reduced model for estimating LD in a double-heterozygote sire: A reduced model can be used after assuming that allele frequencies at the two DNA markers are known without error. It makes easier and faster estimation of linkage disequilibrium and its sampling variance. It can be solved by making use of the EM algorithm as described in Equation 4 but using as input parameters estimates of allele frequencies of M and T (as given by Gomez-Raya 2001):

$$\hat{f}_M = \left(\frac{n_{TTMM,i} + n_{TtMM,i} + n_{ttMM,i}}{n_{TTMM,i} + n_{TtMM,i} + n_{ttMM,i} + n_{TTmm,i} + n_{Ttmm,i} + n_{ttmm,i}}\right)$$

$$\hat{f}_T = \left(\frac{n_{TTMM,i} + n_{TTMm,i} + n_{TTmm,i}}{n_{TTMM,i} + n_{TtMM,i} + n_{ttMM,i} + n_{TTmm,i} + n_{Ttmm,i} + n_{ttmm,i}}\right).$$

A solution when $c = 0$ for the reduced model is a positive root between 0 and 1 of the quadratic: $a\ (\hat{f}^{i}_{TM})^2 + b\hat{f}^{i}_{TM} + z = 0$, where $a = 2N_i$, $b = N_i(1-\hat{f}_M-\hat{f}_T)-2n_{TTMM,i}\ -n_{TtMm,i}$, and $z = -(1-\hat{f}_M-\hat{f}_T)n_{TTMM,i}$. Derivation of the method and an explicit solution for fully linked markers is given in *Appendix B*.

As shown in *Appendix A*, the reduced model provides a simpler approximated sampling variance of the estimates of the disequilibrium parameter for the *i*th family by

$$\mathrm{Var}(\hat{\delta}) \approx \frac{1}{\left[-\left(\partial^2 \ln L_i(\delta|nG)/\partial\delta^2\right)\right]_{\delta=\hat{\delta}}}$$

$$\begin{aligned}\frac{\partial^2 \ln L_i(\delta|nG)}{\partial\delta^2} = &-\frac{n_{TTMM,i}}{[\delta + f_T f_M]^2} - \frac{(1-2c)^2 n_{TTMm,i}}{[(1-c)(-\delta + f_T f_m) + c(\delta + f_T f_M)]^2}\\ &-\frac{n_{TTmm,i}}{[-\delta + f_T f_m]^2} - \frac{(1-2c)^2 n_{TtMM,i}}{[(1-c)(-\delta + f_t f_M) + c(\delta + f_T f_M)]^2}\\ &-\frac{4(1-2c)^2 n_{TtMm,i}}{[(1-c)(2\delta + f_T f_M + f_t f_m) + c(-2\delta + f_T f_m + f_t f_M)]^2}\\ &-\frac{(1-2c)^2 n_{Ttmm,i}}{[(1-c)(-\delta + f_T f_m) + c(\delta + f_t f_m)]^2} - \frac{n_{ttMM,i}}{[-\delta + f_t f_M]^2}\\ &-\frac{(1-2c)^2 n_{ttMm,i}}{[(1-c)(-\delta + f_t f_M) + c(\delta + f_t f_m)]^2} - \frac{n_{ttmm,i}}{[(\delta + f_t f_m)]^2}.\end{aligned}$$

This equation can be used as an approximation to the full model with linkage disequilibrium and allele frequencies estimated from that model.

Estimation of LD Across multiple half-sib families

In most instances, genotype information is available for multiple half-sib families (*e.g.*, data from a granddaughter design project). The likelihood equation to estimate LD across half-sib families is

$$L(\delta, f_T, f_M \mid nG) = \prod_{i=1}^{nf} L_i(\delta, f_T, f_M \mid nG),$$

where *L(δ, f_T, f_M|nG)* is the likelihood for the *i*th half-sib family conditional to genotype marker information (*nG*) and *nf* is the number of families. Note that depending on the sire genotype, allele frequencies for *T* and *M* (double homozygote) or *M* (homo-heterozygote) do not need to be estimated. The EM algorithm can be applied to multiple families by iterating on the four haplotype frequencies:

$$\hat{f}_{TM} = \frac{\sum_{i=1}^{nf}\left(N_i\hat{f}^{i}_{TM}\right)}{\sum_{i=1}^{nf} N_i},$$

$$\hat{f}_{Tm} = \frac{\sum_{i=1}^{nf}\left(N_i\hat{f}^{i}_{TM}\right)}{\sum_{i=1}^{nf} N_i},$$

$$\hat{f}_{tM} = \frac{\sum_{i=1}^{nf}\left(N_i\hat{f}^{i}_{tM}\right)}{\sum_{i=1}^{nf} N_i},$$

$$\hat{f}_{tm} = \frac{\sum_{i=1}^{nf}\left(N_i\hat{f}^{i}_{tm}\right)}{\sum_{i=1}^{nf} N_i}, \tag{5}$$

where equations for haplotype frequencies for each single family varies depending on the sire genotype. For example,

$$f^{i}_{TM} = \frac{1}{N_i}(n_{TTMM,i}),$$

$$f^{i}_{TM} = \frac{1}{N_i}\left(n_{TTMM,i} + \frac{\hat{f}^{i}_{TM}}{\hat{f}_T} n_{TTMm,i}\right),$$

$$\begin{aligned}\hat{f}^{\,i}_{\ TM} = \frac{1}{N_i}\Bigg(n_{TTMM,i} &+ \frac{c\hat{f}^{i}_{TM} n_{TTMm,i}}{c\hat{f}^{i}_{TM} + (1-c)\hat{f}^{i}_{Tm}}\\ &+ \frac{c\hat{f}^{i}_{TM} n_{TtMM,i}}{c\hat{f}^{i}_{TM} + (1-c)\hat{f}^{i}_{tM}}\\ &+ \frac{(1-c)\hat{f}^{i}_{TM} n_{TtMm,i}}{\left[c\left(\hat{f}^{i}_{Tm} + \hat{f}^{i}_{tM}\right) + (1-c)\left(\hat{f}^{i}_{TM} + \hat{f}^{i}_{tm}\right)\right]}\Bigg)\end{aligned}$$

are the equations for haplotype *TM* if the sire is double homozygote, homo-heterozygote, or double heterozygote, respectively. The frequencies for the other haplotypes are as found in Equations 2 and 4 for homo-heterozygote and double heterozygote sires, respectively. Equation 5 can be solved iteratively after giving a starting value to the haplotype frequencies and by estimating in each iteration $\hat{f}_T = \hat{f}^{i}_{Tm} + \hat{f}^{i}_{TM}$ and $\hat{f}_M = \hat{f}^{i}_{TM} + \hat{f}^{i}_{tM}$.

The estimation of the sampling variance for linkage disequilibrium in multiple half-sib families can be carried out by:

$$Var(\hat{\delta}) \approx \frac{1}{\left[-\left(\partial^2 \ln \prod_{i=1}^{nf} L_i(\delta, f_T, f_M|nG)/\partial\delta^2\right)\right]_{\delta=\hat{\delta}}},$$

where second derivatives of the natural logarithm of likelihood varies depending on sire genotype (double homozygote, homo-heterozygote, and double heterozygote) as described in *Appendix A*.

Hypothesis testing of LD in multiple half-sib families

Testing if linkage disequilibrium is different from 0 can be carried out by a likelihood-ratio test. For the *i*th half-sib family the likelihood-ratio test is

$$\mathrm{LRT}_i = -2\ln\frac{L_i\,\mathrm{Null}(\hat{\delta}=0|nG)}{L_i(\delta=\hat{\delta}|nG)},$$

where $L_i\mathrm{Null}(\hat{\delta}=0|nG)$ and $L_i(\delta=\hat{\delta}|nG)$ are the likelihoods for the *i*th family under the null hypothesis ($\delta = 0$) and under the alternative hypothesis with $\delta=\hat{\delta}$.

A likelihood-ratio test across families is

$$\mathrm{LRT}_{\mathrm{joint}} = -2\sum_{i=1}^{nf}\ln\frac{L_i\mathrm{Null}(\hat{\delta}=0|nG)}{L_i(\delta=\hat{\delta}|nG)},$$

which is distributed as a χ^2 with 1 d.f. Here δ is estimated across all families by the EM algorithm (Equation 5).

Bias in estimating LD in half-sibs after ignoring the family structure

In this section, approximate bias for estimating LD in half-sib families using the method of Excoffier and Slatkin (1995) for unrelated individuals and maternal informative haplotypes is derived algebraically. Only sires that are homo-heterozygotes and double heterozygotes might produce progeny in which haplotypes cannot be fully inferred from the genotypes.

Sire homo-heterozygote: Method of Excoffier and Slatkin (1995) for unrelated individuals: Assuming genotype TTMm in the sire, the expected frequency of haplotype TM among half-sib progeny can be approximated by

$$E[\hat{f}_{TM}] \approx \frac{\frac{1}{2}N_i + N_i f_{TM}}{2N_i} = \frac{1}{4}+\frac{1}{2}f_{TM},$$

where $\frac{1}{2}N_i$ comes from the contribution of the *TM* haplotype from the sire and $N_i f_{TM}$ from the contributions of the dams. The total number of haplotypes in the offspring is $2N_i$. The approximated expected frequencies of alleles *T* and *M* are computed following the same rules:

$$E[\hat{f}_T] \approx \frac{N_i + N_i f_T}{2N_i} = \frac{1}{2}+\frac{1}{2}f_T$$

$$E[\hat{f}_M] \approx \frac{\frac{1}{2}N_i + N_i f_M}{2N_i} = \frac{1}{4}+\frac{1}{2}f_M.$$

The expected estimate of the disequilibrium after using the method of Excofier and Slatkin (1995) is

$$\begin{aligned} E[\hat{\delta}] &\approx E[\hat{D}_{TM}] \\ &\approx E[\hat{f}_{TM}] - E[\hat{f}_T]E[\hat{f}_M] \\ &\approx \frac{1}{2}D_{TM} + \frac{1}{8} + \frac{1}{4}f_T f_M - \frac{1}{4}f_M - \frac{1}{8}f_T. \end{aligned}$$

Consequently, the bias after using this method is approximated by

$$\begin{aligned} \mathrm{Bias} &\approx D_{TM} - E[\hat{D}_{TM}] \\ &= \frac{1}{2}D_{TM} - \frac{1}{8} - \frac{1}{4}f_T f_M + \frac{1}{4}f_M + \frac{1}{8}f_T. \end{aligned}$$

Sire homo-heterozygote: Estimation of LD using informative maternal haplotypes in half-sib families: Half-sib progeny from heterozygote sires might not be informative. For example, haplotype *TM* inherited from dams will be informative only in progeny with genotypes *TTMM*.

Therefore, the expected frequency of haplotype *TM* among progeny will be estimated by

$$E[\hat{f}_{TM}] \approx \frac{\frac{1}{2}f_{TM}}{\frac{1}{2}[f_{TM}+f_{Tm}+f_{tM}+f_{tm}]} = f_{TM}.$$

The estimation of haplotype frequencies and linkage disequilibrium is unbiased when the sire is a homo-heterozygote.

Sire double heterozygote: Method of Excoffier and Slatkin (1995) for unrelated individuals: Assuming linkage phase *TM/tm* in the sire, the expected frequency of haplotype *TM* among half-sib progeny can be approximated by

$$E[\hat{f}_{TM}] \approx \frac{\frac{1}{2}N_i(1-c) + N_i f_{TM}}{2N_i} = \frac{1}{4}(1-c)+\frac{1}{2}f_{TM},$$

where $\frac{1}{2}N_i(1 - c)$ and $N_i\, f_{TM}$ are the sire and dams contributions of haplotype *TM* among the offspring.

Similarly, the expected frequencies of alleles *T* and *M* are approximated by

$$E[\hat{f}_T] \approx \frac{\frac{1}{2}N_i + N_i f_T}{2N_i} = \frac{1}{4}+\frac{1}{2}f_T$$

$$E[\hat{f}_M] \approx \frac{\frac{1}{2}N_i + N_i f_M}{2N_i} = \frac{1}{4}+\frac{1}{2}f_M.$$

The expected linkage disequilibrium is

$$\begin{aligned} E[\hat{D}_{TM}] &\approx E[\hat{\delta}] \\ &\approx E[\hat{f}_{TM}] - E[\hat{f}_T]E[\hat{f}_M] \\ &\approx \frac{1}{2}D_{TM} + \frac{1}{4}(1-c) + \frac{1}{4}f_T f_M - \frac{1}{8}\left(\frac{1}{2}+f_T+f_M\right). \end{aligned}$$

Consequently, the bias for using the method of Excoffier and Slatkin (1995) for unrelated individuals is approximated by

$$\begin{aligned} \mathrm{Bias} &\approx D_{TM} - E[\hat{D}_{TM}] \\ &\approx \frac{1}{2}D_{TM} - \frac{1}{4}\left(1-c+f_T f_M\right) + \frac{1}{8}\left(\frac{1}{2}+f_T+f_M\right). \end{aligned}$$

Sire double heterozygote: Estimation of LD using informative maternal haplotypes in half-sib families: The only informative haplotypes that can be traced up to their mothers are from progeny with genotypes *TTMM*, *TTmm*, *ttMM*, and *ttmm*. It is because markers are biallelic and only homozygote progeny can be used to trace inheritance when the sire is a heterozygote. If allele frequencies in the dam population are known, then determining the haplotype with the highest probability is feasible. Nevertheless, for intermediate allele frequencies the probability of inheriting either allele is 0.5. For the calculations below, only informative progeny is used.

Assuming linkage phase *TM/tm*, the expected frequency of informative *TM* haplotypes among progeny is

$$E[f_{TM}] \approx \frac{\frac{1}{2}(1-c)f_{TM}}{\frac{1}{2}(1-c)f_{TM} + \frac{1}{2}(c)f_{Tm} + \frac{1}{2}(c)f_{tM} + \frac{1}{2}(1-c)f_{tm}}.$$

The expected values for the frequencies of alleles *T* and *M* are

$$E[f_T] \approx \frac{(1-c)f_{TM} + cf_{Tm}}{(1-c)f_{TM} + (c)f_{Tm} + (c)f_{tM} + (1-c)f_{tm}}$$

$$E[f_M] \approx \frac{(1-c)f_{TM} + cf_{tM}}{(1-c)f_{TM} + (c)f_{Tm} + (c)f_{tM} + (1-c)f_{tm}}.$$

The expected disequilibrium is

$$E[\hat{D}_{TM}] \approx \frac{(1-c)f_{TM}}{(1-c)f_{TM} + (c)f_{Tm} + (c)f_{tM} + (1-c)f_{tm}} - \frac{[(1-c)f_{TM} + cf_{Tm}][(1-c)f_{TM} + cf_{tM}]}{[(1-c)f_{TM} + (c)f_{Tm} + (c)f_{tM} + (1-c)f_{tm}]^2}.$$

For unlinked loci, $c = 0.5$, the above expression reduces to D_{TM} and the method of informative maternal haplotypes is unbiased.

For $0 < c < 0.5$, the bias for using only maternal inherited haplotypes is approximated by

$$\begin{aligned} \text{Bias} &\approx D_{TM} - E[\hat{D}_{TM}] \\ &\approx D_{TM} - \left[\frac{(1-c)f_{TM}}{(1-c)f_{TM} + (c)f_{Tm} + (c)f_{tM} + (1-c)f_{tm}} \right. \\ &\qquad \left. + \frac{[(1-c)f_{TM} + cf_{Tm}][(1-c)f_{TM} + cf_{tM}]}{[(1-c)f_{TM} + (c)f_{Tm} + (c)f_{tM} + (1-c)f_{tm}]^2}\right]. \end{aligned}$$

Monte Carlo computer simulation

A Monte Carlo computer simulation was carried out to validate methods for estimating LD proposed in this article as well as to compute power. Three scenarios were simulated corresponding to the three possible situations regarding the genotype of the sire: double homozygote, homo-heterozygote, and double heterozygote. In addition, a multifamily situation was also simulated.

Sire double homozygote: A random generator from the uniform distribution was used to assign progeny with the haplotypes *TM*, *Tm*, *tM*, and *tm* according to their probability (frequency): $f_{TM} = \delta + f_T f_M$, $f_{Tm} = -\delta + f_T f_m$, $f_{tM} = -\delta + f_t f_M$, and $f_{tm} = \delta + f_t f_m$, where the allele frequencies f_M, f_m, f_T, f_t, and δ were input parameters. If the drawing of the uniform distribution was between 0 and f_{TM}, then the offspring inherited haplotype *TM* from his dam. If the drawing of the uniform distribution was between f_{TM} and $f_{TM+}\ f_{Tm}$ then the offspring inherited haplotype *Tm* from his dams. Assigning offspring to other haplotypes was done following the same rule.

Sire homo-heterozygote: A random generator from the uniform distribution was used to assign progeny with the genotypes *TTMM*, *TTMm*, *TTmm*, *TtMM*, *TtMm*, and *Ttmm* according to their probability (frequency): $\phi_{TTMM} = \frac{1}{2} f_{TM}$, $\phi_{TTMm} = \frac{1}{2} f_{Tm} + \frac{1}{2} f_{TM}$, $\phi_{TTmm} = \frac{1}{2} f_{Tm}$, $\phi_{TtMM} = \frac{1}{2} f_{tM}$, $\phi_{TtMm} = \frac{1}{2} f_{tM} + \frac{1}{2} f_{tm}$, and $\phi_{Ttmm} = \frac{1}{2} f_{tm}$. If the drawing of the uniform distribution was between 0 and ϕ_{TTMM}, then the offspring had genotype *TTMM*. If the drawing of the uniform distribution was between ϕ_{TTMM} and $\phi_{TTMM+}\ \phi_{TTMm}$ then the offspring genotype was *TTMm*. Assigning other genotypes to offspring was done following the same rule.

Sire double heterozygote: A random generator from the uniform distribution was used to assign progeny with the genotypes *TTMM*, *TTMm*, *TTmm*, *TtMM*, *TtMm*, *Ttmm*, *ttMM*, *ttMm*, and *ttmm* according to their probability (frequency): $\phi_{TTMM} = \frac{1}{2}(1-c)f_{TM}$, $\phi_{TTMm} = \frac{1}{2}(1-c)f_{Tm} + \frac{1}{2}c\, f_{TM}$, $\phi_{TTmm} = \frac{1}{2}c\, f_{Tm}$, $\phi_{TtMM} = \frac{1}{2}(1-c)f_{tM} + \frac{1}{2}c\, f_{TM}$, $\phi_{TtMm} = \frac{1}{2}(1-c)(f_{tm} + f_{TM}) + \frac{1}{2}c(f_{tM} + f_{Tm})$, $\phi_{Ttmm} = \frac{1}{2}(1-c)f_{Tm} + \frac{1}{2}c\, f_{tm}$, $\phi_{ttMM} = \frac{1}{2}c\, f_{tM}$, $\phi_{ttMm} = \frac{1}{2}(1-c)f_{tM} + \frac{1}{2}c\ f_{tm}$, and $\phi_{ttmm} = \frac{1}{2}(1-c)f_{tm}$. If the drawing of the uniform distribution was between 0 and ϕ_{TTMM}, then the offspring had genotype *TTMM*. If the drawing of the uniform distribution was between ϕ_{TTMM} and $\phi_{TTMM+}\phi_{TTMm}$ then the offspring genotype was *TTMm*. Assigning other genotypes to offspring was performed following the same rule.

Subroutines in Fortran 90 were written to estimate linkage disequilibrium with the half-sib methods (HS) described in this article as well as the method of Excoffier and Slatkin (1995) (ES) for unrelated individuals, and by making use of maternal informative haplotypes (MIH). Family sizes of 36 and 500 were used to test the methods in small and large families. Empirical power was computed by sorting within each simulation set according to the likelihood-ratio estimate and finding the percentage of replicates that gave a value higher than the value of the χ^2 with 1 d.f. at a significance level of 0.01.

Multifamily estimation of linkage disequilibrium: A total of six families with sizes 94, 77, 106, 81, 79, and 100 half-sib progeny resembling the sire Norwegian cattle population were simulated (after pooling selected and culled bulls in Table 1 of Gomez-Raya *et al.* 2002). The allele frequencies were intermediate, recombination fraction was 0, 0. 25, or 0.50, and linkage disequilibrium ranged from 0 to 0.25. The sires were simulated as if they were coming from a population with the same linkage disequilibrium and allele frequencies as used to generate the half-sib progeny. To do so, the two haplotypes at each sire were generated following the same principles as above with probabilities according to the simulated frequencies: $f_{TM} = \delta + f_T f_M, f_{Tm} = -\delta + f_T f_m$, $f_{tM} = -\delta + f_t f_M$, and $f_{tm} = \delta + f_t f_m$, in which allele frequencies f_M, f_m, f_T, f_t, and δ were input parameters. Thus, the sire could be a double homozygote, homo-heterozygote, or double heterozygote after assigning the two haplotypes. The half-sib progeny was generated as described in the previous section. Estimation of linkage disequilibrium was carried out using the EM algorithm for multiple families. Empirical power and overall likelihood-ratio test were computed for each simulation set. Each experiment was replicated 10,000 times. A Q-Q plot (using proc qqplot of SAS Inst., Cary, NC) was used to investigate the distribution of $\mathrm{LRT}_{\mathrm{joint}}$ under the null hypothesis (simulated $\delta = 0$) in the situation for $c = 0$.

Genome analyses of LD in a beef cattle half-sib family

A half-sib family consisting of 36 calves from commercial beef cattle at the Gund Ranch in Nevada was used to illustrate and to compare alternative methods for estimation of linkage disequilibrium. The first step was to determine paternity of the calves at the ranch. A set of 25 microsatellites (BMS410, BMS499, BMS650, BMS1244, BMS1634, TGLA227, BMS601, BMS1789, BMS2005, ILSTS081, BMS1315, BMS1226, BMS2573, ILSTS058, TGLA126, CSSM66, SPS115, TGLA53, BM1824, BM2113, ETH3R, TGLA122, INRA023, ETH225, ETH10) was used to assign paternity that was carried out using Cervus software. Total DNA from ear notches of calves and sires was purified using the manufacturer's instructions (Qiagen, CA). The DNA was diluted with AE buffer to 10 ng/μl and stored at −4° prior to genotyping. Primers were diluted to 50 μM and stored at −4°. A primer mix was prepared containing 2 μl of each 50 μM primer set. Each PCR reaction contained a total volume of 15 μl consisting of 1.5 μl of each primer mix, 2 μl water, 4 μl DNA, and 7.5 μl PCR multiplex mix (Qiagen). Gradients were performed to determine the optimal temperature for primer annealing. Amplification was carried out with a TC-512 Thermal Cycler (Techne). The initial denaturation step was performed at 95° for 15 min, followed by 35 cycles of 30 sec at 94°, 1 min and 30 sec at the optimum annealing temperature, and 1 min at 72° with a final extension of 30 min at 60°. Subsequently, 1 μl of PCR product was added to 199 μl water to make a 1:200 dilution. One microliter of this dilution was added to 10 μl of a formamide solution containing 1 ml formamide and 5 μl of ladder and denatured for 5 min at 95°. Genotyping was performed with the Applied Biosystems (ABI) Prism 3730 DNA analyzer.

Table 3 Average estimates of δ in a half-sib family from a homo-heterozygote sire (family size = 36 or 500) with simulated $f_T = 0.5$ and $f_M = 0.5$ and varying linkage disequilibrium (δ)

	Estimation method						
	HS		ES			MIH	
Family size: Simulated δ	36	500	36	500	E(ES)	36	500
0.000	0.001	0.000	0.000	0.000	0.000	0.001	0.000
0.025	0.025	0.025	0.017	0.018	0.013	0.025	0.025
0.050	0.049	0.050	0.034	0.036	0.025	0.050	0.050
0.075	0.074	0.075	0.049	0.051	0.038	0.074	0.075
0.100	0.098	0.100	0.063	0.064	0.050	0.098	0.100
0.125	0.122	0.125	0.075	0.077	0.063	0.122	0.125
0.150	0.146	0.150	0.087	0.088	0.075	0.146	0.150
0.175	0.170	0.175	0.097	0.098	0.087	0.170	0.175
0.200	0.195	0.200	0.106	0.107	0.100	0.194	0.200
0.225	0.218	0.225	0.115	0.116	0.113	0.218	0.225
0.250	0.243	0.250	0.123	0.125	0.125	0.249	0.250

The number of replicates was 10^4. HS, Average estimates using the method derived for half-sibs in this article. ES, Average estimates over replicates of linkage disequilibrium using the algorithm of Excoffier and Slatkin (1995). *E*(ES), Predicted LD using the method of not family structure using the algorithm of Excoffier and Slatkin (1995). MIH, Method of maternal informative haplotypes.

The Illumina bovine 50K BeadChip was used with bull 302 and his 36 calves to compare methods for estimating LD in half-sib families. The genotyping was carried out at the Core Lab of the University of Colorado, Denver. Only SNPs with a call rate >0.80 in at least 24 calves and MAF of 0.10 or more were used. The data were also filtered for SNPs that were not consistent for inheritance from sire to progeny. If a SNP was not consistent for one progeny then the SNP information was discarded for the entire family. Only pairs of SNPs within the same chromosome and within a distance of 50 Mb or less were used for estimating linkage disequilibrium. For the double heterozygote sire, recombination fraction and linkage phase was estimated using the methods proposed by Gomez-Raya (2001). Only SNPs with a recombination fraction of 0.30 or less were used for SNPs in which the sire was double heterozygote. Estimation of disequilibrium was performed using the half-sib method as well as the method of Excoffier and Slatkin (1995) and by making use of maternal informative haplotypes. For comparison of alternative estimation methods of linkage disequilibrium the statistic $r^2 = \delta^2/\{(f_T\ (1 - f_T)f_M\ (1 - f_M)\}$ was used. This statistic is widely used and ranges from 0 to 1, which facilitates comparison among methods. The absolute value of the difference between estimates of either ES or MIH and estimates HS were also used to evaluate discrepancies between methods.

Results

Table 3 shows simulation results for estimating linkage disequilibrium in a half-sib family from a homo-heterozygote sire with 36 or 500 progeny and dam allele frequencies of 0.5 at both SNPs. For these allele frequencies, the maximum possible linkage disequilibrium, δ, is 0.25. The method proposed in Equation 3 of this article (HS) yields identical estimates to the true (simulated) values of linkage disequilibrium with large family size (500). There was very little

Table 4 Average estimates of δ in a half-sib family from a double heterozygote sire (family size = 36 and 500) with simulated $f_T = 0.5$ and $f_M = 0.5$ and varying recombination fraction (*c*) and linkage disequilibrium (δ)

		Simulated *c*					
		0		0.25		0.50	
Simulated δ	Family size	36	500	36	500	36	500
0.000	HS	−0.004	−0.000	−0.000	0.000	0.000	0.000
	ES	0.106	0.108	0.057	0.059	0.000	0.000
	E(ES)		0.125		0.063		0.000
	MIH	0.220	0.250	0.110	0.125	0.000	0.000
	E(MIH)		0.250		0.125		0.000
0.100	HS	0.092	0.099	0.094	0.099	0.094	0.100
	ES	0.166	0.169	0.110	0.112	0.047	0.048
	E(ES)		0.175		0.113		0.050
	MIH	0.229	0.249	0.186	0.187	0.088	0.100
	E(MIH)		0.250		0.188		0.100
0.200	HS	0.188	0.199	0.183	0.199	0.194	0.199
	ES	0.221	0.224	0.158	0.161	0.088	0.090
	E(ES)		0.225		0.163		0.100
	MIH	0.234	0.249	0.213	0.231	0.176	0.200
	E(MIH)		0.250		0.232		0.200

The number of replicates was 10^4. HS, Average estimates using the method derived for half-sibs in this article. ES, Average estimates over replicates of linkage disequilibrium using the algorithm of Excoffier and Slatkin (1995). *E*(ES), Predicted LD using the method of not family structure using the algorithm of Excoffier and Slatkin (1995). MIH, Method of Maternal Informative Haplotypes. *E*(MIH), Predicted LD using the method of Maternal Informative Haplotypes.

bias when the family size is small (36). For the examples in Table 3, the bias using the HS method is <3% . On the other hand, estimates are biased when using the method of Excoffier and Slatkin (1995), which becomes just half of the true disequilibrium at $\delta = 0.25$. The approximation for predicting the expected value for the estimates of linkage disequilibrium using the method of Excoffier and Slatkin (1995) agreed well with the simulation results but tends to underestimate it. On the other hand, the use of maternal informative haplotypes is unbiased, as shown in Table 3 and as proven in the corresponding section of this article.

Table 4 shows simulation results for estimating linkage disequilibrium in a half-sib family from a double heterozygote sire for varying recombination fractions and linkage disequilibrium parameters. The allele frequencies at the two loci were 0.5. Each simulation set was analyzed with the EM algorithm developed in this article (HS) as well as the method of ES and by using MIH from dams. The HS method is asymptotically unbiased with average estimates of disequilibria very close to the simulated (true) parameters for large family sizes (500). For small family sizes (36) the estimates of linkage disequilibrium are slightly downward biased. The method of Excoffier and Slatkin (1995) is severely biased upward at low recombination fractions but becomes biased downward at high recombination fractions. The use of only maternal informative haplotypes to estimate disequilibrium is upward biased at low recombination fractions but becomes unbiased when the markers are unlinked. The approximated expected disequilibrium was very close to what was observed in the simulation for both the method of Excoffier and Slatkin (1995) and when using informative maternal haplotypes from dams. Figures 1 and 2 show expected bias in estimating disequilibrium in a half-sib family from a double heterozygote sire for two scenarios regarding allele frequencies at the DNA markers: $f_T = 0.5, f_M = 0.5$ and $f_T = 0.4$, $f_M = 0.1$. For a low recombination fraction, bias is negative but becomes positive as recombination fraction increases. The effect is more pronounced for loci at intermediate allele frequencies than for loci with allele frequencies closer to fixation.

In many instances, interest is on the amount of progeny needed for detecting linkage disequilibrium. Empirical power for half-sib families in which the sire was a double homozygote, homo-heterozygote, or double heterozygote is shown in Table 5. The standard deviations among replicates for the same simulation sets are given in Table 6. The simulation results are for varying family sizes and true (simulated) linkage disequilibrium parameters. Disequilibrium (δ) of 0.10 was detected with groups of 100 offspring in most situations. The most powerful situation is when the sire is a homozygote at two loci and all haplotypes are informative. Power in a double heterozygote sire family reduces with genetic distance but it is nearly as powerful as the double homozygote for fully linked loci. Power in a homo-heterozygote sire family is always lower than power in a double homozygote sire family. Standard deviation among replicates follows the same trend as power (Table 6). The double homozygote and the double heterozygote (at $c = 0$) sire families had the lowest variation (Table 6). Variation among replicates increases with increasing recombination fraction in double heterozygote families. The estimates of disequilibrium for homo-heterozygote had more variation than double homozygote families. There was good agreement between the observed standard deviation among replicates and the

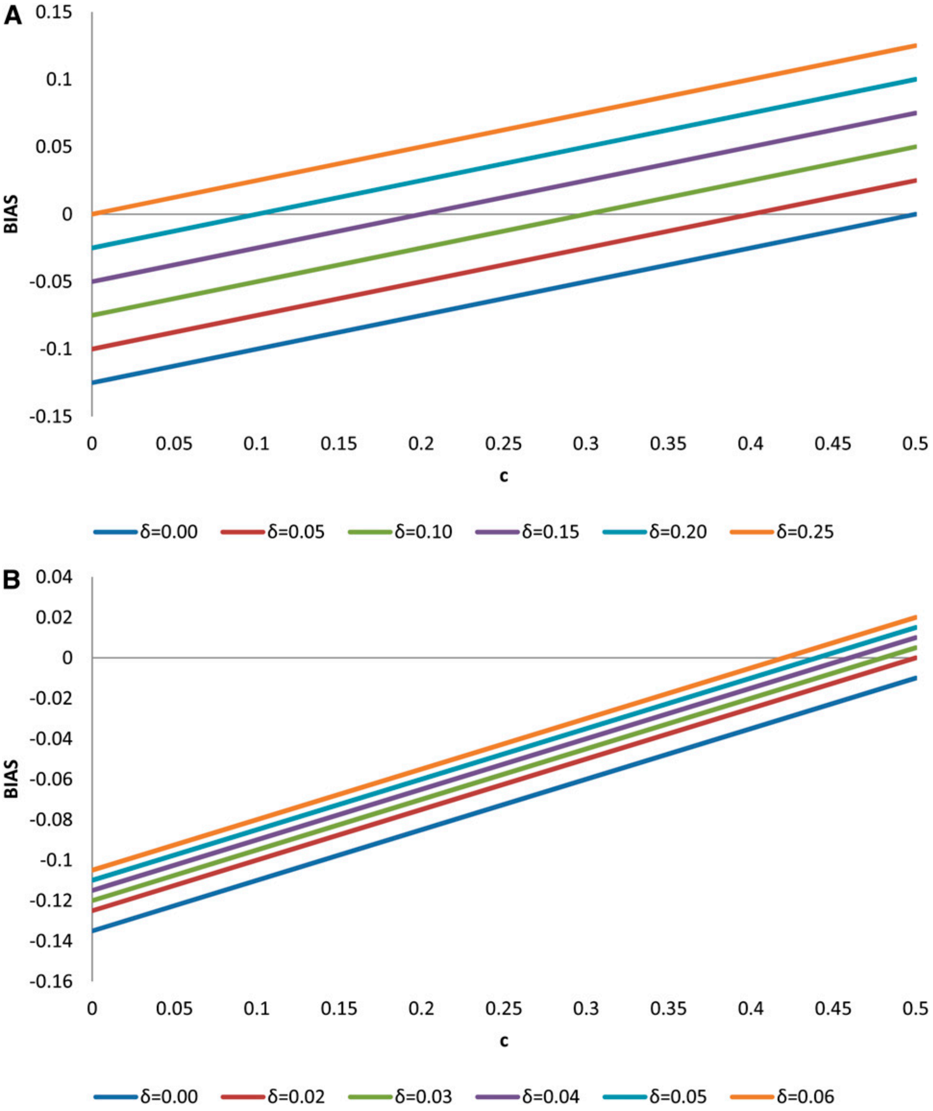

Figure 1 Bias in the estimation of LD using the Excoffier and Slatkin (1995) algorithm in a half-sib family from a double heterozygote sire. (A) $f_T = 0.5$, $f_M = 0.5$, (B) $f_T = 0.4$, $f_M = 0.1$

average of the estimates of sampling standard deviations of δ obtained in each replicate.

The simulation results for estimating LD using multiple sire families are given in Table 7. There was good agreement between simulated and estimated linkage disequilibrium across the range of simulated disequilibrium parameter and recombination fraction. The range of the absolute difference between estimated and simulated δ was between 0.000 and 0.008. A Q-Q plot of the quantiles of the observed distribution of $\text{LRT}_{\text{joint}}$ under the null hypothesis against quantiles from a γ-distribution with shape = 0.5 and scale = 2 is depicted in Figure 3. This gamma distribution is a χ^2-distribution with 1 d.f. The cumulative distribution of $\text{LRT}_{\text{joint}}$ showed larger variation than a χ^2-distribution with 1 d.f.

Linkage disequilibrium was also estimated in a cattle half-sib family using the Illumina 50K BeadChip. There were 0.00189% inconsistencies between genotypes of sire and calves. Table 8 shows the overall estimates of r^2 using HS, ES, and MIH for those situations in which the sire was a homo-heterozygote. There were 314,730 SNP pairs for the entire autosomal genome with average estimates of r^2 of 0.115, 0.067, and 0.111 for HS, ES, and MIH methods. The ES method is downward biased since estimates by this method were around half of their value using the half-sib method. The maternal informative haplotype estimates were slightly lower than those obtained by the half-sib method, which might be due to bias because of reduced family size (noninformative offspring is neglected from these analyses).

Table 9 shows overall estimates of r^2 using HS, ES, and MIH for SNPs for which the sire was a double heterozygote. There were 208,872 SNP pairs. The results using real data support earlier findings showing that the methods of Excoffier and Slatkin (1995) and maternal informative haplotypes were upward biased. The average estimates of r^2 across the genome were of 0.100, 0.267, and 0.925 for HS, ES, and MIH methods.

Figure 4 shows average estimates of r^2 for the three methods of estimation across the entire genome when the distance between the two SNPs is between 10 and 50 Mb.

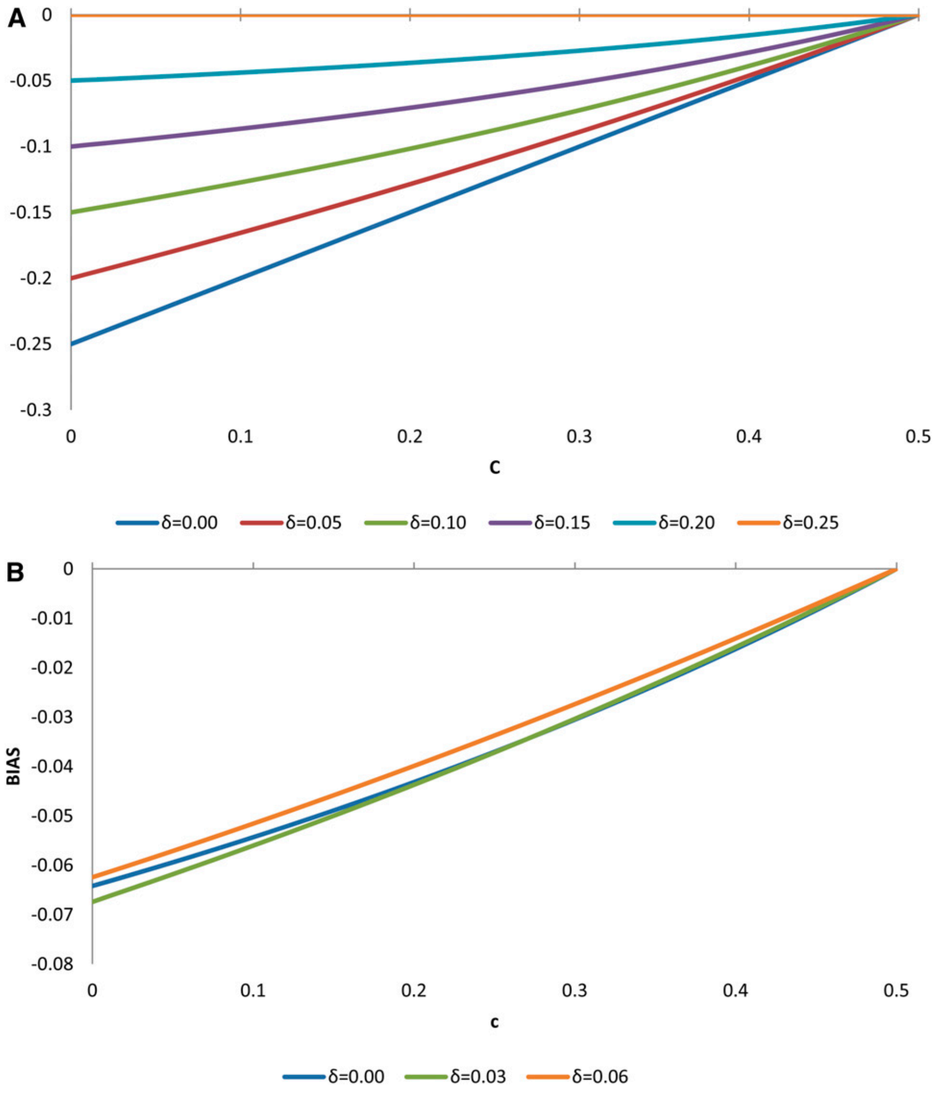

Figure 2 Bias in the estimation of LD using maternal haplotypes in a half-sib family from a double heterozygote sire. (A) $f_T = 0.5$, $f_M = 0.5$, (B) $f_T = 0.4$, $f_M = 0.1$.

A total of 829,042 SNPs pairs were tested and the estimates are a pool of all three possible situations regarding sire genotypes: double homozygote, homo-heterozygote, and double heterozygote. This figure shows again that using either ES or MIH methods give estimates upward biased of linkage disequilibria.

Table 5 Empirical power for estimation of LD in a half-sib families from a double homozygote, homo-heterozygote and a double heterozygote sire for varying family sizes

				Hetero-hetero: Simulated *c*		
Simulated δ	Size	Homo-homo	Homo-hetero	0	0.25	0.50
0	100	0.01	0.01	0.01	0.01	0.01
0.025	100	0.06	0.04	0.05	0.03	0.02
	200	0.12	0.06	0.11	0.05	0.03
	500	0.37	0.16	0.36	0.13	0.07
	1000	0.72	0.37	0.71	0.31	0.16
0.050	100	0.29	0.13	0.27	0.11	0.07
	200	0.60	0.30	0.59	0.24	0.13
	500	0.98	0.73	0.97	0.64	0.38
	1000	1.00	0.97	1.00	0.94	0.73
0.100	100	1.00	0.64	0.93	0.51	0.32
	200	1.00	0.93	1.00	0.87	0.63
	500	1.00	1.00	1.00	1.00	0.97
	1000	1.00	1.00	1.00	1.00	1.00

The allele frequencies were $f_T = 0.5$ and $f_M = 0.5$. The significance level was 0.01. The number of replicates was 10^4.

Table 6 Standard deviation among replicates in the estimation of LD in half-sib families from double homozygote, homo-heterozygote and double heterozygote sire for varying family sizes and $f_T = 0.5$ and $f_M = 0.5$

				Hetero-hetero: Simulated *c*		
Simulated δ	Size	Homo-homo	Homo-hetero	0	0.25	0.50
0	100	0.025 (0.025)	0.035 (0.035)	0.025 (0.025)	0.038 (0.038)	0.052 (0.049)
0.025	100	0.024 (0.025)	0.035 (0.034)	0.025 (0.025)	0.038 (0.038)	0.052 (0.049)
	200	0.017 (0.018)	0.025 (0.025)	0.018 (0.017)	0.027 (0.027)	0.035 (0.035)
	500	0.011 (0.011)	0.016 (0.016)	0.011 (0.011)	0.017 (0.017)	0.023 (0.022)
	1000	0.008 (0.008)	0.011 (0.011)	0.008 (0.008)	0.012 (0.012)	0.016 (0.016)
0.050	100	0.024 (0.025)	0.035 (0.034)	0.024 (0.024)	0.038 (0.037)	0.051 (0.048)
	200	0.017 (0.017)	0.025 (0.024)	0.017 (0.017)	0.027 (0.027)	0.035 (0.034)
	500	0.011 (0.011)	0.016 (0.016)	0.011 (0.011)	0.017 (0.017)	0.022 (0.022)
	1000	0.008 (0.008)	0.011 (0.011)	0.008 (0.008)	0.012 (0.012)	0.015 (0.015)
0.100	100	0.023 (0.021)	0.033 (0.031)	0.023 (0.022)	0.038 (0.037)	0.048 (0.045)
	200	0.016 (0.016)	0.023 (0.022)	0.016 (0.016)	0.027 (0.026)	0.033 (0.032)
	500	0.010 (0.010)	0.015 (0.014)	0.010 (0.010)	0.017 (0.017)	0.021 (0.020)
	1000	0.007 (0.007)	0.010 (0.010)	0.007 (0.007)	0.012 (0.012)	0.015 (0.014)

The number of replicates was 10^4. Values between brackets are average of the estimates of sampling standard deviations of δ obtained in each replicate.

Discussion

Early studies estimating linkage disequilibrium in populations with a half-sib structure were carried out using microsatellites (Farnir *et al.* 2000; Odani *et al.* 2006). These studies used maternal alleles and estimated the most likely haplotype inherited in sons from dams. Although the methods derived in this article are for biallelic loci such as SNPs, a multiallelic marker can always be reduced to a biallelic one after pooling alleles into two groups. As shown by Gomez-Raya (2001) for a double heterozygote sire, the amount of informative progeny in a half-sib family depends on the recombination fraction between the two markers. Thus, the frequency of informative progeny is $c([1-f_t)(1-f_M) + (1-f_T)(1-f_m)]$, and $(1-c)[(1-f_t)(1-f_m) + (1-f_T)(1-f_M)]$ for recombinants and nonrecombinants, respectively. Genotypes among offspring that are informative for tracing inheritance from sires are also informative for tracing alleles inherited from dams (with unknown genotypes). For example, for allele frequencies $f_T = f_M = 0.1$, the frequency of informative recombinant and nonrecombinant progeny is $0.18c$ and $0.82(1-c)$, which means that the closer the markers are, the lower the proportion of informative recombinants among progeny. Therefore, bias in estimating linkage disequilibrium occurs because of the altered proportion of haplotypes that are informative at varying genetic distances. Nevertheless, Farnir *et al.* (2000) used not just the informative haplotypes but the most likely haplotype. Sires carrying alleles at low frequency would allow identification of haplotypes with a higher probability, which will reduce the magnitude of the bias as shown in this article. In another study, also investigating linkage disequilibrium with microsatellites in cattle, Tenesa *et al.* (2003) made use of the method of Excoffier and Slatkin (1995) to

Table 7 Average estimates of linkage disequilibrium (δ) using the EM algorithm for multiple half-sib families together with statistical power at significance level of 0.01

	Simulated *c*					
	0		0.25		0.50	
Simulated δ	δ	Power	δ	Power	δ	Power
0.000	0.0000 (0.013)	0.02	−0.0001 (0.014)	0.02	−0.0001 (0.014)	0.02
0.010	0.0100 (0.013)	0.05	0.0099 (0.014)	0.05	0.0099 (0.014)	0.05
0.020	0.0201 (0.013)	0.20	0.0201 (0.014)	0.17	0.0200 (0.014)	0.16
0.030	0.0301 (0.013)	0.45	0.0301 (0.014)	0.37	0.0301 (0.014)	0.34
0.040	0.0401 (0.013)	0.73	0.0402 (0.014)	0.62	0.0401 (0.014)	0.58
0.050	0.0501 (0.012)	0.91	0.0502 (0.014)	0.83	0.0501 (0.014)	0.79
0.075	0.0751 (0.012)	1.00	0.0751 (0.013)	0.99	0.0752 (0.013)	0.99
0.100	0.1002 (0.013)	1.00	0.1003 (0.013)	1.00	0.1003 (0.013)	1.00
0.125	0.1252 (0.010)	1.00	0.1256 (0.012)	1.00	0.1255 (0.012)	1.00
0.150	0.1502 (0.009)	1.00	0.1506 (0.011)	1.00	0.1506 (0.011)	1.00
0.175	0.1753 (0.008)	1.00	0.1758 (0.010)	1.00	0.1758 (0.010)	1.00
0.200	0.1997 (0.007)	1.00	0.2003 (0.008)	1.00	0.2002 (0.008)	1.00
0.250	0.2457 (0.015)	1.00	0.2443 (0.013)	1.00	0.2453 (0.017)	1.00

The simulated allele frequencies were $f_T = 0.5$ and $f_M = 0.5$. The simulation was carried out for varying recombination fractions (*c*), linkage disequilibria (δ) and resembling the Norwegian cattle population structure. The number of replicates was 10^4. The values between brackets are the average of standard deviations of the estimates of δ.

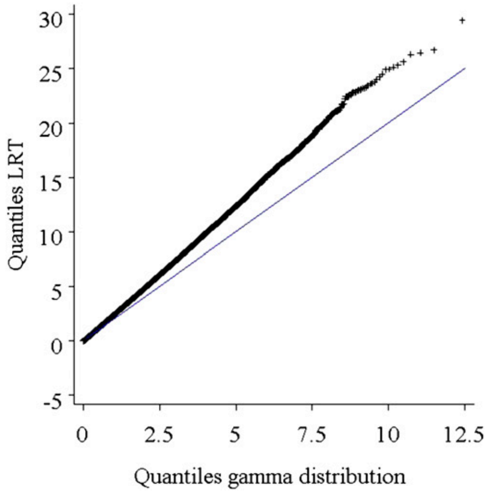

Figure 3 Q-Q plots of likelihood-ratio test using the multi-half EM algorithm on 10^6 replicates. Quantiles LRT are the quantiles from simulated data for $c = 0$ and under the null hypothesis ($\delta = 0$).

estimate linkage disequilibrium. As shown in this article, estimates of LD using that method for unrelated individuals might lead to severe biased estimation when applied in animals with a half-sib structure.

Improvement in the sequencing methods in the last years allowed for the discovery of vast amounts of SNPs in the human and animal genomes (*e.g.*, International Hap Map Consortium 2007). A following step has been the construction of LD maps for the human (Maniatis 2002) and animal genomes (*e.g.*, Khatkar *et al.* 2006). LD maps are based on: (a) estimation of a linkage disequilibrium parameter, ρ, which has the same maximum absolute value as the statistics D' of Lewontin (1964), and (b) use of a model of decay of disequilibrium leading to equations of Malecot's model for isolation by distance (Malecot 1964). Thus, the value of D' is δ/D_{Max} with $D_{\text{Max}} = \min\{ f_T\ (1 - f_M), f_M\ (1 - f_T)\}$. Construction of LD maps are carried out estimating ρ between adjacent SNPs and by using composite maximum likelihood for all pairs of adjacent SNPs. Inferences of the decay of disequilibrium over time are made by $\rho = (1-L)Me^{-\varepsilon d} + L$, where L is a parameter that reflects the residual association at a long distance (d), M is the association at zero distance, and ε is the exponential decline in LD due to recombination over generations. In human genetics, estimation of ρ is performed using unrelated individuals and the Excoffier and Slatkin (1995) algorithm. The construction of LD maps in species with a half-sib family structure like cattle would require methods for the estimation of disequilibrium (δ) as

Table 8 Overall values for estimates of r^2 and abs($r^2_{\text{ES}} - r^2_{\text{HS}}$) and abs($r^2_{\text{MIH}} - r^2_{\text{HS}}$) for pairs of SNPs for which the sire was homo-heterozygote using alternative methods of estimation: r^2_{HS} (half-sib), r^2_{ES} (Excoffier and Slatkin 1995), and r^2_{MIH} (maternal informative haplotypes)

Chromosome	No. of pairs	r^2_{HS}	r^2_{ES}	r^2_{MIH}	abs ($r^2_{ES} - r^2_{HS}$)	abs ($r^2_{MIH} - r^2_{HS}$)
1	23181	0.108	0.061	0.106	0.061	0.015
2	18656	0.108	0.067	0.105	0.062	0.016
3	13730	0.124	0.073	0.120	0.070	0.017
4	16086	0.115	0.068	0.112	0.062	0.014
5	11164	0.132	0.077	0.129	0.075	0.016
6	19674	0.129	0.076	0.126	0.075	0.017
7	11445	0.133	0.076	0.130	0.076	0.016
8	16238	0.111	0.062	0.107	0.061	0.015
9	14253	0.118	0.068	0.114	0.067	0.017
10	14402	0.097	0.057	0.093	0.053	0.012
11	9344	0.132	0.077	0.128	0.076	0.016
12	9914	0.100	0.061	0.096	0.055	0.014
13	11553	0.124	0.069	0.121	0.069	0.014
14	11315	0.121	0.067	0.117	0.069	0.015
15	10410	0.126	0.072	0.122	0.070	0.018
16	7770	0.113	0.067	0.109	0.063	0.014
17	9220	0.101	0.065	0.098	0.056	0.013
18	8359	0.108	0.066	0.104	0.059	0.015
19	9025	0.105	0.065	0.100	0.060	0.015
20	6660	0.109	0.063	0.105	0.059	0.014
21	7337	0.104	0.060	0.099	0.059	0.014
22	6070	0.117	0.068	0.114	0.066	0.015
23	6754	0.106	0.064	0.103	0.061	0.016
24	8585	0.118	0.070	0.113	0.067	0.017
25	7735	0.123	0.071	0.119	0.071	0.016
26	6339	0.115	0.069	0.111	0.064	0.017
27	5468	0.114	0.063	0.110	0.062	0.014
28	7000	0.104	0.065	0.101	0.060	0.013
29	7043	0.101	0.059	0.097	0.057	0.013
Overall	314730	0.115	0.067	0.111	0.065	0.015

abs, the absolute value of the difference.

Table 9 Overall values for estimates of r^2 and abs(r^2_{ES} $-r^2_{HS}$) and abs(r^2_{MIH} $-r^2_{HS}$) for pairs of SNPs for which the sire was double heterozygote using alternative methods of estimation: r^2_{HS} (half-sib), r^2_{ES} (Excoffier and Slatkin 1995), and r^2_{MIH} (maternal informative haplotypes)

Chromosome	No. of pairs	r^2_{HS}	r^2_{ES}	r^2_{MIH}	abs(r^2_{ES} $-r^2_{HS}$)	abs(r^2_{MIH} $-r^2_{HS}$)
1	19344	0.089	0.266	0.933	0.207	0.850
2	12033	0.092	0.269	0.945	0.207	0.859
3	8415	0.122	0.294	0.962	0.216	0.848
4	10439	0.097	0.273	0.919	0.212	0.828
5	7380	0.115	0.287	0.920	0.217	0.814
6	13930	0.120	0.282	0.928	0.214	0.817
7	8369	0.131	0.281	0.940	0.217	0.821
8	10948	0.096	0.269	0.950	0.210	0.862
9	9631	0.106	0.281	0.929	0.216	0.830
10	7212	0.080	0.258	0.900	0.199	0.824
11	5530	0.115	0.283	0.929	0.221	0.823
12	8049	0.080	0.248	0.916	0.198	0.844
13	7982	0.105	0.271	0.923	0.212	0.829
14	5828	0.111	0.265	0.923	0.207	0.823
15	5789	0.113	0.264	0.911	0.210	0.812
16	5286	0.095	0.255	0.898	0.202	0.813
17	6340	0.091	0.270	0.916	0.209	0.830
18	6798	0.103	0.266	0.909	0.200	0.812
19	5334	0.079	0.242	0.894	0.193	0.823
20	4127	0.088	0.258	0.959	0.204	0.878
21	4107	0.083	0.246	0.898	0.200	0.825
22	4146	0.103	0.269	0.932	0.214	0.838
23	5161	0.094	0.262	0.912	0.204	0.828
24	5244	0.089	0.270	0.934	0.213	0.853
25	4703	0.104	0.245	0.889	0.196	0.798
26	3970	0.098	0.260	0.910	0.203	0.821
27	4651	0.092	0.248	0.933	0.205	0.853
28	4214	0.078	0.252	0.913	0.204	0.841
29	3912	0.082	0.239	0.922	0.191	0.847
Overall	208872	0.100	0.267	0.925	0.208	0.834

abs, the absolute value of the difference.

proposed in this article. If δ is biased then ρ should also be biased. If the bias depends upon the distance between the adjacent SNPs as shown here then inferences on population structure and the evolution of the cattle population may not be fully correct. Khatkar *et al.* (2006) carried out a LD map of bovine chromosome 6 using bulls from the Australian Holstein–Friesian. They estimated average coancestry by 0.012 using available pedigree information. Assuming that pedigrees were complete, coancestry was rather small but still might lead to bias in the estimation of the disequilibria currently present in the Australian dairy population.

The square of the correlation of alleles at two loci (r^2) has been widely used in animals with a half-sib structure to estimate linkage disequilibrium (McKay *et al.* 2007; de Roos *et al.* 2008; Hayes *et al.* 2008; Prasad *et al.* 2008; Sargolzaei *et al.* 2008; Bovine Hap Map Consortium 2009; Kim and Kirkpatrick 2009; Qanbari *et al.* 2010). Most of these studies identify phased haplotypes using available information from pedigrees. Haplotypes that could not be phased out were generally ignored. As shown in this article, the proportion of haplotypes that are informative might vary with genetic distance leading to biased estimation of linkage disequilibrium, δ, which would also lead to biased estimates of r^2. The magnitude of the bias depends on how much information from pedigrees can be used for phasing haplotypes and on the distances between the SNPs in the LD analyses. Many of the above studies made inferences about the population structure based on r^2. However, estimates of r^2 might be biased to a different extent for different cattle breeds having a different breeding structure (more or fewer half-sibs families

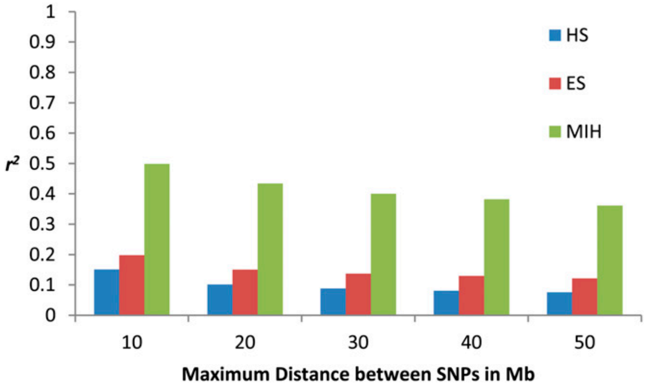

Figure 4 Average values of estimates r^2 across the genome using half-sib (HS), Excoffier and Slatkin (ES), and maternal informative haplotypes (MIH) methods for maximum distances between SNPs of 10, 20, 30, 40, and 50 Mb.

of different sizes). Comparison of r^2 estimated without consideration of the breeding structure in different animal populations should be taken with caution. On the other hand, inferences on past population sizes based on Sved's (1971) equation $E(r^2) = (1 + 4N_ec)^{-1}$ (where N_e is the effective population size) might also be inaccurate if r^2 has been estimated in half-sib families neglecting noninformative haplotypes.

Assumptions of this study were that linkage phase of the sire and recombination fractions were known without error. The linkage phase can be accurately estimated using the same data if progeny groups are not small (>25) and recombination fraction is not too high (<0.30). For other situations, such as those arising by the use of SNP arrays, linkage phase in the sire can be inferred for each of two adjacent SNPs when they are apart at small distances. Reconstruction of haplotypes for all SNPs for each of the two homologous chromosomes of the sire is then feasible. In the same way, the assumption of known recombination fraction will hold for most situations found in practice when SNPs are adjacent, *i.e.*, $c = 0$.

The methods developed in this article are for estimation of linkage disequilibrium present in the dam population and contributing to the half-sib progeny since sire haplotypes are ignored in the computations. In most circumstances, this disequilibrium is the most relevant since sires in dairy and beef cattle are likely related and the number of sire haplotypes is rather small. Nevertheless, if many sire families are available, then haplotype frequencies from sires and dams (estimated among half-sib progeny) can be pooled to obtain a joint estimate of linkage disequilibrium across sexes.

The results of the simulations showed that the proposed method for estimating disequilibrium works well for relatively small family sizes and in multifamily situations. The distribution of likelihood-ratio tests when simulating the null hypothesis showed that it had more variation than a χ^2 with 1 d.f. This is because likelihood equations for multiple sire families make use of a different number of parameters depending on the sire genotype: double homozygote (δ), homo-heterozygote (δ and f_T), and double heterozygote (δ, f_T, and f_M). In practical terms, resampling or simulation methods may be needed for hypothesis testing. The power figures for multifamily situations are also affected, being lower than those reported for this article.

The methods derived in this article were designed for estimating second-order linkage disequilibrium in half-sib families. The same methods and principles used in this article can be applied to the estimation of third- or higher-order linkage disequilibria. These methods may also be incorporated into a more general situation in which pedigrees are incomplete but much information comes from half-sib families. Nevertheless, if genotype information is available only from males (*e.g.*, genotyping information from a granddaughter design), then little information may be gained by incorporating maternal grandsire in the estimation of haplotype frequencies.

The conclusion of this article is that estimation of linkage disequilibrium in populations with a breeding structure of half-sib families must incorporate that structure in their estimation to provide unbiased estimates of the linkage disequilibrium. Inferences on population structure and evolution of cattle or sheep should be based on linkage disequilibria after accommodating the existing half-sib family structure in these populations.

Acknowledgments

I am very grateful to Dr. David Thain and Jon Wilker for the cattle samples at the Gund ranch and to Veronica Kirchoff for the genotyping of microsatellites. Samples for DNA testing were taken from Nevada Agricultural Experiment Station project no. NEV05339, DNA Paternity Testing and Genetic Improvement of Free Range Beef Cattle. The author is also grateful to Luis Alberto Garcia Cortes and Wendy M. Rauw for criticism of the manuscript. I thank two anonymous reviewers for suggestions that improved the manuscript.

Literature Cited

Barendse, W., D. Vaiman, S. J. Kemp, Y. Sugimoto, S. M. Armitage *et al.*, 1997 A medium density genetic linkage map of the bovine genome. Mamm. Genome 8: 21–28.

Bovine Hap Map Consortium, 2009 Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science: 324: 528–532.

Da, Y., and H. A. Lewin, 1995 Linkage information content and efficiency of full-sib and half-sib designs for gene mapping. Theor. Appl. Genet. 90: 699–706.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard, 2008 Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics 179: 1503–1512.

Excoffier, L., and M. Slatkin, 1995 Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12: 921–927.

Farnir, F., W. Coppiettiers, J.-J. Arranz, P. Berzi, N. Cambisano *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. Genome Res. 10: 220–227.

Gomez-Raya, L., 2001 Biased estimation of the recombination fraction using half-sib families and informative offspring. Genetics 157: 1357–1367.

Gomez-Raya, L., H. G. Olsen, H. Klungland, D. I. Våge, I. Olsaker *et al.*, 2002 The use of genetic markers to measure genomics response to selection in livestock. Genetics 162: 1381–1388.

Hayes, B. J., S. Lien, H. Nilsen, H. G. Olsen, P. Berg *et al.*, 2008 The origin of selection signatures on bovine chromosome 6. Anim. Genet. 39: 105–111.

International HapMap Consortium, 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Kappes, M. S., J. W. Keele, R. S. Stone, R. A. McGaw, T. S. Sonstegard *et al.*, 1997 A second-generation linkage map of the bovine genome. Genome Res. 7: 235–249.

Khatkar, M. S., A. Collins, J. A. Cavanagh, R. J. Hawken, M. Hobbs *et al.*, 2006 A first-generation metric linkage disequilibrium map of bovine chromosome 6. Genetics 174: 79–85.

Kim, E.-S., and B. W. Kirkpatrick, 2009 Linkage disequilibrium in the North American Holstein population. Anim. Genet. 40: 279–288.

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

Ma, R. Z., J. E. Beever, Y. Da, C. A. Green, I. Russ *et al.*, 1996 A male linkage map of the cattle (*Bos taurus*) genome. J. Hered. 87: 261–271.

Malecot, G., 1964 *Les mathématiques de l'hérédité*. Masson, Paris.

Maniatis, N., A. Collins, C. F. Xu, L. C. McCarthy, D. R. Hewett *et al.*, 2002 The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc. Natl. Acad. Sci. USA 99: 2228–2233.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts *et al.*, 2007 Whole genome linkage disequilibrium maps in cattle. BMC Genet. 8: 74.

Odani, M., A. Narita, T. Watanabe, K. Yokouchi, Y. Sugimoto *et al.*, 2006 Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. Anim. Genet. 37: 139–144.

Prasad, A., S. D. Mckay, B. Murdoch, P. Stothard, D. Kolbehdari *et al.*, 2008 Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. Anim. Genet. 39: 597–605.

Qanbari, S., E. C. Pimentel, J. Tetens, G. Thaller, P. Lichtner *et al.*, 2010 The pattern of linkage disequilibrium in German Holstein cattle. Anim. Genet. 41: 346–356.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer, 2008 Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91: 2106–2117.

Sved, J., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2: 125–141.

Tenesa, A., S. A. Knott, D. Ward, D. Smith, J. L. Williams *et al.*, 2003 Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. J. Anim. Sci. 81: 617–623.

Våge, D. I., I. Olsaker, H. Klungl, L. Gomez-Raya, and S. Lien, 2000 A male genetic map designed for QTL mapping in Norwegian cattle. Acta Agric. Scand. Sect. Anim. Sci. 50: 56–63.

Communicating editor: H. Zhao

Appendix A: Sampling Variances of the Estimates of Linkage Disequilibrium

The sampling variance of the estimates of the disequilibrium parameter for the *i*th family is

$$\mathrm{Var}(\hat{\delta}) \approx \frac{1}{\left[-\left(\partial^2 \ln L_i(\delta \mid nG)/\partial\delta^2\right)\right]_{\delta=\hat{\delta}}}.$$

The denominator of this equation is obtained by taking the second derivative respect to δ for each likelihood equation, which depends on the sires's genotype.

Sire Double Homozygote

The genotype of the sire is *TTMM*. To obtain an estimate of the sampling variance it is better to use a full-likelihood equation in which all sources of information are used to estimate δ,

$$L_i(\delta|nG) = K(f^i_{TM})^{n_{TM,i}}(f^i_{Tm})^{n_{Tm,i}}(f^i_{tM})^{n_{tM,i}}(f^i_{tm})^{n_{tm,i}},$$

where $n_{TM,i}$, $n_{Tm,i}$, $n_{tM,i}$, and $n_{tm,i}$ are the number of offspring inheriting haplotypes *TM*, *Tm*, *tM*, and *tm* and *K* is a constant.

Taking natural logarithms in the above equation after ignoring the constant, *K*, gives $\ln L_i(\delta|nG) = n_{TM,i} \ln(\hat{f}^i_{TM}) + n_{Tm,i}\ln(f^i_{Tm}) + n_{tM,i}\ln(f^i_{tM}) + n_{tm,i}\ln(f^i_{tm})$. The first two derivatives with respect to δ of this equation are

$$\frac{\partial \ln L_i(\delta|nG)}{\partial\delta} = \frac{n_{TM,i}}{\delta + \hat{f}_T\hat{f}_M} - \frac{n_{Tm,i}}{-\delta + \hat{f}_T\hat{f}_m} - \frac{n_{tM,i}}{-\delta + \hat{f}_t\,\hat{f}_M} + \frac{n_{tm,i}}{\delta + \hat{f}_t\,\hat{f}_m}$$

$$\frac{\partial^2 \ln L_i(\delta \mid nG)}{\partial\delta^2} = -\frac{n_{TM,i}}{(\delta + \hat{f}_T\hat{f}_M)^2} - \frac{n_{Tm,i}}{(-\delta + \hat{f}_T\hat{f}_m)^2} - \frac{n_{tM,i}}{(-\delta + \hat{f}_t\hat{f}_M)^2} - \frac{n_{tm,i}}{(\delta + \hat{f}_t\hat{f}_m)^2}.$$

Sire Homo-heterozygote

Let the sire have genotype *TTMm* at two SNPs, *T/t*, and *M/m*. The likelihood equation for the *i*th family is

$$\begin{aligned} L_i(\delta \mid nG) = K(\phi_{TTMM})^{n_{TTMM,i}}(\phi_{TTMm})^{n_{TTMm,i}}(\phi_{TTmm})^{n_{TTmm,i}}(\phi_{TtMM})^{n_{TtMM,i}} \\ \times\ (\phi_{TtMm})^{n_{TtMm,i}}(\phi_{Ttmm})^{n_{Ttmm,i}}, \end{aligned}$$

where ϕ_j are the probabilities of the *j*th genotype as described in the text. Ignoring the constant and taking natural logarithm of the above expression gives

$$\begin{aligned}\ln L_i(\delta \,|\, nG) \approx\; & n_{TTMM,i} \ln(\phi_{TTMM}) + n_{TTMm,i} \ln(\phi_{TTMm}) + n_{TTmm,i} \ln(\phi_{TTmm}) + n_{TtMM,i} \ln(\phi_{TtMM}) \\ & + n_{TtMm,i} \ln(\phi_{TtMm}) + n_{Ttmm,i} \ln(\phi_{Ttmm}).\end{aligned}$$

The first two derivatives of the above equation with respect to δ are

$$\frac{\partial \ln L_i(\delta \,|\, nG)}{\partial \delta} = \frac{n_{TTMM,i}}{\delta + f_T f_M} - \frac{n_{TTmm,i}}{-\delta + f_T f_m} - \frac{n_{TtMM,i}}{-\delta + f_t f_M} + \frac{n_{Ttmm,i}}{\delta + f_t f_m}$$

$$\frac{\partial^2 \ln L_i(\delta \,|\, nG)}{\partial \delta^2} = -\frac{n_{TTMM,i}}{(\delta + \hat{f}_T \hat{f}_M)^2} - \frac{n_{TTmm,i}}{(-\delta + \hat{f}_T \hat{f}_m)^2} - \frac{n_{TtMM,i}}{(-\delta + \hat{f}_t \hat{f}_M)^2} - \frac{n_{Ttmm,i}}{(\delta + \hat{f}_t \hat{f}_m)^2}.$$

Note that counts of heterozygous offspring for the marker *M/m* are not used and, therefore, do not provide information for estimating disequilibrium.

Sire Double Heterozygote

Let the sire have genotype *TtMm* at two SNPs, *T/t*, and *M/m* and linkage phase (*TM/tm*). As before, $n_{j,i}$ are the genotype counts from offspring from the *i*th sire family (*j* = *TTMM*, *TTMm*, *TTmm*, *TtMM*, *TtMm*, *Ttmm*, *ttMM*, *ttMm*, and *ttmm*). The recombination fraction is *c*, which is assumed to be known without error. The likelihood equation for data of the *i*th half-sib family is

$$\begin{aligned}L_i(\delta, f_T, f_M \,|\, nG) = K & (\phi_{TTMM})^{n_{TTMM,i}} (\phi_{TTMm})^{n_{TTMm,i}} (\phi_{TTmm})^{n_{TTmm,i}} (\phi_{TtMM})^{n_{TtMM,i}} (\phi_{TtMm})^{n_{TtMm,i}} \\ & \times (\phi_{Ttmm})^{n_{Ttmm,i}} (\phi_{ttMM})^{n_{ttMM,i}} (\phi_{ttMm})^{n_{ttMm,i}} (\phi_{ttmm})^{n_{ttmm,i}},\end{aligned}$$

where $\phi_j$is the probability of the *j*th genotype as described in the text. In the reduced model, ignoring the constant and taking natural logarithm of the above expression gives

$$\begin{aligned}\ln L_i(\delta \,|\, nG) \approx\; & n_{TTMM,i} \ln(\phi_{TTMM}) + n_{TTMm,i} \ln(\phi_{TTMm}) + n_{TTmm,i} \ln(\phi_{TTmm}) + n_{TtMM,i} \ln(\phi_{TtMM}) \\ & + n_{TtMm,i} \ln(\phi_{TtMm}) + n_{Ttmm,i} \ln(\phi_{Ttmm}) + n_{ttMM,i} \ln(\phi_{ttMM}) \\ & + n_{ttMm,i} \ln(\phi_{ttMm}) + n_{ttmm,i} \ln(\phi_{ttmm}).\end{aligned}$$

The first two derivatives of the above equation with respect to δ are

$$\begin{aligned}\frac{\partial \ln L_i(\delta | nG)}{\partial \delta} = & \frac{n_{TTMM,i}}{(\delta + f_T f_M)} - \frac{(1-2c) n_{TTMm,i}}{(1-c)(-\delta + f_T f_m) + c(\delta + f_T f_M)} - \frac{n_{TTmm,i}}{-\delta + f_T f_m} \\ & - \frac{(1-2c) n_{TtMM,i}}{(1-c)(-\delta + f_t f_M) + c(\delta + f_T f_M)} \\ & + \frac{2(1-2c) n_{TtMm,i}}{(1-c)(2\delta + f_T f_M + f_t f_m) + c(-2\delta + f_T f_m + f_t f_M)} \\ & - \frac{(1-2c) n_{Ttmm,i}}{(1-c)(-\delta + f_T f_m) + c(\delta + f_t f_m)} - \frac{n_{ttMM,i}}{(-\delta + f_t f_M)} \\ & - \frac{(1-2c) n_{ttMm,i}}{(1-c)(-\delta + f_t f_M) + c(\delta + f_t f_m)} \\ & + \frac{n_{ttmm,i}}{(\delta + f_t f_m)}\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \ln L_i(\delta|nG)}{\partial\delta^2} = & -\frac{n_{TTMM,i}}{[\delta+f_Tf_M]^2}-\frac{(1-2c)^2 n_{TTMm,i}}{[(1-c)(-\delta+f_Tf_m)+c(\delta+f_Tf_M)]^2}-\frac{n_{TTmm,i}}{[-\delta+f_Tf_m]^2}\\
& -\frac{(1-2c)^2 n_{TtMM,i}}{[(1-c)(-\delta+f_tf_M)+c(\delta+f_Tf_M)]^2}\\
& -\frac{4(1-2c)^2 n_{TtMm,i}}{[(1-c)(2\delta+f_Tf_M+f_tf_m)+c(-2\delta+f_Tf_m+f_tf_M)]^2}\\
& -\frac{(1-2c)^2 n_{Ttmm,i}}{[(1-c)(-\delta+f_Tf_m)+c(\delta+f_tf_m)]^2}-\frac{n_{ttMM,i}}{[-\delta+f_tf_M]^2}\\
& -\frac{(1-2c)^2 n_{ttMm,i}}{[(1-c)(-\delta+f_tf_M)+c(\delta+f_tf_m)]^2}\\
& -\frac{n_{ttmm,i}}{[(\delta+f_tf_m)]^2}.
\end{aligned}
$$

Appendix B

Reduced Model for Estimating LD in a Homo-heterozygote Sire Family

In a reduced model, allele frequencies are not estimated simultaneously with haplotype frequencies but are assumed to be known. This likelihood equation assuming known allele frequencies in the dam population is

$$
\begin{aligned}
L_i(\hat{\delta}\,|f_M,nG) = K(\phi_{TTMM})^{n_{TTMM,i}}(\phi_{TTMm})^{n_{TTMm,i}}(\phi_{TTmm})^{n_{TTmm,i}}(\phi_{TtMM})^{n_{TtMM,i}}\\
\times\ (\phi_{TtMm})^{n_{TtMm,i}}(\phi_{Ttmm})^{n_{Ttmm,i}}.
\end{aligned}
$$

Allele frequencies can be estimated using the same data following Gomez-Raya (2001) by

$$\hat{f}_T = \frac{1}{N_i}\,(n_{TTMM,i}+n_{TTMm,i}+n_{TTmm,i})$$

$$\hat{f}_M = \frac{n_{TTMM,i}+n_{TtMM,i}}{n_{TTMM,i}+n_{TtMM,i}+n_{TTmm,i}+n_{Ttmm,i}}.$$

An explicit solution is obtained for the haplotype frequency of TM after rearranging Equation 2 (text):

$$\hat{f}_{TM} = \frac{\hat{f}_T\ n_{TTMM,i}}{N_i\hat{f}_T - n_{TTMm,i}}.$$

The disequilibrium is estimated after substituting $\hat{f}_T$, $\hat{f}_M$, and $\hat{f}_{TM}$ into by $\hat{\delta} = \hat{f}_{TM} - \hat{f}_T\hat{f}_M$.

Reduced Model for Estimating LD in a Double Heterozygote Sire

Following Gomez-Raya (2001), allele frequencies of M and T are estimated from the same data by

$$\hat{f}_M = \left(\frac{n_{TTMM,i}+n_{TtMM,i}+n_{ttMM,i}}{n_{TTMM,i}+n_{TtMM,i}+n_{ttMM,i}+n_{TTmm,i}+n_{Ttmm,i}+n_{ttmm,i}}\right)$$

$$\hat{f}_T = \left(\frac{n_{TTMM,i}+n_{TTMm,i}+n_{TTmm,i}}{n_{TTMM,i}+n_{TtMM,i}+n_{ttMM,i}+n_{TTmm,i}+n_{Ttmm,i}+n_{ttmm,i}}\right).$$

The maximum likelihood equation assuming known allele frequencies is

$$
\begin{aligned}
L_i(\hat{\delta}\,|f_T,f_M,nG) = {} & K(\phi_{TTMM})^{n_{TTMM,i}}(\phi_{TTMm})^{n_{TTMm,i}}(\phi_{TTmm})^{n_{TTmm,i}}(\phi_{TtMM})^{n_{TtMM,i}} \\
& \times\ (\phi_{TtMm})^{n_{TtMm,i}}(\phi_{Ttmm})^{n_{Ttmm,i}}(\phi_{ttMM})^{n_{ttMM,i}}(\phi_{ttMm})^{n_{ttMm,i}}(\phi_{ttmm})^{n_{ttmm,i}},
\end{aligned}
$$

which can be solved by making use of the EM algorithm as described in Equation 4 in the text. For fully linked SNPs (*e.g.*, contiguous SNPs in high density arrays) the recombination fraction between the SNPs is 0, and Equation 4 reduces to

$$\hat{f}^{\,i}_{TM} = \frac{1}{N_i}\left(n_{TTMM,i} + \frac{\hat{f}^{\,i}_{TM} n_{TtMm,i}}{\hat{f}^{\,i}_{TM} + \hat{f}_{tm}}\right).$$

After substituting $\hat{f}_{tm}$ by its value $\hat{f}_{tm} = \hat{f}^{\,i}_{TM} - \hat{f}_T\hat{f}_M + \hat{f}_t\hat{f}_m$ (recall $\hat{\delta} = \hat{f}^{\,i}_{TM} - \hat{f}_T\hat{f}_M$) and rearranging the above equation, a quadratic is obtained,

$$a\left(\hat{f}^{\,i}_{TM}\right)^2 + b\hat{f}^{\,i}{}_{TM} + z = 0,$$

where $a = 2N_i$, $b = N_i(1-\hat{f}_M-\hat{f}_T)-2n_{TTMM,i}-n_{TtMm,i}$, and $z = -(1-\hat{f}_M-\hat{f}_T)n_{TTMM,i}$. This is a conventional second-order polynomial with a real solution between 0 and 1. Note that $\hat{f}_t\hat{f}_m = 1+\hat{f}_{TM}-\hat{f}_T-\hat{f}_M$ so $(1-\hat{f}_M-\hat{f}_T) = -\hat{f}_T\hat{f}_M+\hat{f}_t\hat{f}_m$.