The Role of Pedigree Information in Combined Linkage Disequilibrium and Linkage Mapping of Quantitative Trait Loci in a General Complex Pedigree

S. H. Lee1 and J. H. J. Van der Werf

School of Rural Science and Agriculture, University of New England, New South Wales 2351, Australia

Manuscript received July 2, 2004
Accepted for publication September 20, 2004

ABSTRACT

Combined linkage disequilibrium and linkage (LDL) mapping can exploit historical as well as recent and observed recombinations in a recorded pedigree. We investigated the role of pedigree information in LDL mapping and the performance of LDL mapping in general complex pedigrees. We compared using complete and incomplete genotypic data, spanning 5 or 10 generations of known pedigree, and we used bi- or multiallelic markers that were positioned at 1- or 5-cM intervals. Analyses carried out with or without pedigree information were compared. Results were compared with linkage mapping in some of the data sets. Linkage mapping or LDL mapping with sparse marker spacing (~5 cM) gave a poorer mapping resolution without considering pedigree information compared to that with considering pedigree information. The difference was bigger in a pedigree of more generations. However, LDL mapping with closely linked markers (~1 cM) gave a much higher mapping resolution regardless of using pedigree information. This study shows that when marker spacing is dense and there is considerable linkage disequilibrium generated from historical recombinations between flanking markers and QTL, the loss of power due to ignoring pedigree information is negligible and mapping resolution is very high.

COMBINED linkage disequilibrium (LD) and linkage mapping has been implemented in a variance component approach, and analysis of simulated as well as real data has proven that genomic regions containing quantitative trait loci (QTL) could be narrowed down to within a few centimorgans (Meuwissen and Goddard 2000; Fanir et al. 2002; Grisart et al. 2002; Meuwissen et al. 2002; Lee and Van der Werf 2004). As LD mapping can take into account the great number of historical recombinations reflected by identity-by-descent (IBD) probability between haplotypes, positioning QTL can be very precise even with a relatively small number of animals (Meuwissen and Goddard 2000). Lee and Van der Werf (2004) investigated the efficiency of LD mapping in livestock using records of a few hundred progeny in a half-sib design and reported a high mapping resolution with confidence intervals up to just a few centimorgans.

However, it is noted that in general such a design may not always be available. Rather, a general pedigreed structure is commonly available and often used for mapping of QTL. A general pedigreed structure can span several generations with complex relationships, and ancestors’ genotypes are often unavailable and the number of genotyped progeny is not always enough to deduce parental genotypes (Haley 1999). However, a key process needed in (fine) mapping of QTL is to estimate IBD probabilities on the basis of LD and other available information including observed marker data and pedigree information. This could cause difficulties when predicting IBD in general pedigrees as genotype probabilities are hard to derive when pedigrees are complex and there are many missing genotypic data.

Although the general pedigree structure with missing data is very common, accuracy or efficiency of combined LD and linkage (LDL) mapping for this situation has not been reported. This is because there is no obvious method of pedigree analysis dealing with complex relationships and a large proportion of missing data for multiple closely linked markers. Exact methods for segregation analysis such as pedigree peeling (Elston and Stewart 1971; Cannings et al. 1978) or chromosome peeling (Lander and Green 1987) are well-known algorithms for estimating genotype probabilities on the basis of pedigree. However, the first method increases exponentially in computational complexity with the number of markers, and the latter becomes infeasible with a large proportion of missing data.

The locus sampler (Kong 1991; Heath 1997) uses a modification of the peeling algorithm and is much more efficient and flexible for multilocus problems in a complex pedigree. It computes genotype probabilities using all pedigree information subsequently at each locus, conditional on flanking loci. However, this algorithm requires that pedigrees must be peelable at each locus.
Alternatively, Markov chain Monte Carlo (MCMC) algorithms have been used to estimate genotype probabilities, updating latent variables at a single locus subsequently for each individual, which makes it possible to deal with complex relationships in a pedigree. Examples of algorithms are single-site genotypic samplers (Lange and Mattyyysse 1989; Sheehan et al. 1989) or single-site segregation indicator samplers (Thompson 1994). However, the irreducibility of the Markov chain is not easily guaranteed in complex pedigree structures and mixing problems also appear in using multiple marker loci (Thompson and Heath 1999; Cannings and Sheehan 2002). The meiosis Gibbs sampler (Thompson and Heath 1999) greatly improves mixing by updating segregation indicators jointly for all marker loci. However, this approach is limited to pedigrees of moderate size (<1000) since many MCMC cycles are required for an analysis.

Unfortunately, it is currently infeasible to use all available information for a large complex pedigree with sparse genotypic data (Cannings and Sheehan 2002). One may not be able to use complex relationships, and the loss of information can affect the accuracy of positioning QTL. However, LDL mapping, as mentioned earlier, can use historical recombinations. If mutation age is high, e.g., >100 generations, recorded pedigree of a few generations may contribute little to the information for positioning QTL. In this situation, it is worthwhile to investigate the decrease of accuracy of LDL mapping when ignoring pedigree information. Note that omitting some pedigree information makes analyses much easier and faster.

The aim of this study is to investigate the role of pedigree information in LDL mapping with a general complex pedigree. LDL mapping is carried out for a complex pedigree spanning 5 or 10 generations considering all pedigree information or considering the pedigree information only in the last generation (i.e., ancestors of all animals in the last generation are treated as being unrelated). The power advantage of LDL over linkage alone is shown. The efficiency of LDL mapping with various levels of LD in relation to pedigree information is investigated. In addition, it is shown how to efficiently integrate the case of a complex pedigree with incomplete genotypic data (genotypes are available only for the progeny in the last generation) using Gibbs sampling.

**Materials and Methods**

**Simulation Study:** A population of size \( N_t \) was simulated for 10 marker loci and a QTL for \( t \) generations to generate LD beyond recorded pedigree between QTL and flanking markers. In each generation, the number of male and female parents was \( N_t/2 \) and their alleles were inherited to descendants on the basis of Mendelian segregation using the gene-dropping method (MacCluer et al. 1986). Parents were randomly mated with a total of two offspring for each of \( N_t/2 \) mating pairs. For the QTL, unique numbers were assigned to the base population alleles. In generation \( t \), one of the base alleles surviving with a frequency of >0.1 and <0.9 was randomly chosen and treated as favorable with effect \( \alpha \) compared to other QTL alleles. From generation \( t + 1 \) to generation \( t + 4 \) or \( t + 9 \), pedigree was recorded and polygenic values were simulated. The recorded pedigree had complex relationships between individuals because of random mating and selection. The base parameter for population size was \( N_t = 100 \) (\( t = 100 \)), with \( N_t = 200 \) (\( t = 200 \)) or \( N_t = 800 \) (\( t = 800 \)) as alternative values.

In the multiallelic marker model (e.g., microsatellites), the number of alleles assumed in each marker locus was four and base allele frequencies were all at 0.25. In the biallelic marker model (e.g., single-nucleotide polymorphisms), the number of alleles was two and starting allele frequencies were 0.5. The marker alleles were mutated at a rate of 4 \( \times 10^{-4} \) per generation (Dallas 1992; Weber and Wong 1993; Ellegren 1995). A mutated locus was changed between the two existing alleles for biallelic markers whereas a new allele was added for multiallelic markers.

The role of pedigree information was investigated in linkage mapping and LDL mapping, respectively. Analyses were carried out for a complex pedigree spanning either 5 generations (generations \( t \rightarrow t + 4 \)) or 10 generations (\( t \rightarrow t + 9 \)). IBD probabilities were estimated either considering all relationships across the recorded pedigree or considering only relationships in the last generation (i.e., the parents of the last generation were treated as unrelated).

For a fair comparison between results, phenotypic values were available only for 100 individuals in the last generation in all cases. Phenotypic values for individuals were simulated as

\[
y = \mu + \alpha + u + e.
\]

The population mean (\( \mu \)) was 100, values for polygenic effects (\( u \)) were drawn from \( N(0, \sigma_u^2) \) with \( \sigma_u^2 = 25 \) (\( \lambda \) is numerator relationships among individuals calculated since generation \( t + 1 \)), and values for residual (\( e \)) were from \( N(0, \sigma_e^2) \) with \( \sigma_e^2 = 50 \). The favorable QTL allele had an additive value of 7 (\( \alpha_e = 0 \) and \( \alpha_q = 7 \)) and variance of QTL (\( V_Q \)) ranged from 8.8 to 24.5, depending on QTL allele frequency (0.1 \( \sim \) 0.9 in this study), with \( V_Q = 2p(1 - p)\alpha^2 \) (Falconer and Mackay 1996). The frequency of 0.1 \( \sim \) 0.9 may be reasonable for a QTL that was previously detected by linkage mapping (Meuwissen and Goddard 2000), and loci with more extreme allelic frequencies would contribute little to genetic variation and thus would not be appealing candidates for gene-mapping studies (unless in the case of a rare disease).

**Analysis of Data Sets:** Mixed linear model for detecting QTL: A vector of phenotypic observations simulated from (1) can be modeled as

\[
y = X\beta + Zu + Zq + e,
\]

where \( y \) is a vector of phenotypic observations on the trait of interest, \( \beta \) is a vector of fixed effects, \( u \) is a vector of random polygenic effects for each individual, \( q \) is a vector of random effects due to QTL, and \( e \) are residuals. The random effects \( u \), \( q \), and \( e \) are assumed to be normally distributed with mean zero and variance \( \sigma_u^2 \), \( \sigma_q^2 \), and \( \sigma_e^2 \). \( X \) and \( Z \) are incidence matrices for the effects in \( \beta \), and \( u \) and \( q \), respectively. From (2), the associated variance-covariance matrix of all observations (\( V \)) for a given pedigree and marker genotype set is modeled as

\[
V = ZAZ^\top \sigma_e^2 + Z GRM Z^\top \sigma_q^2 + R,
\]

where \( A \) is the numerator relationship matrix based on additive genetic relationships, GRM is the genotype relationship matrix whose elements are IBD probabilities between individuals computed for a putative QTL position and conditional on marker information, and \( R = k^{\top} k \) (\( I \) is an identity matrix).
Since the value of IBD probabilities between animals depends on the putative QTL position within a tested chromosome region, a number of different GRMs are generated. In this study, we used 10 markers and tested QTL locations at the middle point of each marker bracket; therefore, nine GRMs were generated. Variance components for model parameters ($\sigma^2_I$, $\sigma^2_a$, and $\sigma^2_e$) were estimated with each GRM, using average information restricted maximum likelihood (AIEML; Johnson and Thompson 1995). The maximum values of the log likelihood for the different QTL locations were compared.

**Gibbs sampling scheme for deriving IBD probabilities for GRM:**
With marker genotypes available for all individuals in the recorded pedigree, i.e., parental and progeny’s genotypes are known in each nuclear family, the correct linkage phase can often be assigned with a high certainty (Meuwissen and Goddard 2000). Pong-Wong et al. (2001) reported that if >10 biallelic markers are used, the proportion of individuals having at least one informative marker locus to assign correct phase is >90%. Therefore, the true set of haplotypes is close to the optimal set of haplotypes estimated with the highest likelihood among all candidate sets of haplotypes. Therefore, we used true haplotypes when using complete genotypic data (see also Discussion and Table 2). However, when few genotypes are recorded on parents or further ancestors, there are many uninformative markers and many more possible sets of haplotypes that can have similar likelihoods. For this case of using incomplete genotypic data, all possible states of segregation need to be taken into account in an analysis. The meiosis Gibbs sampler and the locus Gibbs sampler are considered to be suitable for this problem. Both samplers are supposed to give the same result if they work properly. However, the former is more efficient for sparse genotypic data (many missing genotypes) while the latter is more efficient for dense genotypic data (few missing genotypes; Heath 2003). As both use segregation indicators as latent variables, the distribution of segregation indicators is described first, followed by a description of the sampling procedure.

**Distribution of segregation indicators given observed marker data:**
One realization of segregation indicators ($S$) in a pedigree can be expressed in an $M \times L$ matrix whose elements are 0 or 1. If the gene in the $m$th meiosis at the $l$th locus receives the paternal parental allele, the segregation indicator $S_m = 0$, and $S_m = 1$ for the maternal parental allele. The maximum number of possible configurations for $S$ is $2^{|M| \times L}$, where none of the pedigree members are genotyped. The probability of $S$ given observed marker data is

$$pr(S|G) = \frac{pr(G|S)pr(S)}{\sum pr(G|S)pr(S)},$$

where $G$ represents the observed marker data, $pr(S)$ is prior probability of the segregation indicators, $pr(G|S)$ is the probability of the observed marker data given $S$, and the denominator is summed over the probabilities of all possible configurations of $S$. Since the computation of the denominator is not feasible in general pedigrees, a Gibbs sampling scheme is required to obtain the posterior distribution of the segregation indicators.

**Sampling segregation indicators by using the locus Gibbs sampler:**
When pedigree is peelable at each marker locus (e.g., with few missing genotypic data with simple relationships), the locus Gibbs sampler (Kong 1991; Heath 1997) can be used to obtain segregation indicators and reconstructed haplotypes in each sampling round. This method jointly updates the inheritance of all genes for individuals (or meioses) in a single locus. Genotypic configurations are sampled from the distribution of genotype probabilities estimated by recursive peeling (Cannings et al. 1978) at a locus conditional on the segregation indicators of the flanking marker loci (Kong 1991). The sampled genotypic configurations at the locus are converted to segregation indicators for sampling genotypic configuration at the next marker locus. Therefore ordered genotypes and segregation indicators are sampled across all individuals locus by locus. The locus sampler has been implemented in MCMC linkage software “LOKI” (Heath 1997). We obtained one configuration of the segregation indicators using LOKI in each sampling round when using complete genotypic data or incomplete genotypic data without parental relationships.

**Sampling segregation indicators by using the meiosis Gibbs sampler:**
This algorithm makes joint updates for the inheritance of all genes at linked loci, in a single meiosis. For example, for the $m$th meiosis, segregation indicators at all loci can be sampled using a forward-backward algorithm (Thompson and Heath 1999), according to all possible segregation states for the $m$th meiosis, conditional on the other meioses (see Appendix). This sampler was used for incomplete genotypic data with all available relationships in a pedigree.

**Haplotype reconstruction:**
Since LD-based IBD probabilities are derived from haplotype similarity between unrelated base animals, ordered genotypes for base animals are required to reconstruct haplotypes. The ordered genotypes can be sampled on the basis of the distribution of compatible allele assignments to founder genes that are consistent with the sampled segregation indicators (Sobel and Lange 1996). When this procedure is implemented for multiple marker loci, haplotypes for base animals are established. This procedure is performed in each sampling round.

**Initial legal configuration for the Markov chain:**
The meiosis Gibbs sampler requires a starting configuration, consistent with observed marker data, which is essential to start the Markov chain. Heath (1998) used a combined method of pedigree peeling and a genotype elimination algorithm to sample genotype configurations, consistent with the observed marker data. These can be converted to a legal configuration of segregation indicators. However, in the method, each locus should be peelable, which is not always guaranteed in the case of a complex pedigree with many missing genotypic data. Instead of using a peeling-based algorithm, the genotype elimination through inheritance constraint algorithm suggested by Henshall et al. (2001) overcomes the problem of sampling for a complex pedigree with genotypic data at a single locus. After the algorithm samples segregation indicators for each locus independently, the Gibbs procedure obtains the desired conditional distribution, taking into account the linkage between markers.

**IBD probabilities based on LD and linkage information:**
As elements of GRM, IBD probabilities between all members are estimated on the basis of LD and linkage in each sampling round. Sampled haplotypes for base animals are used to estimate LD-based IBD probabilities between unrelated base animals, using the method of Meuwissen and Goddard (2000; 2001). Sampled segregation indicators at multiple loci for descendants are used to recursively estimate IBD probabilities between relatives given LD-based IBD probabilities of base animals, using the method of Wang et al. (1995).

**A brief summary of the sampling procedure:**

Do 1 ~ $N$ cycles:
- Sample segregation indicators for all meioses at all marker loci.
- Sample haplotypes for base animals.
- Estimate IBD based on sampled haplotypes and segregation indicators.

Construct GRM based on IBD.
End do:
- Average GRM.
Joint updates for the whole meiosis or the whole locus sampler result in better mixing properties and the process to be much more reliable than that of a single-site Gibbs sampler (Thompson and Heath 1999). In addition, a sampler updating segregation indicators rather than genotypic configuration has a much smaller state space (Thompson 1994). Therefore, convergence of IBD probabilities could be reached quickly. The Gibbs sampler was carried out for 5000 cycles, discarding the first 1000 cycles. In every tenth sampling round, elements of GRM were estimated and stored. They were averaged after the final sampling round. In the sampling procedure, allele frequency at every marker locus was assumed equal (flat prior distribution). In each sampling round, it took ~1.5 sec to obtain haplotypes and segregation indicators and ~2.5 sec to estimate IBD probabilities in the pedigree spanning five generations (CPU: 2.4 GHz).

RESULTS

Complete genotypic data: Distribution of the estimated position that deviated from the true QTL position in 100 replicates is illustrated in Table 1 for the case when all genotypic information is available on all individuals in the recorded pedigree (generation 100 or 109).

Combined LD and linkage mapping: In LDL mapping, the pattern of this distribution is very similar between the analysis considering all pedigree information (PED) and the analysis ignoring relationships between individuals until the last generation (NPED) for both 5 generations and 10 generations of pedigree. With the LDL method, the most frequently estimated position is in the correct marker bracket and >55% of replicates position the QTL within 3 cM of the true position with biallelic markers and >75% of replicates do the same with multiallelic markers.

Linkage mapping alone: With linkage mapping, the effect of using pedigree information is large as the pattern of the distribution is quite different between PED and NPED. The difference is largest for the pedigree spanning 10 generations. Linkage mapping with biallelic markers and the pedigree spanning 5 generations most frequently estimates the QTL position at the boundaries of the tested region with both PED and NPED. The difference is largest for the pedigree spanning 10 generations. Linkage mapping with biallelic markers and the pedigree spanning 5 generations most frequently estimates the QTL position at the boundaries of the tested region with both PED and NPED, showing there is little information to position the QTL. A similarly poor mapping resolution is obtained in the analysis for the pedigree spanning 10 generations with NPED, but the QTL is more frequently positioned on the true position with PED; ~45% of replicates position the QTL within 3 cM of the true position (Table 1). In the multiallelic marker model with the pedigree spanning 5 genera-
The averaged likelihood ratio: The value of the likelihood ratio (LR) across the genomic region, averaged over replicates with the multiallelic marker model, is plotted in Figure 1. The value of LR is almost flat in linkage mapping with NPED in the pedigree of 5 generations (Figure 1A). Although the overall LR is higher in linkage mapping with PED, the difference between the highest LR and lowest LR is small, showing that not much information for positioning the QTL is provided by the pedigree spanning 5 generations. In the pedigree spanning 10 generations (Figure 1B), it is shown that linkage mapping with NPED has no power to detect the QTL, but linkage mapping with PED gives a considerably higher LR for the correct QTL position and there is an obvious peak, indicating much more information for positioning the QTL is provided when considering all pedigree information compared with considering relationships only in the last generation. This additional information from pedigree must account for the higher frequency of estimated QTL position in the correct marker bracket in linkage mapping with PED compared with NPED in the pedigree of 10 generations (Table 1).

With LDL mapping, the LR curve is clearly peaked and highest at the correct QTL position, showing LDL mapping to be much more powerful than linkage mapping. In both pedigrees spanning 5 and 10 generations, it is also shown that the curves of LR in LDL mapping with PED and NPED are similar, indicating that pedigree information is not so critical in LDL mapping (Figure 1).

The pattern of LR curves with biallelic markers is very similar to that with multiallelic markers although overall LR values are lower (results not shown).

Incomplete genotypic data: With incomplete genotypic data, genotypes were available only for the progeny in the last generation and all ancestral genotypes were missing. Figure 2 presents the distribution of estimated QTL position deviated from the true location and the value of LR averaged over replicates when LDL mapping is carried out with multiallelic markers and a pedigree spanning five generations. The frequency of estimated position near the correct QTL position is reasonably high; >75% of replicates position the QTL within 3 cM of the true position. This result is similar to that with complete genotypic data. Although overall values of LR are reduced compared to that with complete genotypic data, there is an obvious peak, showing that there is sufficient information to locate the QTL at the correct position. As with complete genotypic data, the difference between PED and NPED is not significant.

The results with biallelic markers are similar to those with multiallelic markers in that the distribution of estimated QTL position is similar to that with complete genotypic data, and the LR curve shows an obvious peak at the true QTL position (results not shown).

The effect of marker density and past effective size in relation to LD: Lower LD could arise from either using a lower marker density or a higher effective population size. Figure 3 shows the pattern of LR values with LDL analyses averaged over replicates with 10 multiallelic markers positioned at 5-cM intervals. The value of PED increases slightly with 5 generations of pedigree and the increase is more significant for 10 generations of pedigree, compared to NPED. The value of LR with NPED in each position is similar between 5 generations and 10 generations. When comparing with a marker spacing of 1 cM, the LR values are much lower and the LR curve is flatter, indicating that LDL mapping with a sparse marker spacing does not give sufficient resolution (Figure 3).

Figure 4 shows LR values averaged over replicates when \( N_e = 100, 200, \) or 800 for 100, 200, or 800 generations was simulated. The LR values were substantially decreased with higher values of \( N_e \) and results for PED and NPED were similar for all values of \( N_e \).
The relationship between LD and the length of a chromosomal region that is IBD can be described as

\[ LD = \frac{1}{1 + 4N_e c - 2e - 2N_e c^2 + e^2} \]  

(Sved and Feldman 1973), where \( N_e \) is past effective size and \( e \) is the recombination rate of the chromosomal region. LD is defined here as the probability of the chromosomal region being IBD when two random haplotypes are taken from the population. The observed value of LD based on simulated data agreed with the expected value from (5). The averaged value of LD observed over 100 replicates was 0.19 ± 0.07 (expected value is 0.2) for a chromosome segment of 1 cm with \( N_e = 100 \). The value decreased to 0.04 ± 0.01 (expected value is 0.05) for a segment of 5 cm with \( N_e = 100 \). Note that the LR values were much lower with 5-cM marker intervals than with 1-cM intervals (Figure 3). The value also decreased with higher value of \( N_e \), e.g., 0.11 ± 0.04 (expected value is 0.11) and 0.03 ± 0.01 (expected value is 0.03) for a 1-cM segment with \( N_e = 200 \) and 800, respectively. Note that the LR values were much lower for \( N_e = 800 \) than for \( N_e = 100 \) (Figure 4). Given these results, it is clearly shown that the levels of LD are closely related to the efficiency of LDL mapping. With a lower LD, pedigree information became useful if (and only if) the tested region was wide enough to be sufficiently broken up for 5 or 10 generations. This was the reason that with a lower LD, LDL mapping with PED gave higher accuracy than that with NPED in the marker spacing of 5 cm (Figure 3); however, the difference between PED and NPED was very small with a marker spacing of 1 cm (Figure 4).

**Alternative QTL location:** To test the performance of LDL mapping for QTL not centered in the studied region, a QTL located at one-third of the tested region was investigated. Figure 5 shows the distribution of estimated QTL position and the value of LR averaged over replicates when LDL mapping is carried out with or without pedigree information. As in the case of a centered QTL, the frequency of estimated position as well as the LR is highest at the correct QTL position and the LR curve is fairly peaked.

**DISCUSSION**

While linkage mapping could greatly benefit from additional pedigree information, knowledge of relationships between ancestors was not critical in LDL mapping with closely linked markers (\(~1\) cm). The additional information generated from the recorded relationship in a pedigree was very small compared to the LD information generated from the historical population beyond recorded pedigree. However, when using a lower
marker density (~5 cM), the degree of LD between markers is much decreased and the recorded relationships become more informative. In such cases, it is desirable to use available pedigree information whenever possible.

When parental (ancestral) genotypes are absent, the loss of information for positioning the QTL is small as shown by the limited reduction of LR values in LDL mapping with a 1-cM marker spacing; *i.e.*, the most frequent estimated position was on the true QTL and the LR curve was fairly peaked around the correct QTL position. This implies that parental haplotypes can be reasonably well reconstructed from relatively few genotypes in the last generation in a general pedigree (approximately two to three progeny per family in our study). These results agree with those of Abecasis *et al.* (2000), who reported that in the case that parental genotypes are missing, the loss of power for detecting QTL is negligible when at least four genotyped progeny per family are available. In our study, we used 10 markers, which might help to assign linkage phase more correctly.

With complete genotypic data, we used true haplotypes under the assumption that linkage phase is assigned with high certainty when parental and progeny genotypes are known. Results from true haplotypes and those from sampled haplotypes were compared, using complete genotypic data spanning five generations with a marker spacing of 1 or 5 cM. Table 2 shows high correlations between parameters estimated with the true haplotypes and those with sampled haplotypes for 20 replicates. For all variance components, the correlation between the two results is close to 1. Also for LR values at the true position, the results agreed very well. For the estimated QTL position based on sampled haplotypes, 90–100% of replicates position the QTL within a distance of one marker interval from the estimated QTL position using true haplotypes for both 1- and 5-cM marker spacing. These results show that using true haplotypes is very similar to using sampled haplotypes in complete genotypic data.

Similar results were shown by Morris *et al.* (2004) in that the efficiency of fine mapping was not much reduced using a MCMC approach, compared with using the true haplotypes. The authors also showed that a MCMC approach that considered all possible sets of haplotypes was much more efficient than inferring a set of haplotypes based on maximum likelihood. This was probably due to the fact that the likelihood was flat with respect to different haplotype configurations because no pedigree was used in deriving these haplotypes. Inferring haplotypes in the case that parents’ and progeny’s genotypes and their relationships are fully known is expected to give better results, as was also shown by Pong-Wong *et al.* (2001).

Since the locus sampler is based on a modification of the peeling algorithm that considers all compatible states simultaneously in a single locus, it does not have...
TABLE 2

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_1^2$</th>
<th>$\sigma_2^2$</th>
<th>$\sigma_3^2$</th>
<th>LR on the true QTL</th>
<th>Position (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-cM PED</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>100</td>
</tr>
<tr>
<td>1-cM NPED</td>
<td>0.98</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>100</td>
</tr>
<tr>
<td>5-cM PED</td>
<td>0.98</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>90</td>
</tr>
<tr>
<td>5-cM NPED</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>90</td>
</tr>
</tbody>
</table>

Complete genotypic data spanning five generations with a marker spacing of 1 or 5 cM.

* Proportion of replicates with sampled haplotypes positioning the QTL within one marker interval of estimated QTL position with true haplotypes.

Because the number of genotyped progeny per parent was low enough not to constrain parental allelic types. Therefore, there were few or no noncommunicating classes in the Markov chain. Moreover, nearly identical results from two very different MCMC approaches proved the process to be reliable.

With sparse marker spacing (>5 cM) too wide to find LD information between QTL and flanking markers, it is necessary to consider all pedigree information as the accuracy with PED and NPED was much different especially in a deeper pedigree (Figure 3). The meiosis Gibbs sampler can be an efficient tool to deal with complex pedigree information in such cases. However, in some data structures where allelic types of founders in a recorded pedigree are fully constrained by direct (if founders were genotyped) or indirect observations (founders having a large number of genotyped progeny), the problem of reducibility in the meiosis sampler would occur. Block updating segregation indicators for a number of relatives simultaneously can be a way to increase irreducibility although the large number of relatives can make it infeasible to compute genotype probabilities (i.e., the number of segregation states is $2^{2n^2}$ with $n$ the number of individuals for which segregation indicators are updated jointly). A random-walk approach (Sobel and Lange 1996) could be applied to noncommunicating classes using the Metropolis algorithm (Metropolis et al. 1953). Further study is required to increase irreducibility in the meiosis sampler.

In linkage mapping especially with biallelic markers, the QTL was estimated more frequently at the boundaries than at the center in the tested region and the distribution of estimated positions did not seem normal (Table 1). If the length of the tested region was increased from 9 to 19 cM with the same data, estimated QTL positions were more normally distributed. Figure 6 shows that the replicates positioning the QTL at the boundaries in the tested region of 9 cM were actually the sum of the estimates beyond this boundary. Probably the most likely position was outside the tested region of 9 cM and the QTL was estimated at the boundary of the tested region that was closest to that point. It is expected that the estimated positions would be completely normally distributed if the tested region is wide enough (>~100 cM). Note that the distribution of estimated

---

**Figure 6.** Different distribution of estimated QTL position in linkage mapping with biallelic markers when the tested region is extended. Darkly shaded bar, tested region of 9 cM; lightly shaded bar, tested region of 19 cM.
position within \(-3 \text{ cM} \) or \(3 \text{ cM} \) from the true position is very similar for either length of the tested regions.

Similar results were observed in the study of Mueuwissen and Goddard (2000). When LD information was reduced, the frequency of the estimated position became higher at the boundaries. In the study of Sarry et al. (2002), when LD information content was low, the proportion of replicates positioned at the boundaries was much larger than that of those positioned in the center regions. However, when there was more information, the estimated positions were normally distributed in a small region.

Although the QTL effect was fixed, QTL heritability ranged from 0.11 to 0.25 because of different QTL allele frequencies. However, the performance of mapping is not substantially affected by QTL heritability unless QTL allele frequency is extreme (<0.2 or >0.8). The estimated position is similarly distributed near the correct position for all values of heritability >0.15 (result not shown). This is because the QTL effect was constant and the effect as proportion of phenotypic SD was nearly equal for all values of heritability (ranging from 0.7 to 0.76). As shown by Lee and Van der Werf (2004), the accuracy substantially decreased with a lower phenotypic SD (e.g., <0.45\(\sigma_p\)); however, as the number of records used for fine mapping increased, the accuracy became reasonably high. In this study, the number of records was only 100; therefore, the accuracy could easily be improved with a larger number of records.

Linkage disequilibrium between QTL and markers was generated from a large number of generations (in this study, \(N_e = 100, 200, \) or \(800 \) for 100, 200, or 800 generations). A more correct simulation would simulate a number of background genes with a mutation model during the population history. However, it is arbitrary when deciding the number of background genes and their effect. Instead, we simply derived normally distributed polygenic effects for the last generation only, assuming the founders in the recorded pedigree were unrelated. Since polygenic effects are not directly related to the efficiency of positioning QTL unless they are linked or there is epistasis, the assumption of unrelated founders when generating polygenic effects is not expected to affect the main results.

In this study, generally, LDL mapping was much more powerful than linkage mapping alone in positioning the QTL. If there were useful levels of LD (\(>0.1\)), pedigree information was not important. However, with lower levels of LD due to sparser marker spacing, the efficiency of LDL mapping decreased and pedigree information became more informative. With higher population effective size, LD also decreases and a denser marker spacing is needed for LDL mapping to be efficient. Pedigree information will generally be less informative if the size of the region considered is too small to allow sufficient recombinations during the recorded pedigree. Using the meiosis or the locus Gibbs sampler, complex relationship with missing genotypic data could be efficiently integrated for fine mapping of QTL.

We are thankful for helpful discussion with Bruce Tier and Miguel Perez-Enciso. Useful comments from reviewers are much appreciated. This study was supported by Australian Wool Innovation and a University of New England research assistance.

**LITERATURE CITED**


Communicating editor: C. Haley

APPENDIX

Forward-backward algorithm in the meiosis sampler: Following Thompson and Heath (1999), we describe how to jointly sample segregation indicators of all genes at linked loci, in a single meiosis.

Forward working: In the forward working, the cumulative probability (Q) for the segregation indicator Sm l, is computed, conditional on all meioses at marker loci up to and including marker locus l except the mth meiosis itself, which is updating at the current stage. The working order is from the first marker to the last marker (1 ~ L),

\[ Q(x) = \Pr(S_{m,l} = x \mid S_{m-1,l}, G_{l}, S_{m-1,r}, G_{r}) , \] (A1)

where \( x = 0 \) (being transmitted from paternal) or 1 (from maternal), \( S_{m-1,l} \) = all segregation indicators at locus \( l \) except the \( m \)th meiosis, \( G_{l} \) is the observed marker data at locus \( l \), \( S_{m-1,r} \) is all segregation indicators from locus 1 to locus \( l - 1 \) except the \( m \)th meiosis, and \( G_{r} \) is the observed marker data from locus 1 to \( l - 1 \).

The right-hand side in (A1) can be divided by two parts as

\[ Q(x) = \Pr(S_{m,l} = x \mid S_{m-1,l}, G_{l}) \Pr(S_{m,l} = x \mid S_{m-1,r}, G_{r}) . \] (A2)

The first part in the right-hand side in (A2) can be obtained, using Bayes theorem:

\[ \Pr(S_{m,l} = x \mid S_{m-1,l}, G_{l}) = \frac{\Pr(G_{l} \mid S_{m,l} = x, S_{m-1,l}) \Pr(S_{m,l} = x)}{\sum_{i=0}^{1} \Pr(G_{l} \mid S_{m,l} = x, S_{m-1,l}) \Pr(S_{m,l} = x)} . \] (A3)

\( \Pr(S_{m,l} = x) \) is a prior probability with a value of 0.5; therefore, (A3) can be simplified as

\[ \Pr(S_{m,l} = x \mid S_{m-1,l}, G_{l}) \propto \Pr(G_{l} \mid S_{m,l} = x, S_{m-1,l}) . \] (A4)

The second part of the right-hand side in (A2) can be computed using the cumulative probability of the previous locus and recombination rate between locus \( l \) and the previous locus \( l - 1 \),

\[ (S_{m,l} = x \mid S_{m-1,r}, G_{r}) \propto Q_{l-1}(x)(1 - \theta_{l-1}) + Q_{l-1}(1 - x)\theta_{l-1} , \] (A5)

where \( \theta_{l-1} \) is the recombination rate between the locus \( l \) and \( l - 1 \). Note that for the first marker locus, the right-hand side in (A5) is negligible (=1) because there is no previous marker.

From (A1), (A4), and (A5),

\[ Q_{l}(x) \propto \Pr(G_{l} \mid S_{m,l} = x, S_{m-1,l}) \{Q_{l-1}(x)(1 - \theta_{l-1}) + Q_{l-1}(1 - x)\theta_{l-1}\} . \] (A6)

The term \( \Pr(G_{l} \mid S_{m,l} = x, S_{m-1,l}) \) can be efficiently computed, using a descent graph (see Sobel and Lange 1996; Thompson and Heath 1999; Bureau 2001). When forward working is completed, we have the cumulative probability of the segregation indicator for the last locus, which takes into account all possible segregation states for the \( m \)th meiosis at locus \( l \), conditional on all observed marker data and segregation states for all other meioses, that is,
is negligible; therefore, previous locus can be sampled conditional on the already sampled marker locus ($S_{i+1} \sim S_{i}$) and using the cumulative probability for locus $l$ that was computed in the forward working. The sampling order is the second last locus to the first locus ($L - 1 \sim 1$):

$$
\text{pr}(S_{m,l} = x | S_{\delta_{i-1}}, G_i, S_{\delta_{i-1}}, G_{i'}), S_{\delta_{i-2}}, \ldots, S_{m,l}) \propto Q_i(x) | S_{m,l} = x | 0_1 + (1 - |S_{m,l} - x|) (1 - \theta_i).
$$

(A7)

**Backward sampling:** In backward sampling, the segregation indicator $S_{m,l}$ is sampled conditional on the already sampled marker locus ($S_{i+1} \sim S_{m,l}$) and using the cumulative probability for locus $l$ that was computed in the forward working. The sampling order is the second last locus to the first locus ($L - 1 \sim 1$):

**A numerical example:** Table A1 shows a simple pedigree with genotypic data for three markers (M1 $\sim$ M3). Table A2 shows one legal configuration of segregation indicators in the pedigree as sampled in the first round. Joint updates for the third meioses (paternal gametes for animal 4 in italic letters) using the forward-backward algorithm are shown as an example. It is assumed that each marker has four alleles and allele frequencies are equal (0.25), and the recombination rate between each marker pair is 0.1.

The first term in the right-hand side in (A6) can be estimated for each marker locus using a descent graph:

$$
\text{pr}(G_i | S_{3,1} = 0, S_{\delta_{i-1}}) = 0, \quad \text{pr}(G_i | S_{3,1} = 1, S_{\delta_{i-1}}) = 0.0039
$$

$$
\text{pr}(G_i | S_{3,2} = 0, S_{\delta_{i-2}}) = 0.016, \quad \text{pr}(G_i | S_{3,2} = 1, S_{\delta_{i-2}}) = 0.016
$$

$$
\text{pr}(G_i | S_{3,3} = 0, S_{\delta_{i-3}}) = 0.008, \quad \text{pr}(G_i | S_{3,3} = 1, S_{\delta_{i-3}}) = 0.016.
$$

For example, $\text{pr}(G_2 | S_{3,2} = 1, S_{\delta_{i-2}}) = 0.25 \times 0.25 \times 0.25 = 0.016$.

The second term in the right-hand side in (A6) can be easily obtained using the cumulative probability for the previous locus $l - 1$ and the recombination rate between locus $l$ and $l - 1$. For the first marker, the second term is negligible; therefore, $Q_i(0) = 0$ and $Q_i(1) = 1$. For the second marker,

$$
Q_i(0) = \frac{0.016(Q_i(0)(1 - \theta_i) + Q_i(1)\theta_i)}{0.016(Q_i(0)(1 - \theta_i) + Q_i(1)\theta_i) + 0.016(Q_i(1)(1 - \theta_i) + Q_i(0)\theta_i)} = 0.1
$$

$$
Q_i(1) = \frac{0.016(Q_i(1)(1 - \theta_i) + Q_i(0)\theta_i)}{0.016(Q_i(0)(1 - \theta_i) + Q_i(1)\theta_i) + 0.016(Q_i(1)(1 - \theta_i) + Q_i(0)\theta_i)} = 0.9.
$$

For the last marker,
From $Q_3$, either 0 or 1 can be sampled for $S_{3,3}$. Now let the value of 1 be sampled for the last locus.

From (A8) in the backward sampling,

$$\text{pr}(S_{3,2} = 0 | S_{\text{all}} = 3, 2^*, G_2, S_{3,3}) = 0.012, \quad \text{pr}(S_{3,2} = 1 | S_{\text{all}} = 3, 2^*, G_2, S_{3,3}) = 0.988.$$ 

Now, let the value of 1 be sampled for the second locus:

$$\text{pr}(S_{3,1} = 0 | S_{\text{all}} = 3, 1^*, G_1, S_{3,2}, S_{3,3}) = 0, \quad \text{pr}(S_{3,1} = 1 | S_{\text{all}} = 3, 1^*, G_1, S_{3,2}, S_{3,3}) = 1.$$ 

The value of 1 is sampled for the first locus.

Thus, new updated segregation indicators for the third meioses are 1, 1, and 1 for M1, M2, and M3, respectively. For the other meioses, the same method for joint updates can be used.