Genet. Sel. Evol. 38 (2006) 25–43

DOI: 10.1051/gse:2005025

Original article

An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree

Sang Hong LEE*, Julius H.J. VAN DER WERF

School of Rural Science and Agriculture, UNE, Armidale, NSW2351, Australia

(Received 20 April 2005; accepted 5 October 2005)

Abstract – Variance component (VC) approaches based on restricted maximum likelihood (REML) have been used as an attractive method for positioning of quantitative trait loci (QTL). Linkage disequilibrium (LD) information can be easily implemented in the covariance structure among QTL effects (*e.g.* genotype relationship matrix) and mapping resolution appears to be high. Because of the use of LD information, the covariance structure becomes much richer and denser compared to the use of linkage information alone. This makes an average information (AI) REML algorithm based on mixed model equations and sparse matrix techniques less useful. In addition, (near-) singularity problems often occur with high marker densities, which is common in fine-mapping, causing numerical problems in AIREML based on mixed model equations. The present study investigates the direct use of the variance covariance matrix of all observations in AIREML for LD mapping with a general complex pedigree. The method presented is more efficient than the usual approach based on mixed model equations and robust to numerical problems caused by near-singularity due to closely linked markers. It is also feasible to fit multiple QTL simultaneously in the proposed method whereas this would drastically increase computing time when using mixed model equation-based methods.

quantitative trait loci / fine-mapping / linkage disequilibrium / average information / genotype relationships matrix

1. INTRODUCTION

Variance component (VC) approaches have been widely used to detect the existence of variation associated with quantitative trait loci (QTL) [1, 3, 10, 13, 15, 38]. The idea behind the approaches is to obtain identity by descent (IBD) coefficients between relatives for the QTL, based on

* Corresponding author: slee7@une.edu.au

marker and pedigree data, and maximize the likelihood of phenotypic data given these IBD coefficients at each putative QTL position. 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 [27, 28]. In the fine-mapping method, IBD coefficients between unrelated founders in a recorded pedigree are estimated based on haplotype similarity using the genedropping method [26] or the coalescence method. This allows utilizing unknown relationships beyond the recorded pedigree as well as known relationships, possibly allowing to estimate QTL position within a smaller region, *e.g.* within a few cM [30].

Data sets used for fine-mapping are typically not very large (a few hundred genotyped animals with records) because many marker loci with denser spacing are required. However, there is much richer information in the (co) variance structure (more non-zero elements in the variance covariance matrix). Such a complex and dense (co) variance structure makes a sparse matrix technique less useful. Therefore, the usual and efficient REML algorithm based on the mixed model equations (MME) using sparse matrix techniques (*e.g.* [12, 19]) may be less suitable for fine-mapping of QTL. Moreover, the complicated and subtle process to maximize computational efficiency in the MME-based REML, *i.e.* by avoiding the inverse of the variance covariance matrix of the observations, is prone to have numerical problems since the matrix of covariance coefficients between haplotypes based on closely linked markers can reach near-singularity. Direct use of the variance covariance matrix of the observations ($\mathbf{V}$) and its inverse would be robust to such problems because the $\mathbf{V}$ is an aggregate of (co) variance matrices of all components, which is more guaranteed to be positive definite. The dimension of the $\mathbf{V}$ is usually smaller than that of the MME, especially when there are many ancestors without records. Another advantage in using the $\mathbf{V}$ is to easily fit multiple QTL simultaneously since the dimension of the matrix is not changed. By contrast, the dimension of the MME rapidly increases with the number of QTL fitted.

The aim of this study was to investigate the efficiency of a REML algorithm with the direct use of the $\mathbf{V}$ compared to the usual REML algorithm based on MME when the combined LD and linkage (LDL) mapping is used with a general complex pedigree.

2. MATERIALS AND METHODS

2.1. Mixed linear model

The genetic model that is used in the VC approach with combined LD and linkage mapping is relatively general. A vector of phenotypic observations is written as a linear function of the effects of QTL, a polygenic term representing the sum of other unidentified genetic effects, fixed effects and residuals. The model can be expressed as,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \sum_{i=1}^{NQ} \mathbf{Z}\mathbf{q}_i + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is a vector of N observations on the trait of interest, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of n random polygenic effects for each animal, NQ is the number of QTL, $\mathbf{q}_i$ is a vector of n random effects due to ith putative QTL and $\mathbf{e}$ are residuals. The random effects ($\mathbf{u}$, $\mathbf{q}_i$ and $\mathbf{e}$) are assumed to be normally distributed with mean zero and variance $\mathbf{A}\sigma_u^2$, $\mathbf{G}_i\sigma_{q_i}^2$ and $\mathbf{R}$, where $\mathbf{A}$ is the numerator relationship matrix based on additive genetic relationships based on recorded pedigree, $\mathbf{G}_i$ is the genotype relationship matrix whose elements are IBD probabilities at ith QTL, and $\mathbf{R}$ is the covariance matrix among residual effects, assumed diagonal in this study. $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices for the effects $\boldsymbol{\beta}$ and $\mathbf{u}$ and $\mathbf{q}_i$, respectively. It is assumed that there is no correlation between $\mathbf{u}$ and $\mathbf{q}_i$, $\mathbf{u}$ and $\mathbf{e}$, and $\mathbf{q}_i$ and $\mathbf{e}$, and no correlation among $\mathbf{q}_1 \sim \mathbf{q}_{NQ}$. The associated variance covariance matrix of all observations ($\mathbf{V}$) given the observed pedigree and marker genotypes is modeled as

$$\mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_u^2 + \sum \mathbf{Z}\mathbf{G}_i\mathbf{Z}'\sigma_{q_i}^2 + \mathbf{R}. \tag{2}$$

The mixed model equations (MME) pertaining to (1) are

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \cdots & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\left(\mathbf{A}\sigma_u^2\right)^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \cdots & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \left(\mathbf{G}_1\sigma_{q_1}^2\right)^{-1} & \cdots & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \cdots & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\left(\mathbf{G}_{NQ}\sigma_{q_{NQ}}^2\right)^{-1} \end{bmatrix}$$

$$\times \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{q}}_1 \\ \vdots \\ \hat{\mathbf{q}}_{NQ} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \vdots \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{3}$$

2.2. Estimation of IBD probabilities

2.2.1. Finding inheritance states and haplotype construction

IBD coefficients are estimated based on the pattern of possible inheritance states and associated haplotype configurations [17, 34, 36]. 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 using maximum likelihood approaches. IBD coefficients are then estimated based on the most likely haplotypes. The other is to apply a Markov chain Monte Carlo (MCMC) algorithm to surface all possible inheritance states and haplotype configurations. IBD probabilities are estimated every MCMC cycle. Averaged IBD probabilities over all cycles would provide estimates based on the posterior distribution given the observed data [24].

2.2.2. LD information

IBD coefficients are based on similarity of haplotypes unrelated through known pedigree. These patterns of similarity can be derived by using a gene-dropping method [26] or the coalescence method introduced by Meuwissen and Goddard [27, 28]. An assumption of a mutation age of 100 generations and a past effective size of 100 can be used to estimate IBD coefficients since they are usually unknown. Results have been shown to be robust against such assumptions by exploring a range of values for effective size and mutation age as well as for populations that have a decreasing effective size [22, 23, 27]. Therefore, IBD coefficients between unrelated founder haplotypes in recorded pedigrees can be approximated and these are non-zero values.

2.2.3. Combined LD and linkage information

Using the IBD probabilities between unrelated haplotypes, IBD probabilities between related haplotypes in the following generations can be recursively estimated using known pedigree information [32,39,40]. Therefore, IBD probabilities between all haplotypes can be estimated based on joint information from LD and linkage. Note that all resulting coefficients have a non-zero value.

2.3. Genotype relationship matrix

The number of variance covariance coefficients between all haplotypes at each position is $2n \times 2n$ (n is the number of animals). From a computational

perspective, it is useful to transform the coefficients between all haplotypes to a covariance between individual genotypes without loss of information [32]. Therefore, the number of IBD coefficients are reduced to $n \times n$. The following equation is used [37],

$$\mathbf{G} = 0.5\ \mathbf{KHK}' \text{ with } \mathbf{K} = \mathbf{I}_n \otimes [1, 1].$$

Where **G** is the genotype relationship matrix, **H** is the haplotype (gametic) relationship matrix, $\mathbf{I}_n$ is an identity matrix with rank n, $\otimes$ is the Kronecker product of two matrices and [1, 1] is a 1 by 2 vector.

2.4. G and the position of the QTL

For a given data set, the similarity of marker haplotypes changes with the position of the QTL. Therefore, a different **G** exists for each putative QTL position on a tested chromosomal region. For each **G**, the maximum value of the log likelihood and the variance components are estimated for the corresponding position on a chromosome. Therefore, each position has a maximum value for the log likelihood for model parameters at the given QTL position. The most likely QTL position can be estimated as the position with the highest maximum likelihood value across all positions.

2.5. REML estimates using an average information algorithm

2.5.1. Calculation of log likelihood

By assuming multivariate normality of the data with mean vector **Xb** and variance covariance matrix **V**, the resulting likelihood can be written and some numerical procedure can be used to estimate the variance components. The log likelihood for the model in (1) can be obtained using the following equation [16].

$$\ln L = -\frac{1}{2}\left[\ln|\mathbf{V}| + \ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{P}\mathbf{y}\right] \tag{4}$$

where ln is a natural log, and | | refers to the determinant of the associated matrices. Note that **X** is a full rank incidence matrix. The **P** matrix is defined as,

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}. \tag{5}$$

Alternatively, avoiding the inverse of the **V** matrix, the following equation can be used [31],

$$\ln L = -\frac{1}{2}\Big[\left(N_r - N_\beta - N_u - N_q\right)\ln\hat{\sigma}_e^2 + \ln|\mathbf{C}| + \ln|\mathbf{A}| + \ln|\mathbf{G}| + n\ln\hat{\sigma}_u^2 + n\ln\hat{\sigma}_q^2 + \mathbf{y}'\mathbf{P}\mathbf{y}\Big] \quad (6)$$

where N_r, N_β, N_u and N_q is the number of records, fixed effects, polygenic effects and QTL effects and C is the coefficient matrix in the MME.

2.5.2. Average information REML

As a method to obtain REML estimates, the average information algorithm (AIREML [11, 19]) is used. The average information coefficients are the average of the observed and expected information matrices from the Newton-Raphson and Fisher scoring method (see App. A). The equation for the iterative algorithm is,

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} + \left(\mathbf{AI}^{(k)}\right)^{-1}\frac{\partial \mathbf{L}}{\partial \boldsymbol{\Theta}}\Big|_{\boldsymbol{\Theta}^{(k)}} \quad (7)$$

where $\boldsymbol{\Theta}$ is a column vector of variance components (σ_e^2, σ_u^2 and σ_q^2), and k is the iteration round. AI is the average of the observed and expected information matrices which consists of the second derivatives of the log likelihood function with respect to the variance components, which can be written as,

$$\mathbf{AI} = \frac{1}{2}\begin{bmatrix} \mathbf{y}'\mathbf{PPPy} & \mathbf{y}'\mathbf{PA}^*\mathbf{PPy} & \mathbf{y}'\mathbf{PG}^*\mathbf{PPy} \\ \mathbf{y}'\mathbf{PA}^*\mathbf{PPy} & \mathbf{y}'\mathbf{PA}^*\mathbf{PA}^*\mathbf{Py} & \mathbf{y}'\mathbf{PA}^*\mathbf{PG}^*\mathbf{Py} \\ \mathbf{y}'\mathbf{PG}^*\mathbf{PPy} & \mathbf{y}'\mathbf{PA}^*\mathbf{PG}^*\mathbf{Py} & \mathbf{y}'\mathbf{PG}^*\mathbf{PG}^*\mathbf{Py} \end{bmatrix} \quad (8)$$

where $\mathbf{A}^* = \mathbf{ZAZ}'$ and $\mathbf{G}^* = \mathbf{ZGZ}'$.

$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\Theta}}$ is a column vector with first derivatives of the log likelihood function with respect to each variance component, which is,

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\sigma}_i^2} = -\frac{1}{2}tr\left(P\frac{\partial \mathbf{V}}{\partial \boldsymbol{\sigma}_i^2}\right) + \frac{1}{2}\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\sigma}_i^2}\mathbf{P}\mathbf{y}.$$

The first derivatives for each of the three VC of the QTL model are,

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\sigma}_u^2} = -\frac{1}{2}(tr(\mathbf{PA}^*) - \mathbf{y}'\mathbf{PA}^*\mathbf{Py}) \tag{9a}$$

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\sigma}_q^2} = -\frac{1}{2}(tr(\mathbf{PG}^*) - \mathbf{y}'\mathbf{PG}^*\mathbf{Py}) \tag{9b}$$

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\sigma}_e^2} = -\frac{1}{2}(tr(\mathbf{P}) - \mathbf{y}'\mathbf{PPy}). \tag{9c}$$

In general cases, the use of the MME (3) is very efficient to obtain AI coefficients and the first derivatives because one can avoid inverting matrix **V** (see App. B and C). However, in fine-mapping, the QTL part in the MME becomes dense, substantially reducing the computational efficiency of this procedure.

2.6. Comparison of the direct and MME method

The **AI** matrix (8) and the first derivatives (9) can be obtained in two different ways. One is to directly compute the inverse of the **V** matrix, and then the **P** matrix (direct method). The other is to solve MME equations and use matrix products (MME-based methods, see App. B). In the direct method, the main part of the computations is to invert **V**, needed for matrix **P** for each REML iteration round (6). The dimension of the **V** matrix is $N_r \cdot N_r$ and matrix inversion requires order $(N_r)^3$ computations. However, the partial inverse of the coefficient matrix corresponding to polygenic and QTL effects is necessary in the MME-based methods. The dimension of the coefficient matrix corresponding to polygenic and QTL effects is $(N_u + N_q)\cdot(N_u + N_q)$ and partial inversion requires computations slightly more than order $(N_u)^3 + (N_q)^3$. In a fine-mapping model where **G** is a dense matrix, the direct method is expected to be more computationally efficient unless the number of records is far greater than the number of animals. Note that the number of animals is often much larger than the number of records with genotypic data. For example, in a pedigree spanning several generations, often only the last one or two generations would be phenotyped and genotyped for mapping studies. Another advantage in the direct method is to easily fit multiple QTL simultaneously. This is because the dimension of **V** (2) is not changed by increasing the number of QTL. In contrast, the dimension in the MME (3) rapidly increases with the number of QTL, *i.e.* $\left(N_u + \sum_{i=1}^{NQ} N_{q_i}\right) \cdot \left(N_u + \sum_{i=1}^{NQ} N_{q_i}\right)$.

2.7. Simulation study

One hundred generations of a population with an effective size of 100 was simulated for 11 markers and a QTL. In each generation, the number of male and female parents was 50 and their alleles were transmitted to descendents based on Mendelian segregation using the genedropping method [26]. Parents were randomly mated with a total of 2 offspring for each of 50 mating pairs. For the QTL, one of the base alleles surviving with a frequency of more than 0.1 and less than 0.9 was randomly chosen and treated as favourable with effect α compared to other QTL alleles in generation 100. The number of base alleles in each marker locus was 4 and starting allele frequencies were all at 0.25. The marker alleles were mutated at a rate of 4×10^{-4} per generation [5,7,41]. Therefore, this historical population would have an equilibrium distribution of alleles in all marker loci and LD between the QTL and closely flanking markers. Note that pedigree information is available only from generation 101 onwards.

The computational efficiency of the direct method and the MME-based method was investigated. Analyses were carried out for a pedigree spanning two generations (generation 100 $\sim$ 101), 5 generations (100 $\sim$ 104) and 10 generations (100 $\sim$ 109). By default, random selection and mating was carried out, resulting in many marriage loops and inbreeding loops, therefore a complex pedigree. Note that even with a pedigree of two generations, complex marriage loops heavily affect the covariance structure (*i.e.* reducing sparsity of the inverse of the covariance matrix). For a simple pedigree with each parent having one mate only, marriage and inbreeding loops do not exist. We generated 50 unrelated full sib families each with two progeny and compared with a complex pedigree. For a fair comparison between results, marker genotypes and phenotypic values were only available for animals in the last two generations. Phenotypic values were simulated as $y_i = \mu + \alpha_i + u_i + e_i$. The mean of population (μ) was 100, values for u_i were drawn from $N(0, A\sigma_u^2)$ with $\sigma_u^2 = 25$, and values for e_i were from $N(0, \sigma_e^2)$ with $\sigma_e^2 = 50$. The favorable QTL allele had an additive value of 7 ($\alpha_0 = 0$ and $\alpha_1 = 7$), therefore, QTL variance ranged from 8.8 to 24.5 with $V_{QTL} = 2p(1-p)\alpha^2$ [8]. To evaluate the effects of marker densities on computational stability and efficiency, eleven markers were positioned at 10, 1 or 0.1 cM intervals.

3. RESULTS

3.1. Computational efficiency for the MME and direct method

Table I shows the computing time averaged over 10 replicates per REML iteration for each method with a general complex pedigree spanning 2,

Table I. Computational statistics averaged over replicates for each method with complex pedigrees.

Data set	2 generations[a]	5 generations[a]	10 generations[a]
	Inbreeding coefficient in the last generation		
	0	0.014	0.039
	No. of non-zero elements in MME[b]		
LDL mapping	42800	254887	1091920
Linkage mapping	32951	240620	1063153
	No. of non-zero elements in inverse of **G**		
LDL mapping	40000	250000	1000000
Linkage mapping	30152	235732	971233
	No. of non-zero elements in inverse of **NRM**[c]		
	796	2883	89916
	Comp. time (s) per iteration with the direct method[d]		
	0.09 (< 0.001)[e]	0.09 (< 0.001)	0.15 (< 0.001)
	Comp. time (s) per iteration with the MME method		
LDL mapping	0.19 (< 0.001)	4.8 (0.02)	43 (0.5)
Linkage mapping	0.13 (0.004)	4.5 (0.02)	41 (0.3)

[a] Genotypes and phenotypes are only available for the last two generations (100 animals per generation).

[b] Total number of elements in MME $\sim \left(N_u + N_q\right)^2$.

[c] Elements in the inverse of **NRM** are the same for LDL mapping and linkage mapping.

[d] Computing time with the direct method is the same for LDL mapping and linkage mapping.

[e] Standard error over 10 replicates in the bracket.

5 or 10 generations. The results show that the computational effort per iteration round of the direct method is lower in all cases than that of the MME method. When the number of animals and the number of records are the same (*e.g.* $N_r = N_u = N_q = 200$), the computing time of the direct method is around $1.5 \sim 2$ times lower than that of the MME method. With a pedigree spanning 5 generations where $N_r = 200$ and $N_u = N_q = 500$, the direct method is about 50 times faster than the MME method. With a pedigree of 10 generations with $N_r = 200$ and $N_u = N_q = 1000$, the direct method is around 270 times faster than the MME method. As expected, the direct method performed at a similar computational speed regardless of the number of non-zero elements in the

Table II. Computational statistics averaged over replicates for each method with complex or simple pedigree.

Data set	Complex pedigree[a]	50 full sib families[b]
	Proportion of non-zero elements in MME	
LDL mapping	42800	42704
Linkage mapping	32951	3499
	No. of non-zero elements in inverse of **G**	
LDL mapping	40000	40000
Linkage mapping	30152	795
	No. of non-zero elements in inverse of **NRM**	
	796	700
	Comp. time (s) with the direct method	
	0.09 (< 0.001)[c]	0.09 (< 0.001)
	Comp. time (s) with the MME method	
LDL mapping	0.19 (< 0.001)	0.18 (< 0.001)
Linkage mapping	0.13 (0.004)	0.02 (< 0.001)

[a] Random selection and random mating with effective size of 100 of 2 generations (number of animals are 200).
[b] 50 families each with 2 offspring (number of animals are 200).
[c] Standard error over 10 replicates in the bracket.

MME matrix while the computing time of the MME method rapidly increases with a larger number of non-zero elements.

With a general complex pedigree, the inverse of **G** is very dense even in linkage mapping (*e.g.* 76% of non-zero elements in a pedigree spanning two generations). Therefore, a sparse matrix technique is not useful. With a more complex pedigree, the proportion of non-zero elements increases (95% and 97% for a pedigree of 5 and 10 generations). This may explain a smaller difference of computing time between LDL mapping and linkage mapping with a more complex pedigree. Table II shows the number of non-zero elements and computing time both for a complex pedigree spanning two generations and a simple pedigree (of 50 unrelated full sib families). Although the number of animals are the same (200), the number of non-zero elements in the inverse of **G** is much smaller for a simple pedigree than for a complex pedigree when using linkage mapping. In this case, a sparse matrix technique is very useful and the computing time of the MME method is lower than that of the direct method (*e.g.* for linkage mapping with 50 full sib families).

Table III. Proportion of replicates having numerical problems[a] in the procedure.

Marker density	10 cM	1 cM	0.1 cM
Direct method	0	0	0
MME method	0	0.24	0.69

[a] Numerical problems: $\ln L$ (6) cannot be obtained due to singularity or dependency of the coefficient matrix. Estimations were carried out for 10 putative QTL positions in 10 simulated data sets, therefore, 100 estimations were replicated.

3.2. Stability of the direct method and MME method

Table III shows the proportion of replicates having numerical problems to obtain the log likelihood and variance components when LDL mapping is carried out. Note that if **G** is non-positive definite, a bending algorithm is used to ensure positive definiteness for **G** [35]. Thus, positive definite **G** is used for both methods. When marker density is low (> 10 cM), the log likelihood and parameters are estimable in all replicates for both the direct and MME method. When marker density is higher (1 cM or 0.1 cM), the MME method often faces numerical problems, *i.e.* $\ln |C|$ in (6) cannot be obtained, therefore, $\ln L$ (6) cannot be estimated. This is probably due to the fact that very high marker density increases the likelihood of the coefficient matrix to be singular. The **G** matrix is an explicit part of the MME, and even though this matrix is bent to become positive definite, the Gaussian elimination procedure could still face very small pivotal values and there is a good chance of negative values for determinants. These problems rarely or never occur for the direct method implementing the **V** that is the sum of all covariance matrices.

3.3. Comparison with a standard VC software 'ASReml'

Table IV compares the computing time per iteration for the direct method with that for ASReml [12]. In LDL mapping, the direct method is much more efficient than ASReml especially when using a complex pedigree spanning 10 generations. In linkage mapping alone, the direct method performs better with a complex pedigree, however, with a simple pedigree, ASReml is more efficient. It is noted that the performance of ASReml is similar to that of our MME method (Tabs. I and II). This is because both programs use the extended MME to obtain AI coefficients and the inverse of the coefficient matrix with a sparse matrix technique (see App. B). Optimal ordering of sparse structures and utilization of the reordering matrix [6] may be more efficiently optimised in ASReml than in our MME method, making ASReml perform better than

Table IV. Comparison of computing time (s) averaged over replicates to estimate variance components between the direct method and 'ASReml'.

	Simple pedigree	General complex pedigrees		
	50 full sib families	2 generations	5 generations	10 generations
	Direct method			
	0.09 (< 0.001)[a]	0.09 (< 0.001)	0.09 (< 0.001)	0.15 (< 0.001)
	ASReml			
LDL	0.3 (< 0.001)	0.33 (0.001)	5.7 (0.02)	50 (0.3)
Linkage alone	< 0.01 (< 0.001)	0.2 (0.01)	5.3 (0.03)	46 (0.2)

[a] Standard error over 10 replicates in the bracket.

the MME method for a simple pedigree. However, in general, both programs perform similarly in that the computational effort of the MME-based procedure is high with LDL mapping and complex pedigrees whereas it is much lower with a simple pedigree structure in linkage mapping.

4. DISCUSSION

This study presented a REML procedure suitable for fine-mapping of QTL with a complex pedigree. Because the coefficient matrix is dense with LDL mapping and a complex pedigree, MME-based methods are less computationally efficient and sparse matrix techniques are not very useful. Besides, with closely linked multi markers in fine-mapping (*e.g.* marker spacing $<\sim$ 1 cM), MME-based methods could face numerical problems because of dependency or near-singularity in the MME. These problems were not observed in the direct method.

It is common that genotypic and phenotypic observations are available only on relatively few animals from the last few generations in an entire pedigree (genotypes and phenotypes for ancestors are missing). In such situations when the number of animals in the pedigree is similar or greater than the number of available observations, the direct method is computationally efficient since the **V** describes the variance covariance structure between observations, taking into account all the ancestral relationships.

When multiple QTL are involved in phenotypes of a trait, it is useful to analyze a number of QTL simultaneously [18, 20, 29, 42, 43]. It is straightforward to simultaneously include a number of QTL positions in the model. The computing time for a multiple QTL analysis is not much different from that of a single QTL analysis if the direct method is used; the dimension of the **V** is not

changed and the computing time for inverting the **V** is the same in the multiple or single QTL model. However, in MME-based methods, the computing time rapidly increases with a larger number of QTL positions fitted since the dimension of MME linearly increases with additional QTL positions included in the model. For example, with a complex pedigree of two generations, computing time per iteration is 0.11, 0.12 and 0.13 (s) for the direct method and 1.6, 4.7 and 19 (s) for ASReml when fitting 2, 3 and 5 multiple QTL simultaneously using LDL mapping. These results were expected given (2) and (3).

Since estimates were based on the same likelihood equation (7), theoretically, AI coefficients and first derivatives should have the same value in each method if starting value and convergence criteria are the same. In this study with a complex pedigree of two generations, the average number of iterations over replicates was 4.4 (±0.27) for the direct method and 6.4 (±0.27) for ASReml. This difference was probably due to the fact that the first and second derivatives were estimated from two different coefficient matrices (*e.g.* **V** matrix and MME) and the computational procedure was different. The small number of iterations for the direct method is probably due to the fact that the procedure is more stable, therefore quicker to reach convergence than the MME-based method (see Sect. 3.2). However, the estimated variance components and maximum likelihood for the direct and ASReml agreed well when the procedure was completed without numerical problems (results not shown).

One could consider using IBD coefficients between haplotypes rather than between animals, which can make the inverse of the matrix (**H**) more sparse. However, this would be advantageous only if the proportion of zero elements is much higher in **H** than in **G**. This is not the case in LDL mapping in which IBD coefficients between all haplotypes (of base animals and descendents as well) are non-zero elements. The inverse of the **H** matrix is a part of MME. The MME must be inverted every iteration. Unless the proportion of zero elements is very large in the MME, inverting MME is computationally heavy whether partial matrix theory with a recursive method (*e.g.* [9, 12]) is used or not. This is because all non-zero elements must be involved in inverting MME. As a matter of fact, the **H** matrix has order $2N_q \times 2N_q$ which makes MME bigger, *e.g.* the dimension of MME with the **H** matrix is $\sim (N_u + 2N_q) \times (N_u + 2N_q)$. However, the dimension of MME with the **G** matrix is $\sim (N_u + N_q) \times (N_u + N_q)$. In linkage mapping with a relatively simple pedigree, the proportion of zero elements is much higher in **H** than **G**. This results in a very sparse structure in the inverse of **H** and the MME. In this case, zero elements can be skipped in the computation (using sparse matrix techniques), therefore a considerable proportion of elements does not have to be involved in inverting MME (saving

computational cost). However, in LDL mapping, $\mathbf{H}$ and $\mathbf{H}^{-1}$ do not have any zero elements. Therefore, more than $2N_q \times N_q$ elements must be involved in inverting MME. Note that the number of elements reduces to $N_q \times N_q$ when using $\mathbf{G}$, which is therefore computationally more efficient.

In the MME method, we used the package AMD version 1.1 [2] for an optimal ordering of sparse matrices. We found this procedure very useful and dramatically reducing the computing time if and only if there was a considerable proportion of zero elements in the MME (*e.g.* linkage mapping with simple pedigree). However, when LDL mapping was used, the computing time was not changed. Note that the optimal ordering had been already done before the iteration started, the time for the optimal ordering was not included in the computing time per iteration.

5. CONCLUSION

The direct method is generally suitable for fine-mapping of QTL with closely linked markers and a complex pedigree where genotypes and observations are available for the last few generations. Efficient algorithms also make use of proper statistical testing techniques such as permutation testing [4] in fine-mapping more feasible. Only when linkage mapping is applied with a simple pedigree and sparse marker spacing, will the MME method have a similar efficiency. In addition, the direct method has the potential to easily accommodate multiple QTL because the dimension of $\mathbf{V}$ is not affected by the number of QTL.

ACKNOWLEDGEMENTS

We are thankful for valuable comments from the reviewers. This study was supported by Australian Wool Innovation and a University of New England research assistantship.

REFERENCES

[1] Almasy L., Blangero J., Multipoint quantitative-trait linkage analysis in general pedigrees, Am. J. Hum. Genet. 62 (1998) 1198–1211.
[2] Amestoy P.R., Davis T.A., Duff I.S., Amd version 1.1 user guide, 2004.
[3] Amos C.I., Robust variance components approach for assessing genetic linkage in pedigrees, Am. J. Hum. Genet. 54 (1994) 535–543.
[4] Churchill G.A., Doerge R.W., Empirical threshold values for quantitative trait mapping, Genetics 138 (1994) 963–971.

[5] Dallas J.F., Estimation of microsatellite mutation rates in recombinant inbred strains of mouse, Mamm. Genome 3 (1992) 452–456.
[6] Duff I.S., Erisman A.M., Reid J.K., Direct method for sparse matrix, Oxford, Clarendon Press, 1989.
[7] Ellegren H., Mutation rates at porcine microsatellite loci, Mamm. Genome 6 (1995) 376–377.
[8] Falconer D.S., Mackay T.F.C., Introduction to quantitative genetics, 4th edn., Longman, 1996.
[9] Fernando R.L., Grossman M., Marker assisted selection using best linear unbiased prediction, Genet. Sel. Evol. 21 (1989) 467–477.
[10] George A.W., Visscher P.M., Haley C.S., Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach, Genetics 156 (2000) 2081–2092.
[11] Gilmour A.R., Thompson R., Cullis B.R., Average information REML: An efficient algorithm for variance parameters estimation in linear mixed models, Biometrics 51 (1995) 1440–1450.
[12] Gilmour A.R., Cullis B.R., Welham S.J., Thompson R., Asreml reference manual, Orange Agriculture Institute, New South Wales, Australia, 2004.
[13] Goldgar D.E., Multipoint analysis of human quantitative genetic variation, Am. J. Hum. Genet. 47 (1990) 957–967.
[14] Graser H.-U., Smith S.P., Tier B., A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood, J. Anim. Sci. 64 (1987) 1362–1370.
[15] Grignola F.E., Hoeschele I., Tier B., Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. Methodology, Genet. Sel. Evol. 28 (1996) 479–490.
[16] Harville D.A., Maximum likelihood approaches to variance component estimation and to related problems, J. Am. Stat. Assoc. 72 (1977) 320–338.
[17] Heath S.C., Markov chain Monte Carlo segregation and linkage analysis for oligogenic models, Am. J. Hum. Genet. 61 (1997) 748–760.
[18] Jansen R.C., Interval mapping of multiple quantitative trait loci, Genetics 135 (1993) 205–211.
[19] Johnson D.L., Thompson R., Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information, J. Dairy Sci. 78 (1995) 449–456.
[20] Kao C.H., Zeng Z.B., Teasdale R.D., Multiple interval mapping for quantitative trait loci, Genetics 152 (1999) 1203–1216.
[21] Lange K., Westlake J., Spence M.A., Extensions to pedigree analysis. Iii. Variance components by the scoring method, Ann. Hum. Genet. 39 (1976) 485–491.
[22] Lee S.H., van der Werf J.H.J., Efficient designs for fine-mapping of quantitative trait loci using linkage disequilibrium and linkage, Proceedings of the Association for the Advancement of Animal Breeding and Genetics 15 (2003) 9–13.

[23] Lee S.H., van der Werf J.H.J., The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage, Genet. Sel. Evol. 36 (2004) 145–161.
[24] Lee S.H., van der Werf J.H.J., The role of pedigree information in combined linkage disequilibrium and linkage mapping of quantitative trait loci in a general complex pedigree, Genetics 169 (2005) 455–466.
[25] Lynch M., Walsh B., Genetics and analysis of quantitative traits, 1st edn., Sinauer Associates, Sunderland, 1998.
[26] MacCluer J.W., VanderBerg J.L., Raed B., Ryder O.A., Pedigree analysis by computer simulation, Zoo Biol. 5 (1986) 147–160.
[27] Meuwissen T.H.E., Goddard M.E., Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci, Genetics 155 (2000) 421–430.
[28] Meuwissen T.H.E., Goddard M.E., Prediction of identity by descent probabilities from marker-haplotypes, Genet. Sel. Evol. 33 (2001) 605–634.
[29] Meuwissen T.H.E., Goddard M.E., Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data, Genet. Sel. Evol. 36 (2004) 261–279.
[30] Meuwissen T.H.E., Karlsen A., Lien S., Olsaker I., Goddard M.E., Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping, Genetics 161 (2002) 373–379.
[31] Meyer K., Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm, Genet. Sel. Evol. 21 (1989) 317–340.
[32] Pong-Wong R., George A.W., Woolliams J.A., Haley C.S., A simple and rapid method for calculating identity-by-descent matrices using multiple markers, Genet. Sel. Evol. 33 (2001) 453–471.
[33] Searle S.R., Casella G., McCulloch C.E., Variance components, John Wiley & Sons, New York, NY, 1992.
[34] Sobel E., Lange K., Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics, Am. J. Hum. Genet. 58 (1996) 1323–1337.
[35] Sorensen A.C., Pong-Wong R., Windig J.J., Woolliams J.A., Precision of methods for calculating identity-by-descent matrices using multiple markers, Genet. Sel. Evol. 34 (2002) 557–579.
[36] Thompson E.A., Heath S.C., Estimation of conditional multilocus gene identity among relatives, in: Seller-Moiseiwitsch (Ed.), Statistics in molecular biology and genetics, ims lecture notes, Institute of Mathematical Statistics, American Mathematical Society, Providence, RI, 1999, pp. 95–113.
[37] Tier B., Solkner J., Analysing gametic variation with an animal model, Theor. Appl. Genet. 85 (1993) 868–872.
[38] Van Arendonk J.A., Tier B., Bink M.C., Bovenhuis H., Restricted maximum likelihood analysis of linkage between genetic markers and quantitative trait loci for a granddaughter design, J. Dairy Sci. 81 Suppl. 2 (1998) 76–84.
[39] Van Arendonk J.A., Tier B., Kinghorn B.P., Use of multiple genetic markers in prediction of breeding values, Genetics 137 (1994) 319–329.

[40] Wang T., Fernando R.L., van der Beek S., Grossman M., van Arendonk J.A.M., Covariance between relatives for a marked quantitative trait locus, Genet. Sel. Evol. 27 (1995) 251–274.
[41] Weber J.L., Wong C., Mutation of human short tandem repeats, Hum. Mol. Genet. 2 (1993) 1123–1128.
[42] Zeng Z.B., Precision mapping of quantitative trait loci, Genetics 136 (1993) 1456–1468.
[43] Zeng Z.B., Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci, Proc. Natl. Acad. Sci. USA 90 (1993) 10972–10976.

APPENDIX A: NEWTON-RAPHSON AND FISHER SCORING ALGORITHM

The Newton-Raphson algorithm obtains the REML estimates using the following equation [25].

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} - \left(\mathbf{H}^{(k)}\right)^{-1} \frac{\partial \mathbf{L}}{\partial \boldsymbol{\Theta}} \Big|_{\boldsymbol{\Theta}^{(k)}} \tag{A1}$$

where $\boldsymbol{\Theta}$ is a column vector of VC (σ_e^2, σ_u^2 and σ_q^2), k is the iteration round, $\frac{\partial \mathbf{L}}{\partial \boldsymbol{\Theta}}$ is a column vector of the first derivatives of the log likelihood function with respect to each variance component, and $\mathbf{H}$ is the Hessian matrix which consists of the second derivatives of the log likelihood function with respect to the variance components. In the Fisher scoring method, the inverse of the Hessian matrix in (A1) is replaced by minus its expected value [25].

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} + (\mathbf{F}^{(k)})^{-1} \frac{\partial \mathbf{L}}{\partial \boldsymbol{\Theta}} \Big|_{\boldsymbol{\Theta}^{(k)}} . \tag{A2}$$

The derivation of the Hessian matrix and the Fisher information matrix has been described in several studies [21, 25, 33]. **AI** (8) is the average of the Hessian and Fisher information matrix.

APPENDIX B: CALCULATION OF THE ELEMENTS OF THE AI MATRIX USING MME-BASED METHODS

Use of MME

The MME in (3) can be used to estimate the elements in the **AI** matrix as in [19]. This method is potentially useful to avoid calculating the inverse of matrix $\mathbf{V}$. To simplify notation, we define $\mathbf{W}_i = \frac{\partial \mathbf{V}}{\partial \sigma_i^2} Py$ as working variates for

all components ($i = u, q$ and e). The working variates for our QTL model are calculated as,

$$\mathbf{W}_u = \frac{\partial \mathbf{V}}{\partial \boldsymbol{\sigma}_u^2}\mathbf{P}\mathbf{y} = \frac{1}{\boldsymbol{\sigma}_u^2}\mathbf{Z}\hat{\mathbf{u}} \tag{A3a}$$

$$\mathbf{W}_q = \frac{\partial \mathbf{V}}{\partial \boldsymbol{\sigma}_q^2}\mathbf{P}\mathbf{y} = \frac{1}{\boldsymbol{\sigma}_q^2}\mathbf{Z}\hat{\mathbf{q}} \tag{A3b}$$

$$\mathbf{W}_e = \frac{\partial \mathbf{V}}{\partial \boldsymbol{\sigma}_e^2}\mathbf{P}\mathbf{y} = \frac{1}{\boldsymbol{\sigma}_e^2}\hat{\mathbf{e}} \tag{A3c}$$

where $\hat{\mathbf{u}}$ and $\hat{\mathbf{q}}$ are solutions from the MME, and $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}} - \mathbf{Z}\hat{\mathbf{q}}$.

The matrix $\mathbf{P}$ is a projection matrix transforming the observation vector ($\mathbf{y}$) into residuals (*e.g.* $\hat{\mathbf{e}} = \mathbf{P}\mathbf{y}$). Furthermore, $\mathbf{P}\mathbf{W}_e$ is the vector of residuals obtained when $\mathbf{W}_e$ is taken as the observation vector instead of $\mathbf{y}$. In this manner, $\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{P}\mathbf{y}$ can be obtained as the scalar product of the scaled residual vector ($\mathbf{W}_e$) and the new residual vector ($\mathbf{P}\mathbf{W}_e$) obtained when $\mathbf{W}_e$ is taken as the observation vector, *i.e.* $\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{P}\mathbf{y} = \mathbf{W}_e'\mathbf{P}\mathbf{W}_e$. All elements in the **AI** matrix can be obtained in the same manner.

Use of Gaussian elimination on extended MME

Gilmour *et al.* [12] used an extended MME with Gaussian elimination to estimate the elements of the **AI** matrix. We construct an **M** matrix including the QTL term as an extension of the MME in (3).

$$\mathbf{M} = \begin{bmatrix} \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} & \mathbf{y}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{y}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{y}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + (\mathbf{A}\boldsymbol{\sigma}_u^2)^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + (\mathbf{G}\boldsymbol{\sigma}_q^2)^{-1} \end{bmatrix} \tag{A4}$$

After performing Gaussian elimination, the first row and first column, **M** (1,1), equals $\mathbf{y}'\mathbf{P}\mathbf{y}$ [14]. If $\mathbf{y}$ is replaced by $\mathbf{W}_e$ from (A3c), then **M** (1,1) after Gaussian elimination equals $\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{P}\mathbf{y}$ that is one of the elements in the **AI** matrix. If $\mathbf{y}$ is replaced by $\mathbf{W}_q$ from (A3b), then **M** (1,1) after Gaussian elimination equals $\mathbf{y}'\mathbf{P}\mathbf{G}^*\mathbf{P}\mathbf{G}^*\mathbf{P}\mathbf{y}$ that is another element in the **AI** matrix. All elements in the **AI** matrix can be calculated in the same manner.

APPENDIX C: CALCULATION OF THE FIRST DERIVATIVES USING MME-BASED METHODS

In the MME-based method, since matrix P is not explicitly calculated, the following equations similar to [19] can be used to replace (9 a, b and c).

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\sigma}_u^2} = -\frac{1}{2}\left[\frac{N_u}{\boldsymbol{\sigma}_u^2} - \frac{tr(\mathbf{A}^{-1}C^{uu})}{\boldsymbol{\sigma}_u^4} - \left(\frac{\hat{e}}{\boldsymbol{\sigma}_e^2}\right)'\left(\frac{Z\hat{u}}{\boldsymbol{\sigma}_u^2}\right)\right] \tag{A5a}$$

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\sigma}_q^2} = -\frac{1}{2}\left[\frac{N_q}{\boldsymbol{\sigma}_q^2} - \frac{tr(\mathbf{G}^{-1}\mathbf{C}^{qq})}{\boldsymbol{\sigma}_q^4} - \left(\frac{\hat{e}}{\boldsymbol{\sigma}_e^2}\right)'\left(\frac{Z\hat{q}}{\boldsymbol{\sigma}_q^2}\right)\right] \tag{A5b}$$

$$\frac{\partial \mathbf{L}}{\partial \boldsymbol{\sigma}_e^2} = -\frac{1}{2}\left[\frac{N_r - nf}{\boldsymbol{\sigma}_e^2} - \left(N_u - \frac{tr(\mathbf{A}^{-1}\mathbf{C}^{uu})}{\boldsymbol{\sigma}_u^2}\right)\frac{1}{\boldsymbol{\sigma}_e^2} - \left(N_q - \frac{tr(\mathbf{G}^{-1}\mathbf{C}^{qq})}{\boldsymbol{\sigma}_q^2}\right)\frac{1}{\boldsymbol{\sigma}_e^2} - \left(\frac{\hat{e}'\hat{e}}{\boldsymbol{\sigma}_e^4}\right)\right] \tag{A5c}$$

where C^{uu} and C^{qq} is the partition of the inverse of the coefficient matrix corresponding polygenic effects (u) and QTL effect (q).

To access this journal online:
www.edpsciences.org