# FINE-MAPPING OF QUANTITATIVE TRAIT LOCI USING COMBINED LINKAGE DISEQUILIBRIUM AND LINKAGE WITH GENERAL PEDIGREES

**S.H. Lee and J.H.J. Van der Werf**

The Institute for Genetics and Bioinformatics, School of Rural Science and Agriculture, University of New England, Armidale, NSW2351

## SUMMARY

Linkage disequilibrium (LD) information from closely linked markers has been widely used to fine-map quantitative trait loci (QTL). Steps needed for fine-mapping of QTL include inferring inheritance state of markers (haplotyping), estimating identity by descent (IBD) at QTL, and implementing the covariance structure based on IBD into a statistical model. Appropriate pedigree-genotype analysis can estimate more accurate inheritance state and haplotype configuration, which results in more accurate IBD probabilities. A variance component approach can implement covariance coefficients based on IBD probabilities in a mixed linear model to fit additive, dominance and epistasis QTL effects. We have also implemented a reversible jump Markov chain Monte Carlo (MCMC) method for multiple QTL mapping to account for confounding effects between closely linked QTL. Simulation results have shown that this method can improve the precision of mapping QTL.

## INTRODUCTION

The use of genetic markers has made it possible to map quantitative trait loci (QTL) underlying phenotypic variation of complex traits. Interval mapping methods based on linkage information have positioned numerous QTL in crossbred or outbred populations within ~ 30 centimorgan (cM) confidence intervals (Andersson *et al.* 1994; Georges *et al.* 1995). Due to the relatively large confidence intervals, it has been infeasible to identify the causal genes.

Linkage disequilibrium (LD) from dense markers can be used to narrow down the candidate region for QTL. Many QTL studies have shown how to use LD information for decreasing the confidence intervals of the QTL region (Riquet *et al.* 1999; Meuwissen and Goddard 2000; 2001; Meuwissen *et al.* 2002).

The use of LD information requires haplotypes which can be inferred from genotypes and pedigree. Identity by descent (IBD) probabilities between founders of the pedigree at the QTL can be derived from the LD information. IBD probabilities between relatives within the pedigree at the QTL can be estimated from inheritance states (linkage information). IBD probabilities from combined LD and linkage can give useful information about the covariance structure of phenotypic variation of quantitative traits. Different covariance structures at each different putative QTL used in a mixed linear model (variance component approaches) can give important information about the position of the genuine QTL.

When there are multiple linked QTL, confounding effects may decrease the accuracy and precision in QTL mapping (Haley and Knott 1992). Moreover, there may be intra- and inter-locus allelic interaction between QTL, and the interaction should be as important as additive QTL effects in complex traits (Carlborg and Haley 2004). In this paper, we show that the simultaneous mapping of

multiple QTL can lead to more accurate estimates of locations and effects of QTL as well as their interactions.

**Pedigree-genotype analysis.** Pedigree-genotype analysis can infer inheritance states, genotype configurations or haplotype configurations. These latent variables give critical information for gene mapping. In linkage mapping, IBD coefficients between relatives can be estimated based on the inheritance states, i.e. pedigree-based IBD probabilities. In association mapping, IBD coefficients at the QTL between founders can be estimated based on the haplotype similarity, i.e. LD-based IBD probabilities (Meuwissen and Goddard 2000). The pedigree-genotype analysis is naturally using the information from recorded pedigree and marker information. Complex relationships and missing genotypes in a general pedigree result in a very large state space, too large for exact likelihood methods (Elston and Stewart 1971; Lander and Green 1987). Therefore, Markov chain Monte Carlo (MCMC) algorithms have been widely used for pedigree-genotype analysis.

**Gibbs samplers.** One MCMC approach, the single site Gibbs samplers sequentially sample genotypes for an individual at a single locus, conditional on genotypes for the other individuals and the other loci. This makes it possible to deal with a large state space generated from a general pedigree with missing genotypes (Sheehan et al. 1989; Thompson 1994). However, reducibility problems often occur, i.e. the Markov chain cannot reach some states (mixing problems) especially when using multiple marker loci (Thompson and Heath 1999; Canning and Sheehan 2002). By updating state variables jointly for all loci in a single meiosis (the meiosis Gibbs sampler, Thompson and Heath 1999), or for all individuals at a single loci (the locus Gibbs sampler, Heath 1997), the mixing property greatly improves. However, reducibility problems are still generated in the meiosis sampler when founder allelic types are determined by direct or indirect observations (Thompson and Heath 1999). In the locus sampler, the joint updating requires the exact likelihood methods at a single loci, which cannot deal with a general pedigree with missing genotypes.

**Random walk approach.**The random walk approach suggested by Sobel and Lange (1996) remedied the reducibility problems by taking multiple moves of the random walk which allow the chain to pass through illegal configurations of segregation indicators on its way between legal configurations of segregation indicators. However, illegal or less likely configurations are often proposed, which are mostly rejected by the Metropolis mechanism. Thus, the computational efficiency of the random walk approach is much less than that of the meiosis Gibbs sampler where updated variables are always accepted.

**Combined random walk and meiosis Gibbs sampler.** An algorithm has been developed to combine the merits from both the random walk approach and the meiosis Gibbs sampler (Lee et al. 2005). The meiosis sampler is used for all sites where updated variables are always accepted, therefore, the variables are more frequently updated at the same time (computational efficiency is high if there is no reducibility problem). If there are (potentially) reducible sites, the random walk approach is applied. Combining these two approaches gives a higher computational efficiency than the random walk approach alone. Besides that, joint updates of segregation indicators for all marker loci help mixing, and therefore improve accuracy. Further, reducibility problems in the meiosis sampler alone can be remedied with the random walk approach (see Lee et al. 2005).

**Combinatorial optimization algorithm for haplotyping.** Haplotyping describes the process of finding the most likely haplotype configuration given pedigree and genotypic data. To achieve this, optimization methods are used rather than sampling all possible haplotype configurations. Optimization methods are based on finding efficiently combinations of solutions that ate most likely,

hence combinatorial optimization. Examples of such algorithms are simulated annealing and evolutionary algorithm.

**Simulated annealing.** A widely used statistical approach for haplotype reconstruction is simulated annealing (SA) (Kirkpatrick et al. 1983) which has been implemented in the linkage software, SimWalk2 (Sobel and Lange 1996). SimWalk2 uses a random walk approach and an annealing process to find the optimal haplotypes. SimWalk2 constitutes a flexible and efficient algorithm for haplotyping and probably the only one used for a general complex pedigree with incomplete genotypes. However, it needs a very large number of sequential evaluations and it is not always guaranteed that the most likely solutions are found within the arbitrarily determined number of evaluations.

**Evolutionary algorithm**. Evolutionary algorithms (EA) (Holland 1975) constitute an efficient tool for combinatorial optimization problems. A number of parallel solutions are respectively updated (evolved) by changing the variables within each solution (EA-mutation), or recombining them from different solutions (EA-recombination), and the most favorable solutions are selected (EA-selection). Compared to SA, EA may be competitive in efficiently finding an optimal solution. The main advantage of EA is its potential to parallellise computations because the algorithm can be divided across multiple CPUs. This would substantially reduce computing time. The computational efficiency linearly increases with the number of CPU (result not shown). In addition, the search mechanism in EA can make it easier to diagnose convergence compared to that in SA.

**IBD estimation.** The covariance structure of additive, dominance or epistatic effects due to causal genes can be predicted from IBD probabilities. These covariance coefficients can be implemented in a mixed linear model to detect variation associated with QTL having additive, dominance or epistatic effects. IBD probabilities are estimated based on the pattern of inheritance states and haplotype configurations. To derive such patterns, two kinds of approaches can be used. One is to find an optimal haplotype configuration with the highest likelihood given observed data (e.g. using Evolutionary algorithm). IBD coefficients are then estimated based on the most likely haplotypes. The other is to apply a MCMC algorithm to surface all possible inheritance states and haplotype configurations based on the posterior distribution (e.g. Combined random walk and meiosis Gibbs sampler). IBD probabilities are estimated every MCMC cycle. Averaged IBD probabilities over all cycles would provide estimates based on the posterior distribution given observed data (Lee and Van der Werf 2006a; 2006b). This approach would give unbiased estimates especially with many missing genotypes although in that case many sampling rounds would be needed.

**LD information.** IBD coefficients are based on similarity of haplotypes unrelated through known pedigree. These patterns of similarity can be derived by using a genedropping method (MacCluer et al. 1986) or the coalescence method introduced by Meuwissen and Goddard (2000; 2001). An assumption of a mutation age of 100 generations and a past effective size of 100 can be used to estimate IBD coefficients as they are usually unknown. Results have been shown to be robust against such assumptions about mutation age and effective size (Meuwissen and Goddard 2000; Lee and Van der Werf 2004).

**Combined LD and linkage information.** Using the IBD probabilities between unrelated haplotypes, IBD probabilities between related haplotypes in the following generations can be estimated using known pedigree information, for example using the methods of Fernando and Grossman (1989). Therefore, IBD probabilities between all haplotypes can be estimated based on joint information from LD and linkage (e.g. Meuwissen et al. 2002; Lee and Van der Werf 2005).

**Estimation of relationship matrices.** There are four IBD probabilities between any pair of individuals i and j in the pedigree which are the probabilities of paternal or maternal QTL allele of individual i being IBD to the paternal or maternal allele of individual j, given marker genotypes (Liu *et al.* 2002), i.e. $pr(Q_i^x \equiv Q_j^x \mid g)$ where $x$ is paternal ($P$) or maternal ($M$) QTL allele for each individual. From the probabilities, the additive genetic relationship coefficient ($G$) between animals i and j at the QTL is,

$$[G]_{ij} = \frac{1}{2}\left[pr(Q_i^P \equiv Q_j^P \mid g) + pr(Q_i^P \equiv Q_j^M \mid g) + pr(Q_i^M \equiv Q_j^P \mid g) + pr(Q_i^M \equiv Q_j^M \mid g)\right]$$

The dominance relationship coefficient ($D$) between animals i and j at the QTL is,

$$[D]_{ij} = \left[pr(Q_i^P \equiv Q_j^P \mid g) \cdot pr(Q_i^M \equiv Q_j^M \mid g) + pr(Q_i^P \equiv Q_j^M \mid g) \cdot pr(Q_i^M \equiv Q_j^P \mid g)\right]$$

The Hadamard products of appropriate matrices can be used to fit epistatic terms in a model (Mitchell *et al.* 1997).

**Multiple QTL analysis with a variance component approach**

Variance component (VC) approaches have been widely used to detect existence of variation associated with quantitative trait loci (QTL) (Grignola *et al.* 1996; George *et al.* 2000). The idea behind the approaches is to obtain IBD coefficients between relatives for the QTL (see section IBD estimation), and maximize the likelihood of phenotypic data in a mixed linear model implementing these IBD coefficients at each putative QTL position. The QTL position can be estimated with maximum likelihood (ML) or restricted maximum likelihood (REML) at the location with the highest likelihood value across the chromosome. This idea has been extended to a fine-mapping method using linkage disequilibrium (LD) generated from closely linked markers (Meuwissen and Goddard 2000; 2001) where LD-based IBD coefficients are used.

**Mixed linear model** A vector of phenotypic observations is written as a linear function of fixed effects, a polygenic term representing the sum of other unidentified additive genetic effects, the additive and dominance effects due to $n$ QTL, epistatic interaction among the QTL, and residuals. The model can be written as,

$$y = X\beta + Zu + \sum_{i=1}^{n}(Za_i + Zd_i) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(Za_ia_j + Za_id_j + Zd_ia_j + Zd_id_j) + e \qquad (1)$$

where $y$ is a vector of $N_r$ observations on the trait of interest, $\beta$ is a vector of fixed effects, $u$ is a vector of $N_u$ random polygenic effects for each animal, $a_i$ and $d_i$ are a vector of $N_a$ additive and non-additive random effects due to the $i$th putative QTL, $a_ia_j$, $a_id_j$, $d_ia_j$ and $d_id_j$ are a vector of epistatic interactions between the $i$th and $j$th putative QTL, and $e$ are residuals. The random effects ($u$, $a_i$, $d_i$, $a_ia_j$, $a_id_j$, $d_ia_j$ and $d_id_j$ and $e$) are assumed to be normally distributed with mean zero and variance $A\sigma_u^2$, $G_i\sigma_{q_i}^2$, $D_i\sigma_{d_i}^2$, $G_iG_j\sigma_{a_ia_j}^2$, $G_iD_j\sigma_{a_id_j}^2$, $D_iG_j\sigma_{d_ia_j}^2$, $D_iD_j\sigma_{d_id_j}^2$, and $I\sigma_e^2$, where $A$ is a numerator relationship matrix, $G_i$ and $D_i$ is an additive genetic and dominance relationship matrix at the $i^{th}$ putative QTL position, $M_iM_j$ is the Hadamard product of the matrix $M_i$ and $M_j$, and $I$ is a $N_r$ - order identity matrix. $X$ and $Z$ are incidence matrices (for the effects $\beta$ and $u$, $a_i$, $d_i$, $a_ia_j$, $a_id_j$, $d_ia_j$ and $d_id_j$ respectively). The associated variance covariance matrix ($V$) of all observations given pedigree and marker genotypes is modeled as

$$V = ZAZ'\sigma_u^2 + \sum_{i=1}^{n}(ZG_iZ'\sigma_{a_i}^2 + ZD_iZ'\sigma_{d_i}^2) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(ZG_iG_jZ'\sigma_{a_ia_j}^2 + ZG_iD_jZ'\sigma_{a_id_j}^2 + ZD_iG_jZ'\sigma_{d_ia_j}^2 + ZD_iD_jZ'\sigma_{d_id_j}^2) + I\sigma_e^2$$

**VC approach with an empirical Bayesian method** As the number of QTL and their positions are unknown in (1) a decision is required in relations to which putative QTL positions should be chosen, and how many QTL should be simultaneously fitted in a model. Model selection strategies have been developed and applied for this problem in gene mapping (Sillanpaa and Corander 2002). Stepwise selection (Jansen 1993; Kao *et al.* 1999) is a standard model selection technique where the Akaike information criterion (Akaike 1969) or Bayesian information criterion (Schwarz 1978) can be used to correct the model likelihood for the number of parameters fitted. Randomized approaches, such as MCMC or genetic algorithms, have also been proposed to find an optimal model (Sillanpaa and Arjas 1998; Calborg *et al.* 2000)

Green (1995) proposed a reversible jump MCMC which allows the Markov chain to surface state space across different model dimensions according to the correct posterior distribution. This is a generalization of Metropolis-Hastings methods (Metropolis *et al.* 1953; Hastings 1970) dealing with model selection problems. This technique has been used in multiple QTL analysis to estimate the number of QTL and their positions in linkage mapping (Heath 1997; Sillanpaa and Arjas 1998). Lee and Van der Werf (2006a) proposed the use of a reversible jump MCMC in a variance component approach using combined LD and linkage information to simultaneously map multiple QTL with additive

The ratio of additive variance over phenotypic variance is ~ 0, 0.07, 0 and 0 for the first, second, third and fourth QTL. The ratio of dominance variance over phenotypic variance is close to 0 for all QTL. The ratio of epistatic variance over phenotypic variance is 0.16 for the pair of the first and second QTL, and 0.19 for the pair of the third and fourth QTL and dominance effects in a small region. In the process, the QTL model is firstly defined by the number of QTL and their positions which are sampled from a proposal distribution. In a second step, REML estimates for the model parameters are obtained for a given QTL model. The proposed variables and model parameters are accepted or rejected, according to the acceptance ratio derived from the proper posterior distribution across different model dimensions. Hence, a REML procedure is used nested within a Bayesian reversible jump MCMC. This is an empirical Bayesian approach.

The empirical Bayesian approach described above can be extended to simultaneously fit additive, dominance and epistasis as in (1). Figure 1 shows that the full model (1) gives a higher mapping resolution than any other reduced model when a complex interaction exist between two closely linked QTL.

**DISCUSSION**

Given pedigree and genotypic data, an appropriate analysis is required to obtain reliable results in fine-mapping of QTL. As a preliminary analysis, Mendelian inconsistency generated from pedigree or genotype error should be corrected (e.g. data cleaning). Given the cleaned data, appropriate pedigree-genotype analysis can estimate accurate inheritance states and haplotype configuration. This results in reliable IBD estimation. Covariance coefficients based on the IBD probabilities can then be implemented in the VC approach to fit additive, dominance and epistatic QTL effects. Shade and confounding effects between closely linked QTL can be corrected using the multiple QTL analysis with the reversible jump MCMC. This process would give an accurate and precise fine-mapping of QTL.

For whole genome scan with a dense marker map (e.g. > ~ 10,000 SNP), it may not be feasible to use the VC approach for such large number of markers. However, rapid development of computer processor and parallel computing will make it possible for the VC approach to deal with such large data in the near future.

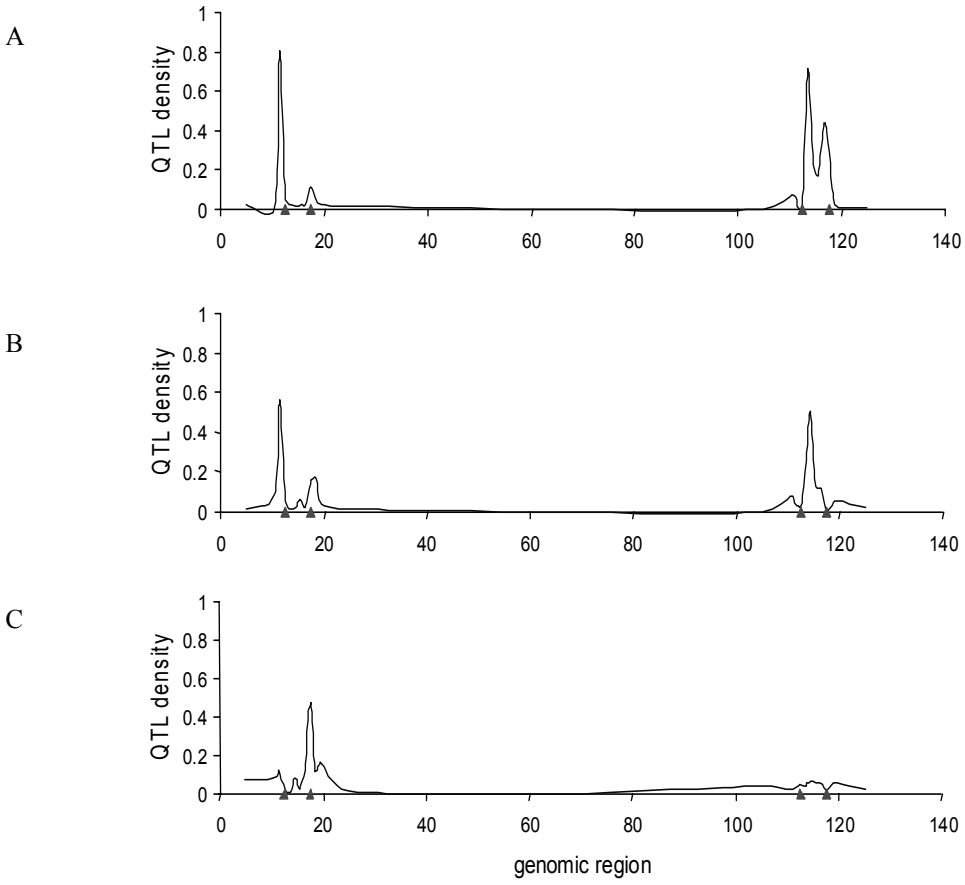


**Figure 1. The posterior QTL density with full model including additive, dominance and epistasis (A), additive and dominance model (B), and additive model (C) for two pairs of interacting closely linked QTL. Triangle shows the true QTL positions.**

**REFERENCES**

Akaike, H. 1969. *Ann Inst Statist Math* **21:** 243.
Andersson, L., Haley, C. S., Ellegren, H., Knott, S. A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfors-Lilja, I., Fredholm, M., Hansson, I., Hakansson, J., and Lundstrom, K. 1994. *Science* **263:** 1771.
Calborg, O., Andersson, L., and Kinghorn, B. P. 2000. *Genetics* **155**: 2003.
Carlborg, O., and Haley, C. S. 2004. *Nat Rev Genet* **5**: 618.
Elston, R. C., and Stewart, J. 1971. *Hum Hered* **21**: 523.
Fernando, R. L., and Grossman, M. 1989. *Genet Sel Evol* **21**: 467.
George, A. W., Visscher, P. M., and Haley, C. S. 2000. *Genetics* **156**: 2081.
Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., Sargeant, L. S., Sorensen, A., Steele, M. R., Zhao, X., Womack, J. E., and Hoeschele, I. 1995. *Genetics* **139:** 907.
Green, P. 1995. *Biometrika* **82**: 711.
Grignola, F. E., Hoeschele, I., and Tier, B. 1996. *Genet Sel Evol* **28**: 479.
Haley, C. S., and Knott, S. A. 1992. *Heredity* **69:** 315.
Hastings, W. K. 1970. *Biometricka* **57:** 97.
Heath, S. C. 1997. *Am J Hum Genet* **61***:* 748.
Holland, J. H. 1975. University of Michigan Press.
Jansen, R. C. 1993. *Genetics* **135**: 205.
Kao, C.-H., Zeng, Z.-B., and Teasdale, R. D. 1999. *Genetics* **152**: 1203.
Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. 1983. *Science* **220**: 671.
Lander, E. S., and Green, P. 1987. *Proc Natl Acad Sci U S A* **84:** 2363.
Lee, S. H., Van der Werf, J. H., and Tier, B. 2005. *Genetics* **171:** 2063.
Lee, S. H., and Van der Werf, J. H. J. 2004. *Genet Sel Evol* **36**: 145.
Lee, S. H., and Van der Werf, J. H. J. 2005. *Genetics* **169**: 455.
Lee, S. H., and Van der Werf, J. H. J. 2006a. *Genetics* **173**: 2329.
Lee, S. H., and Van der Werf, J. H. J. 2006b. *Genetics* **174**: 1009.
Liu, Y., Jansen, G. B., and Lin, C. Y. 2002. *Genet Sel Evol* **34**: 657.
MacCluer, J. W., VanderBerg, J. L., Raed, B., and Ryder, O. A. 1986. *Zoo Biology* **5**: 147.
Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. 1953. *J Chem Phys* **21**: 1087.
Meuwissen, T. H. E., and Goddard, M. E. 2000. *Genetics* **155**: 421.
Meuwissen, T. H. E., and Goddard, M. E. 2001. *Genet Sel Evol* **33**: 605.
Meuwissen, T. H. E., Karlsen, A., Lien, S., Olsaker, I., and Goddard, M. E. 2002. *Genetics* **161**: 373.
Mitchell, B. D., Ghosh, S., L., S. J., Birznieks, G., and Blangero, J. 1997. *Genetic Epidemiology* **14**: 1017.
Riquet, J., Coppieters, W., Cambisano, N., Arranz, J. J., Berzi, P., Davis, S. K., Grisart, B., Farnir, F., Karim, L., Mni, M., Simon, P., Taylor, J. F., Vanmanshoven, P., Wagenaar, D., Womack, J. E., and Georges, M. 1999. *Proc Natl Acad Sci U S A* **96**: 9252.
Schwarz, G. 1978. *Ann Stat* **6:** 461.
Sheehan, N. A., Possolo, A., and Thompson, E. A. 1989. Am J Hum Genet **45** (Suppl): A248.
Sillanpaa, M. J., and Arjas, E. 1998. *Genetics* **148:** 1373.

Sillanpaa, M. J., and Corander, J. 2002. *Trends Genet* **18:** 301.
Sobel, E., and Lange, K. 1996. *Am J Hum Genet* **58:** 1323.
Thompson, E. A. 1994. *Stat Sci* **9:** 355.
Thompson, E. A., and Heath, S. C. 1999. In:" Statistics in Molecular Biology and Genetics", editors F. Seillier-Moiseiwitsch , IMS Lecture notes. p 95, Institute of Mathematical Statistics, American Mathematical Society, Providence, RI.