Genet. Res., Camb. (2011), **93**, *pp.* 333–342. © Cambridge University Press 2011
doi:10.1017/S0016672311000152

How to deal with genotype uncertainty in variance component quantitative trait loci analyses

XIA SHEN[1,2]*, LARS RÖNNEGÅRD[2,3] AND ÖRJAN CARLBORG[1,3]
[1] *The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden*
[2] *School of Technology and Business Studies, Dalarna University, Borlänge, Sweden*
[3] *Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden*

(*Received 10 October 2010; revised 4 March 2011; first published online 18 July 2011*)

Summary

Dealing with genotype uncertainty is an ongoing issue in genetic analyses of complex traits. Here we consider genotype uncertainty in quantitative trait loci (QTL) analyses for large crosses in variance component models, where the genetic information is included in identity-by-descent (IBD) matrices. An IBD matrix is one realization from a distribution of potential IBD matrices given available marker information. In QTL analyses, its expectation is normally used resulting in potentially reduced accuracy and loss of power. Previously, IBD distributions have been included in models for small human full-sib families. We develop an Expectation–Maximization (EM) algorithm for estimating a full model based on Monte Carlo imputation for applications in large animal pedigrees. Our simulations show that the bias of variance component estimates using traditional expected IBD matrix can be adjusted by accounting for the distribution and that the calculations are computationally feasible for large pedigrees.

1. Introduction

Variance component models have played an important role in detecting quantitative trait loci (QTL) for the last couple of decades in both animal breeding (Fernando & Grossman, 1989; Goddard, 1992; Arendonk *et al*., 1994; Wang *et al*., 1995; George *et al*., 2000) and human genetics (Goldgar, 1990; Schork, 1993; Fulker & Cardon, 1994; Olson, 1995; Xu & Atchley, 1995; Blangero *et al*., 2001). To construct the variance–covariance matrix of the random QTL effect, identity-by-descent (IBD) probabilities are required. The IBD probabilities describe the correlation structure between individuals with respect to the frequency of their shared (common) alleles. The genetic variance component estimates, and the corresponding likelihoods, are usually calculated using an estimated IBD matrix.

The IBD matrix can be estimated from marker information using either deterministic (Wang *et al*., 1995; Pong-Wong *et al*., 2001; Besnier & Carlborg, 2007) or stochastic algorithms (Thompson & Heath, 1999; Pérez-Enciso *et al*., 2000; Mao & Xu, 2005). All these methods actually calculate an average IBD matrix, where each entry is the average frequency of shared alleles, based on partially informative markers. Namely, all the IBD values are known in the statistical models. Instead of using the average IBD matrix, which we refer to as the *expectation method* (Xu, 1996), the uncertainty of the IBD matrix itself may also be included in the likelihood. Such a method accounting for the uncertainty of the IBD matrix is referred to as the *distribution method*. The likelihood function that the distribution method uses is called the *full likelihood function*, in contrast to the expectation method that uses an approximated likelihood function.

Comparison of distribution methods with expectation methods has been a thoroughly investigated problem in human genetics, especially for regression models in QTL analysis and genome-wide association studies (GWAS). Using genotype imputation, GWAS can gain power at positions with uncertain genotypes (Marchini & Howie, 2010) where the expectation method gives good power and accuracy (Kutalik *et al*., 2011) by using SNP probabilities as covariates. For QTL analyses based on regression models, the QTL

* Corresponding author: The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden.
E-mail: xia.shen@lcb.uu.se

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

effect is treated as fixed, and several studies have applied the idea of a full likelihood function (Elston & Stewart, 1971; Morton & Maclean, 1974; Lander & Botstein, 1989), which is referred to as the maximum likelihood (ML) method in QTL analysis and has been implemented in, for instance, MAPMAKER-QTL (Lander & Botstein, 1989). The implementation is based on an Expectation–Maximization (EM) algorithm (Dempster *et al*., 1977). However, ML estimates based on a regression model can be approximated very well by the simple Haley–Knott (HK) regression (Haley & Knott, 1992), which is the corresponding expectation method using line-origin probabilities as covariates.

In random effect models, the QTL effect is regarded as random, and considering a full likelihood method is still important to avoid losing statistical power (Schork, 1993). Replacing the IBD matrix using its expectation can only approximate the ML estimates, and the approximation was shown, by means of simulations, to be non-negligible in the analyses of sib-pairs (Kruglyak & Lander, 1995). To resolve these problems, a weighted likelihood approach has been implemented in the software package Mx (Eaves *et al*., 1996) for the analysis of small human pedigrees where the probability of IBD states are used as weights. However, knowing the distribution of the IBD matrix is crucial for deriving the full likelihood function. In human full-sib studies, the closed form of the joint distribution of the additive IBD matrix and the dominance IBD matrix has been derived, but this is feasible only for pedigrees including small full-sib families (Gessler & Xu, 1996; Xu, 1996). These earlier studies show that the full likelihood function is statistically more powerful and often gives higher likelihood at the QTL.

Three problems were raised from previous studies. First, for animal pedigrees, deriving the distribution of the IBD matrix is infeasible due to the large size. Therefore, approximating the full likelihood function using a Monte Carlo strategy is a reasonable idea but has not been implemented (Xu, 1996). Second, full-sib studies in humans calculate IBD probabilities for F_1 individuals. For a crossing design in animals, where e.g. F_2 individuals are studied, deriving the IBD distribution is difficult even for small pedigrees. An application of the distribution method to F_2 individuals has therefore not been investigated before, even though the theory was claimed to be able to extend to different kinds of crosses. Third, when there is inbreeding, for instance, in an F_2 intercross, diagonal elements of the IBD matrix need to be adjusted. After adjusting for inbreeding, the full likelihood theory still holds. If the marker density is low or the markers are partially informative, the difference between the distribution method and the expectation method might be substantial for experimental crosses as well.

The aim of our study is to evaluate the performance of the distribution method in animal intercross designs by assessing the magnitude and direction of bias for the expectation method. We try to account for the above problems and investigate the full likelihood function for animal pedigrees. The rest of this paper is arranged as follows. We first describe the statistical model that our study is based on and introduce the theory about the full likelihood. Two illustrative examples of F_2 pedigrees are simulated, where one is used to show the difference from full-sib studies, and the other is used to show the performance of the distribution method in adjusting the bias of heritability estimates. We compare the distribution method with the expectation method for a real experimental dataset, with simulations based on real genotypes for comparing the power of the two methods. The paper is concluded by discussing possible applications and suggesting future developments.

2. Methods and materials

(i) *Models and likelihoods*

We consider the variance component QTL model (Fernando & Grossman, 1989; Goldgar, 1990):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon, \tag{1}$$

where $\mathbf{y}$ is the trait response vector for N individuals, β is the fixed effect vector, γ is the multivariate normal-distributed random QTL effect with a zero mean and variance–covariance matrix $\mathbb{V}(\gamma) = \frac{1}{2}\sigma_g^2\mathbf{I}_q$ and ε is the normal-distributed error term with a zero mean and variance–covariance matrix $\mathbb{V}(\epsilon) = \sigma_e^2\mathbf{I}_N$. $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices. σ_g^2 is the genotypic variance and σ_e^2 is the residual variance. $\mathbf{Z}$ relates individuals and their inherited allelic substitution effects. Given the IBD matrix $\mathbf{\Pi}$, the variance–covariance matrix of the phenotype $\mathbf{y}$ is $\mathbf{V} = \mathbb{V}(\mathbf{y}) = \sigma_g^2\mathbf{\Pi} + \sigma_e^2\mathbf{I}$. The relationship between $\mathbf{\Pi}$ and $\mathbf{Z}$ is given by (Rönnegård & Carlborg, 2007):

$$\mathbf{\Pi} = \frac{1}{2}\mathbf{Z}\mathbf{Z}'. \tag{2}$$

To adjust the bias of variance component estimates due to estimating the fixed effects, restricted maximum likelihood (REML) is commonly used instead of ML. If $\theta = (\sigma_g^2, \sigma_e^2)'$ is the vector of variance components, the likelihood function is given by

$$\begin{aligned}\mathcal{L}(\theta|\Pi, \mathbf{y}) = {} & |2\pi\mathbf{V}|^{-1/2} \\ & \cdot \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})\right) \\ & \cdot \left|\frac{\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}}{2\pi}\right|^{-1/2}.\end{aligned} \tag{3}$$

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

If the alleles cannot be traced unambiguously through the pedigree, e.g. because the markers are not fully informative, the conditional expectation of $\mathbf{\Pi}$ has been used as a known matrix for the estimation of likelihood (3) (expectation method). Regarding $\mathbf{\Pi}$ as random, a joint distribution $f(\mathbf{y},\mathbf{\Pi}|\theta)$ is considered for estimating the likelihood (distribution method), i.e. the full likelihood. In this paper, the terms 'distribution method' and 'full likelihood method' are used interchangeably. ℓ is used to denote the logarithm of the corresponding likelihood $\mathcal{L}$.

Given the incomplete marker information, there is a probability space in which the IBD matrix $\mathbf{\Pi}$ is distributed. The expectation method uses an approximated likelihood

$$\mathcal{L}_{\mathrm{E}} = \mathcal{L}(\theta|\mathbf{y}, \mathbb{E}[\mathbf{\Pi}]), \tag{4}$$

where the variation of $\mathbf{\Pi}$ is not considered since $E[\mathbf{\Pi}]$ is inserted as a known matrix. Instead of calculating the expected IBD matrix $E[\mathbf{\Pi}]$, the full likelihood function takes the variation of $\mathbf{\Pi}$ into account by considering the joint distribution of θ and $\mathbf{\Pi}$. Inference of θ should be made from the marginal likelihood of θ integrating out $\mathbf{\Pi}$. Hence, based on profile likelihood (3), the distribution method uses the marginal likelihood

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} &= \mathcal{L}(\theta|\mathbf{y}) \\ &= \sum_{\mathbf{\Pi}} \mathcal{L}(\theta, \mathbf{\Pi}|\mathbf{y}) \\ &= \sum_{\mathbf{\Pi}} \mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi})\mathcal{P}(\mathbf{\Pi}) \\ &= \mathbb{E}_{\mathbf{\Pi}}[\mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi})].\end{aligned} \tag{5}$$

Thus, the difference between $\mathcal{L}_{\mathrm{E}}$ and $\mathcal{L}_{\mathrm{D}}$ is the difference between calculating the function of an expectation and calculating an expectation of the function. Non–linearity of function $\mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi})$ with respect to $\mathbf{\Pi}$ affects the difference between $\mathcal{L}_{\mathrm{E}}$ and $\mathcal{L}_{\mathrm{D}}$.

(ii) *Computation of the full likelihood*

Since the distribution of $\mathbf{\Pi}$ is rather complicated, marginal likelihood (5), involving an expectation with respect to $\mathbf{\Pi}$, is hardly derivable unless the number of individuals is extremely small. Therefore, we propose a Monte Carlo strategy that approximates likelihood (5) by

$$\tilde{\mathcal{L}}_{\mathrm{D}}(\theta|\mathbf{y}) \approx \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi}_i), \tag{6}$$

where m is the number of imputed IBD matrices drawn based on the marker information. Each impute $\mathbf{\Pi}_i$ corresponds to an incidence matrix $\mathbf{Z}_i$, and eqn (2) holds, so that $\frac{1}{2}\mathbf{Z}_i\mathbf{Z}_i' = \mathbf{\Pi}_i$. $\tilde{\mathcal{L}}_{\mathrm{D}}(\theta|\mathbf{y})|_{\theta=\hat{\theta}}$ converges to $\mathcal{L}_{\mathrm{D}}(\theta|\mathbf{y})|_{\theta=\hat{\theta}}$ as $m\to\infty$, where $\hat{\theta}$ is the ML estimate of $\mathcal{L}_{\mathrm{D}}(\theta|\mathbf{y})$.

The estimate of θ is identical for all the imputes of $\mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi}_i)$, namely, instead of maximizing each imputed likelihood, the entire sum $\sum_{i=1}^{m}\mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi}_i)$ needs to be maximized. The first and second derivatives of $\log \mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi}_i)$ with respect to θ have closed solutions (Harville, 1977). Let $\ell = \log \tilde{\mathcal{L}}(\theta|\mathbf{y})$ and $\ell_i = \log \mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi}_i)$. ℓ is the target log-likelihood to maximize. Using the derivatives $\partial\ell_i/\partial\theta$ and $\partial^2\ell_i/\theta\partial\theta'$, a Newton–Raphson-based EM algorithm can be used to estimate θ.

Algorithm. *Given a set of imputed IBD matrices (or corresponding incidence matrices), estimation of variance components θ using the full likelihood function can be made via the following steps:*

(i) *Find an initial estimate* $\hat{\theta}_{\mathbf{0}}$.
(ii) *Loop on k until convergence.*

$$\hat{\theta}_k = \hat{\theta}_{k-1} - \delta\left(\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right)^{-1}_{\theta=\hat{\theta}_{k-1}}\left(\frac{\partial\ell}{\partial\theta}\right)_{\theta=\hat{\theta}_{k-1}}, \tag{7}$$

where δ is a step size constant and

$$\frac{\partial\ell}{\partial\theta} = \sum_{i=1}^{m} w_i\frac{\partial\ell_i}{\partial\theta}, \tag{8}$$

$$\begin{aligned}\frac{\partial^2\ell}{\partial\theta\partial\theta'} = \sum_{i=1}^{m} w_i\left(\left(\frac{\partial\ell_i}{\partial\theta}\right)\left(\frac{\partial\ell_i}{\partial\theta}\right)' + \frac{\partial^2\ell_i}{\partial\theta\partial\theta'}\right) \\ -\left(\frac{\partial\ell}{\partial\theta}\right)\left(\frac{\partial\ell}{\partial\theta}\right)'.\end{aligned} \tag{9}$$

In eqns (8) and (9), the weights are defined as

$$w_i = \frac{\mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi}_i)}{\sum_{i=1}^{m}\mathcal{L}(\theta|\mathbf{y}, \mathbf{\Pi}_i)}. \tag{10}$$

(iii) *Take the converged estimate $\hat{\theta}$ as the final variance component estimates.*

This algorithm is based on Newton–Raphson iterations but is an EM algorithm since the weights in gradient and hessian need to be updated using the current variance component estimates. There are two advantages with this implementation. First, imputing incidence matrices ($\mathbf{Z}_i$) is much easier than creating the IBD matrices ($\mathbf{\Pi}_i$) that is not required in the proposed algorithm. Second, for large pedigrees with a few founders, when the marker is partially informative, the rank of $\mathbb{E}[\mathbf{\Pi}]$ is much greater than the rank of $\mathbf{Z}_i$. A low rank of $\mathbf{Z}_i$ increases the computational efficiency for maximizing the likelihood function (Rönnegård *et al.*, 2007).

(iii) *Simple illustrative examples*

Two pedigrees were simulated for showing different properties of the method. Pedigree I was simulated for showing that the non-negligible bias from the

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

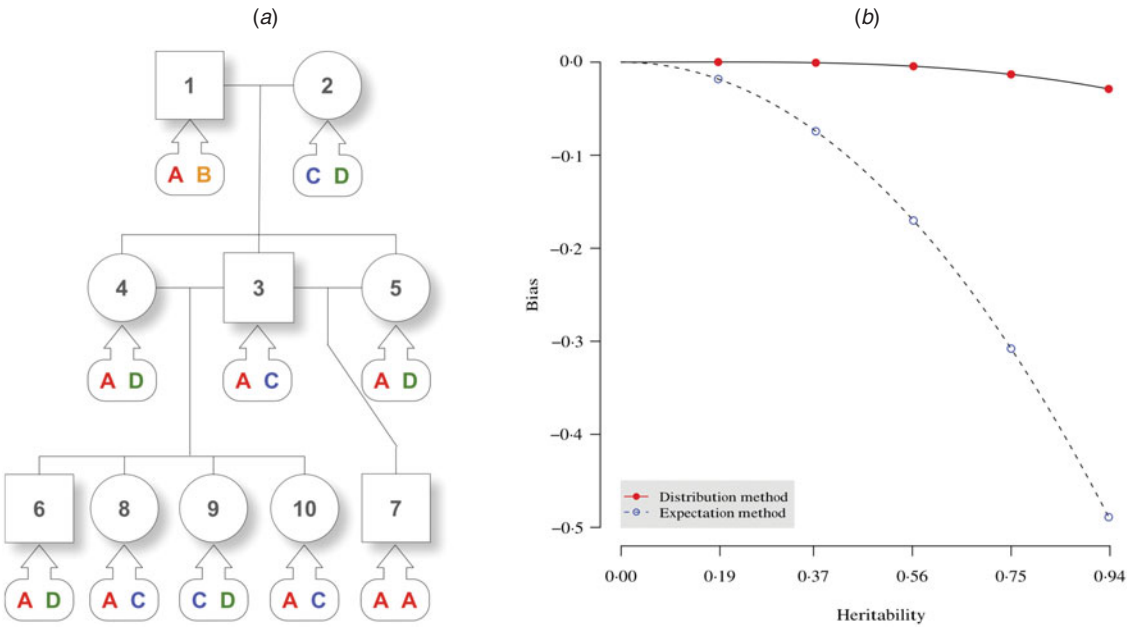

Fig. 1. Pedigree I – a simulated F_2 intercross with 10 individuals including two founders. (*a*) A three-generation intercross with five offspring, three parents and two grandparents. Squares and circles denote male and female animals, respectively, with indices inside. Bubbles with arrows pointing to each animal indicate the true genotypes that are not observed. (*b*) For the distribution method and the expectation method, asymptotic trend of bias in estimating the narrow sense heritability is displayed with respect to the heritability of the trait.

expectation method could be adjusted by the distribution method. Pedigree II was simulated for showing that the proven relationship between the likelihood estimate from the expectation method and that from the distribution method for full-sib families is not true for F_2 intercrosses.

(a) *Pedigree I*

In an F_2 intercross (Fig. 1*a*), a single pair of grandparents were mated to produce one male and two female F_1 progeny. They were thereafter mated to obtain five F_2 offspring. At a putative QTL, genotypes were simulated for each individual, and there are four alleles (A, B, C and D) throughout the pedigree. The phenotypic value for individual i was simulated by

$$y_i = \mu + \gamma_{i1} + \gamma_{i2} + \epsilon_i, \tag{11}$$

where $\mu = 50$, $\epsilon_i \sim \mathcal{N}(0, 1)$, and γ_{i1} and γ_{i2} correspond to the paternal and maternal allele substitution effects. Five sets of allele substitution effects were simulated according to five different σ_g^2 values. For instance, given the sample variance of $\sigma_g^2 = 15$, the allelic effects can be assigned as $\gamma_A = 3$, $\gamma_B = 6$, $\gamma_C = 9$ and $\gamma_D = 12$, which give a consistent estimate for the genetic variance, and the simulated phenotypic values were 64·87, 56·21, 61·28, 69·20 and 62·08 for individuals 6–10, respectively, for this particular set of allelic effects. In Fig. 1*a*, the kinship information is known but the genotypes are not observed. The (narrow sense) heritability (Lynch & Walsh, 1997) of the studied trait is defined as the intra-class correlation by

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}, \tag{12}$$

which measures how large proportion of the trait variation is determined by inheritance. We used both the expectation and the distribution method to estimate h^2 using the five animals in the F_2 generation. Note that equivalently, this example compares the variance component estimates using the two methods given an uninformative marker (no marker information) in QTL analyses.

(b) *Pedigree II*

One male was mated to three females to produce six F_1 individuals, and the F_1s were mated to obtain 10 F_2 offspring (Fig. 2). At a putative QTL, allele substitution effects for the eight alleles of the founders were simulated from a normal distribution with a zero mean and a standard deviation of 3, and they were inherited through the pedigree (Table 1). The phenotypes were calculated by summing the allele substitution effects, an overall mean of 200 and an error term drawn from $\mathcal{N}(0, 1)$. Assuming a complete link to the QTL, two sets of marker genotypes were simulated (Markers I and II). We computed the log-likelihood function using both the expectation and the distribution method.

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

Table 1. *The tabular form of pedigree II. Two markers that have a complete link to the QTL were simulated.* ℓ_D *and* ℓ_E *are log-likelihood from the distribution method and the expectation method, respectively. Given marker I,* $\ell_D > \ell_E$*, while given marker II,* $\ell_D < \ell_E$.

				Marker genotypes			True allelic effects	
ID	Sire ID	Dam ID	Sex	Marker I	Marker II	Phenotypic values	Paternal	Maternal
1	0	0	M	*gg*	*Gg*	188·5349	−6·5498	−5·4410
2	0	0	F	*gg*	*Gg*	194·1120	−4·6096	1·6403
3	0	0	F	*GG*	*gg*	202·1803	1·6783	1·7518
4	0	0	F	*GG*	*GG*	210·5896	4·3443	6·4353
5	1	2	M	*gg*	*GG*	195·1616	−6·5498	1·6403
6	1	4	M	*Gg*	*Gg*	200·2790	−5·4410	6·4353
7	1	4	F	*Gg*	*GG*	198·2070	−6·5498	4·3443
8	1	3	F	*Gg*	*Gg*	194·9456	−6·5498	1·7518
9	6	8	M	*gg*	*Gg*	191·1532	−5·4410	−6·5498
10	5	7	F	*gg*	*GG*	194·0887	1·6403	−6·5498
11	5	7	F	*gg*	*GG*	193·7124	1·6403	−6·5498
12	6	7	M	*GG*	*GG*	211·0225	6·4353	4·3443
13	6	8	F	*Gg*	*gg*	197·7793	−5·4410	1·7518
14	6	8	F	*GG*	*GG*	208·6419	6·4353	1·7518
15	5	8	F	*gg*	*GG*	195·7440	1·6403	−6·5498
16	6	7	M	*Gg*	*Gg*	199·2124	−5·4410	4·3443
17	5	8	M	*gg*	*GG*	193·6632	1·6403	−6·5498
18	6	8	F	*Gg*	*GG*	200·4391	6·4353	−6·5498
PIC				0·3515	0·3047			
ℓ_D				−15·4621	−25·4558			
ℓ_E				−16·9452	−21·6901			

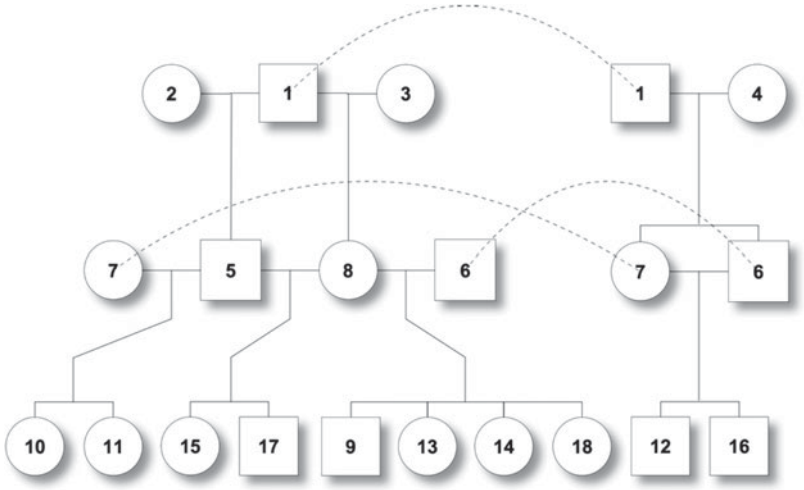

Fig. 2. Pedigree II – a simulated F_2 intercross with 18 individuals including four founders. Squares and circles denote male and female animals, respectively, with indices inside. Each dashed curve connects the same individual for the purpose of clear display.

(iv) *Analyses of experimental data*

(a) *Simulation using real genotypes*

An F_2 cross was bred from two European wild boars mated to eight large white sows (Andersson *et al.*, 1994). Four F_1 boars were mated to 22 F_1 sows to produce 191 recorded F_2 offspring in 26 families. The genetic information on chromosome 6 came from 22 genotyped micro-satellite markers at 0·0, 8·6, 36·6, 49·7, 50·5, 62·9, 79·2, 80·4, 83·7, 84·1, 84·8, 90·6, 95·4, 100·7, 101·9, 115·9, 116·7, 119·0, 120·2, 124·0, 127·0 and 170·9 cM. In order to investigate the power of the distribution method and the expectation method for this real dataset, we simulated a QTL at 25 cM harboured by the two flanking markers at 8·6 and 36·6 cM. This simulated QTL position has low marker information since it is in the middle of the long-interval between two markers with low information

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

content. IBD probabilities at the QTL were calculated taking into account all the marker information across the chromosome. We simulated the phenotype under three different scenarios for the heritability, 90%, 10% and 0%. For each scenario, 1000 replicates were simulated, where for each replicate, we calculated the LRT statistics both for the expectation and the distribution methods.

(b) *Analysis of a real phenotype*

A meat quality trait (reflectance value, EEL) recorded in the pig cross is strongly affected by the halothane gene located on chromosome 6 at position 80·4 cM. One of the founder boars was heterozygote (Hal^N/Hal^n) for this gene while all the other founders were homozygotes (Hal^N/Hal^N) for the wild-type allele. In our analyses, we include sex, litter and slaughter weight as fixed effects (Knott *et al*., 1998). The causal mutation underlying this QTL had previously been evidenced by Lundström *et al*. (1995), and the dataset is here used to illustrate the properties of both the expectation and the distribution method. The statistic used for QTL scan was the likelihood-ratio test (LRT) statistic that which is equivalentto the maximized likelihood function value but always non-negative. A permutation test was performed to obtain a significance threshold.

To understand the informativeness of each marker, the polymorphism information content (PIC) was used and determined using the following equation:

$$\mathrm{PIC}=1-\sum_{i=1}^{t}p_i^2-2\left(\sum_{i=1}^{t-1}\sum_{j=i+1}^{t}p_i^2p_j^2\right), \tag{13}$$

where p_i is the frequency of the ith allele and t is the number of alleles (Botstein *et al*., 1980).

3. Results and discussion

(i) *Simple illustrative examples*

(a) *Pedigree I*

The conventional expectation method estimates h^2 using likelihood (4), where the variance–covariance matrix of the genetic random effect is the relationship matrix in this example, since it is assumed that no marker information is known. Hence, for the F_2 individuals the expected IBD matrix is

$$\mathbb{E}[\mathbf{\Pi}]=\begin{pmatrix}1{\cdot}250 & 0{\cdot}625 & 0{\cdot}750 & 0{\cdot}750 & 0{\cdot}750\\ 0{\cdot}625 & 1{\cdot}250 & 0{\cdot}625 & 0{\cdot}625 & 0{\cdot}625\\ 0{\cdot}750 & 0{\cdot}625 & 1{\cdot}250 & 0{\cdot}750 & 0{\cdot}750\\ 0{\cdot}750 & 0{\cdot}625 & 0{\cdot}750 & 1{\cdot}250 & 0{\cdot}750\\ 0{\cdot}750 & 0{\cdot}625 & 0{\cdot}750 & 0{\cdot}750 & 1{\cdot}250\end{pmatrix} \tag{14}$$

The simulated true value of h^2 is given by $s_\gamma^2/(s_\gamma^2+\sigma_e^2)$, where $\gamma=(\gamma_A, \gamma_B, \gamma_C, \gamma_D)'$, i.e. the vector of allelic effects, and s represents the standard deviation.

Table 2. *Heritability estimates for pedigree I. Using the distribution method, variance components were estimated with different number of Monte Carlo imputes. Compared to the simulated true value, the heritability estimated by the distribution method had much less bias than that estimated by the expectation method.*

Method	No. imputes	Estimated h^2*	Bias
Distribution method	100	0·8462	−0·0913
	200	0·8669	−0·0706
	500	0·8829	−0·0546
	1000	0·8955	−0·0420
	2000	0·9037	−0·0338
	5000	0·9086	−0·0289
	10 000	0·9087	−0·0288
Expectation method	—	0·4485	−0·4890
Simulated value	—	0·9375	—

* Narrow sense heritability, defined as $\sigma_g^2/(\sigma_g^2+\sigma_e^2)$.

The aim here was to estimate the heritability h^2 by estimating the variance components σ_g^2 and σ_e^2. For the simulation with $\sigma_g^2=15$, the simulated true h^2 is 0·9375. The estimate by the distribution method converged to 0·9087 using 10 000 Monte Carlo imputes, and the bias was therefore, −0·0288 (Table 2). The expectation method using the expected IBD matrix (14) gave an estimate of 0·4485 and a bias of −0·4890. The comparison of all the five sets of simulated allelic effects shows an asymptotic trend of the bias by both methods, with respect to heritability (Fig. 1*b*). The results show that the difference between the two methods increases with the heritability, and this is consistent with the comparison done in full-sib families (Xu, 1996). From this example, the distribution method tends to be robust in variance component estimation and more reliable than the expectation method. For large sample or low heritability, the difference between the two methods will become small; however, the distribution method still gains power, which we show in the analyses of experimental data.

(b) *Pedigree II*

The purpose of this simulated example was to show that the relationship $\ell_D>\ell_E$ that has been proved for full-sib families (Xu, 1996) does not always hold, not to compare the two methods. We show here that this inequality does not hold for F_2 intercross designs. Based on the information of *marker I*, the maximized log-likelihood value of the distribution method was greater than that of the expectation method.

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

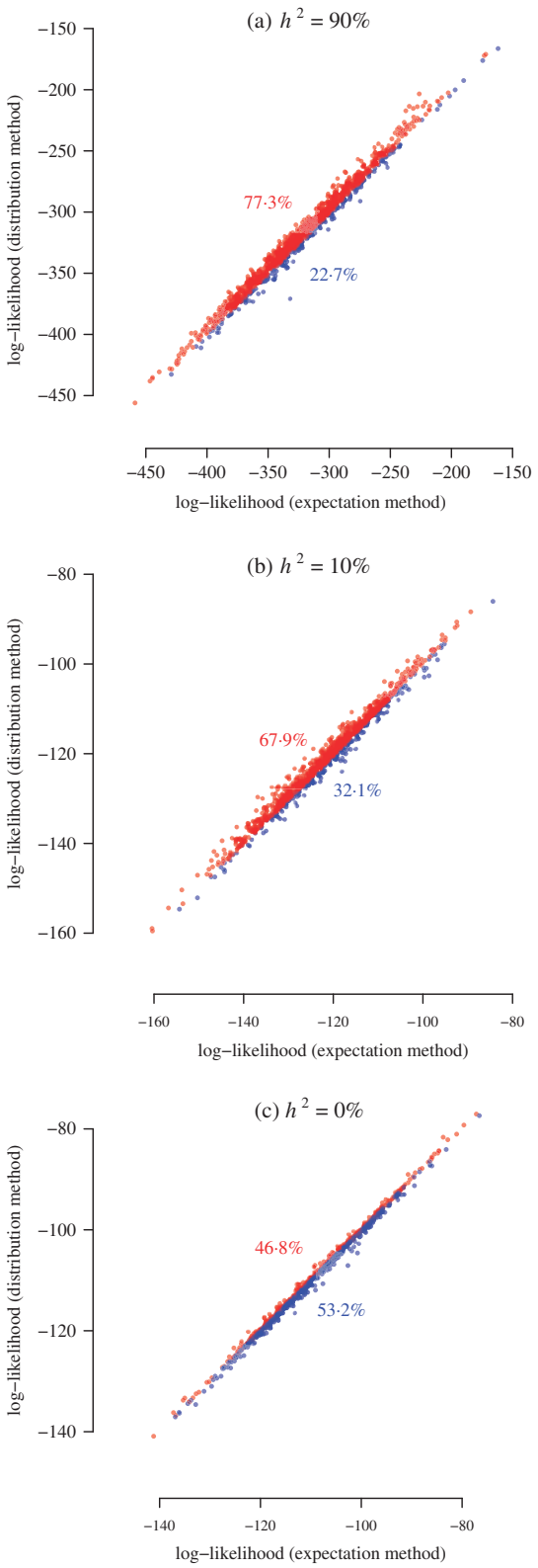

Fig. 3. Simulation results for power of interval mapping. A QTL was simulated at 25 cM of pig chromosome 6. The two markers flanking the interval harbouring the QTL are located at 8·6 cM and 36·6 cM. The real experimental genotypes for 191 F_2 individuals were used to simulate phenotypes, assuming a narrow sense heritability of 0·9, 0·1, and 0. 1000 simulations were used for comparing the log-likelihood from the expectation and the distribution methods. The points above/below the diagonal are in red/blue, indicating that the distribution method has larger power than the expectation method (*a*, *b*), or that the distribution method has lower false positive rate than the expectation method (*c*). The numbers in colour show the corresponding percentages of the sets of points.

However, the inequality was reversed for *marker II* (Table 1).

It has been claimed that the distribution method is more powerful in QTL detection due to larger likelihood estimates. It is not always true that larger likelihood estimates will lead to more powerful QTL detection, because the likelihood estimates for non-QTL positions might also be inflated. Thus, a QTL detection method gains power when the interval harbouring the QTL produces a relatively larger likelihood than those intervals not harbouring any QTL. As shown above, the inequality $\ell_D > \ell_E$ does not necessarily hold in F_2 intercrosses, but this does not mean that the expectation method is potentially more powerful in QTL detection. In order to assess this question for the two methods, a larger simulation and a full genome scan need to be performed (see the next subsection).

(ii) *Analyses of experimental data*

(a) *Simulation using real genotypes*

For each of the 1000 replicates, we compared the log-likelihood values (equivalent to LRT statistics) from the two methods, namely, the power of detecting the simulated QTL harboured by two flanking partially informative markers (Fig. 3). Since the sample size in this simulation was much bigger than the previous two illustrative examples, the approximation of the expectation method was better so that the result from it is closer to the distribution method. However, for a heritable trait, the distribution method is more powerful than the expectation method (Fig. 3*a*, *b*). When the heritability is low (10%), the distribution method still gains power (Fig. 3*b*). When the heritability is 0% (no QTL effect, non-heritable), the distribution method produces significantly less (*P*-value $= 2{\cdot}3 \times 10^{-14}$ from a Wilcoxon test) false positives than the expectation method (Fig. 3*c*). This suggests that the distribution method has a tendency to improve location accuracy in a variance component QTL scan, which is also found in the analysis of a real phenotype (see below).

(b) *Analysis of a real phenotype*

A variance-component–based QTL scan was performed on chromosome 6 for the European wild boar × large white intercross (Fig. 4). Both the

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

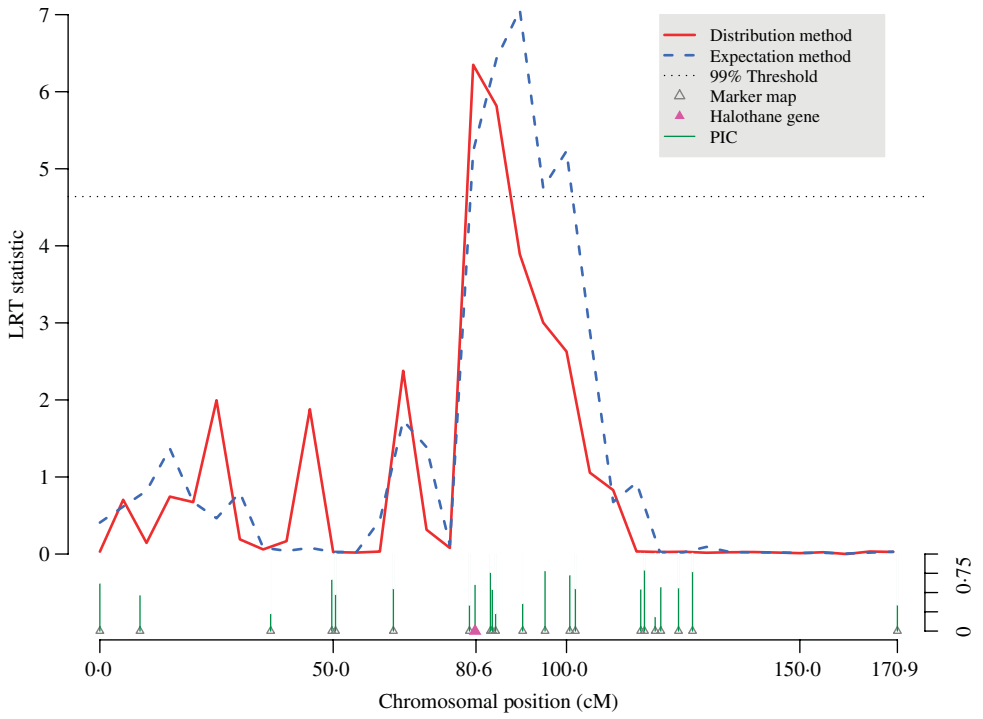

Fig. 4. QTL scan using LRT statistic along pig chromosome 6. The meat quality trait (reflectance value, EEL) is strongly affected by the halothane gene located at 80·4 cM on the chromosome. By adjusting bias of likelihood estimates, thedistribution method refined the peak of the traditional variance component QTL scan using the expectation method and thereby shortened the confidence interval for the QTL. Information for each micro-satellite marker is shown as their PIC.

expectation and the distribution method detected a significant signal but with different intervals harbouring the QTL. The conventional expectation method provided an interval longer than 20 cM, but the scan using the full likelihood function refined the peak and shortened the interval down to around 10 cM.

As shown in our first illustrative example above, the expectation method has a tendency of underestimating the genetic variance at a QTL. Furthermore, it might also overestimate the genetic variance at tested loci with no QTL that are linked to the QTL. Based on our simulations, analyses and the results from previous studies (Gessler & Xu, 1996), we conclude that the expectation method is a compromising approximation, which loses power of localizing QTL compared with the full likelihood method.

(iii) *Further extensions and developments*

Epistasis has been emphasized to be common and important in genetic control of complex traits (Carlborg & Haley, 2004). A potential extension for our implemented algorithm is to extend it for detecting epistasis, and the idea is straightforward. Consider the variance component model including epistatic effects as

$$\mathbf{y}=\mathbf{X}\beta+\mathbf{Z}_{\mathrm{A}}\mathbf{v}_{\mathrm{A}}+\mathbf{Z}_{\mathrm{B}}\mathbf{v}_{\mathrm{B}}+\mathbf{Z}_{\mathrm{AB}}\mathbf{v}_{\mathrm{AB}}+\varepsilon, \tag{15}$$

where $\mathbf{v}_{\mathrm{A}}$ and $\mathbf{v}_{\mathrm{B}}$ are the main random QTL effects of loci A and B, with $\mathbb{C}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}})=0$, and $\mathbf{v}_{\mathrm{AB}}$ is the random interaction effect, where $\mathbb{C}$ denotes covariance. The IBD matrix $\mathbf{\Pi}_{\mathrm{AB}}$ for the epistatic effects in the variance structure of model (19), $\mathbb{V}(\mathbf{y})=\mathbf{\Pi}_{\mathrm{A}}\sigma^2_{\mathrm{A}}+\mathbf{\Pi}_{\mathrm{B}}\sigma^2_{\mathrm{B}}+\mathbf{\Pi}_{\mathrm{AB}}\sigma^2_{\mathrm{AB}}+\mathbf{I}\sigma^2_e$, can be calculated as the Hadamard product of the two IBD matrices at loci A and B (Stern *et al.*, 1996), i.e. $\mathbf{\Pi}_{\mathrm{AB}}=\mathbf{\Pi}_{\mathrm{A}}\circ\Pi_{\mathrm{B}}$. If the expectation method is applied, $\mathbf{\Pi}_{\mathrm{AB}}$ is estimated as $\mathbb{E}[\mathbf{\Pi}_{\mathrm{A}}]\circ\mathbb{E}[\mathbf{\Pi}_{\mathrm{B}}]$, which is not always a good approximation if the two loci are linked (Rönnegård *et al.*, 2008). Using our proposed algorithm, after getting a number of Monte Carlo imputes for the IBD matrices atloci A and B, the full likelihood can be approached by

$$\mathbb{E}_{\mathbf{\Pi}_{\mathrm{A}}\circ\mathbf{\Pi}_{\mathrm{B}}}[\mathcal{L}(\theta|\mathbf{y},\mathbf{\Pi}_{\mathrm{A}}\circ\mathbf{\Pi}_{\mathrm{B}})]\approx\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\theta|\mathbf{y},\mathbf{\Pi}_{\mathrm{A}i}\circ\mathbf{\Pi}_{\mathrm{B}i})\,. \tag{16}$$

Substituting the sum into eqn (6), the same EM algorithm can be applied.

In the sample space of $\mathbf{\Pi}$, there exists a model-based 'best' IBD matrix based on the maximum full likelihood. This implies, after profiling out (estimating and inserting back into likelihood) the fixed effects, as in eqn (3), that the best IBD matrix and its corresponding estimate of $\hat{\theta}$ maximize the joint

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

likelihood $\mathcal{L}(\theta, \mathbf{\Pi}|\mathbf{y})$. By maximizing the joint likelihood with respect to θ and $\mathbf{\Pi}$ simultaneously, it is possible to infer the best IBD matrix and thereby the best variance component estimates. Nevertheless, our algorithm is not implemented for maximizing the joint likelihood but the marginal. More advanced methods, e.g. those used in phylogenetic tree estimation, using the optimality criterion of ML, often under a Bayesian framework, might be utilized to estimate variance components and the incidence matrix jointly (Felsenstein, 2004). Identifying the optimal IBD matrix is NP-hard[1] using these methods, and so heuristic search and optimization methods most likely need to be used to identify a reasonably good incidence matrix that fits the data. In genetic applications, using Bayesian methods, Gibbs sampling can be used to estimate the incidence matrix; however, it is not guaranteed that the chain is irreducible in large complex problems, and even if the chain is irreducible, mixing can be quite slow (Sorensen & Gianola, 2002). Bayesian computation can be introduced when it is possible and reasonable to joint-estimate the model, otherwise, inference should focus on the variance components. Hence, we suggest that our suggested full likelihood method should be used rather than Bayesian joint estimation of variance components and the true IBD matrices at each putative QTL position.

In principle, the distribution method derived in this paper can be extended to any general pedigree or population, as long as the genotype probabilities can be calculated. Namely, the algorithm itself has no assumption on the design. Using the genotype probabilities calculated from half-sib designs, backcrosses, advanced intercross lines, or even imputed SNP data from population–based studies, the distribution method can be applied in all situations. The key motivation for using the distribution method is to deal with genotype uncertainty. In experimental crosses, the Monte Carlo step samples IBD matrices, whereas for imputed data in GWAS, we might sample IBS (identity-by-state) matrices. For GWAS using real–typed SNPs, there is no need to apply a full likelihood method as there is no difference between the distribution method and the expectation method when the real genotypes are observed.

4. Conclusion

We have evaluated a full likelihood approach based on variance component QTL models for intercross populations. By means of simulation, we have shown that the full likelihood method is able to correct bias of the traditional variance component method using expected IBD matrices. Also, we used simulations to show that the previously reported relationship between the likelihoods of the distribution and expectation methods does not always hold in general pedigrees. The full likelihood method was compared with the expectation method in experimental cross data, and the former was found to be able to improve the precision of QTL detection. The algorithm described in this paper has been implemented in a package written in R (R Development Core Team, 2010) and is available on request.

[1] Non-deterministic polynomial-time hard: a term indicating high computational difficulty in computational complexity theory (Garey & Johnson, 1979).

Xia Shen is funded by a Future Research Leaders grant from Swedish Foundation for Strategic Research (SSF) to Örjan Carlborg. Lars Rönnegård is funded by the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS). We thank Carl Nettelblad for providing his programme particularly for our implementation. Leif Andersson is acknowledged for sharing the experimental pig data.

References

Andersson, L., Haley, C., Ellegren, H., Knott, S., Johansson, M., Andersson, A.-E. L., Edfors-Lilja, K. I., Fredholm, M., Hansson, I., Håkansson, J. & Lundström, K. (1994). Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* **263**, 1771–1774.

Arendonk, V., Tier, B. J. A. M. & Kinghorn, B. (1994). Use of multiple genetic markers in prediction of breeding values. *Genetics* **137**, 319–329.

Besnier, F. & Carlborg, Ö. (2007). A general and efficient method for estimating continuous IBD functions for use in genome scans for QTL. *BMC Bioinformatics* **8**, 440.

Blangero, J., Williams, J. & Almasy, L. (2001). Variance component methods for detecting complex trait loci. *Advances in Genetics* **42**, 151–181.

Botstein, D., White, R., Skolnick, M. & Davis, R. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* **32**, 314–331.

Carlborg, Ö. & Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies? *Nature Reviews, Genetics* **5**, 618–625.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* **39**, 1–38.

Eaves, L., Neale, M. & Maes, H. (1996). Multivariate multipoint linkage analysis of quantitative trait loci. *Behavior Genetics* **26**, 519–525.

Elston, R. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Fernando, R. & Grossman, M. (1989). Marker-assisted selection using best linear unbiased prediction. *Genetics, Selection, Evolution* **21**, 467–477.

Fulker, D. & Cardon, L. (1994). A sib-pair approach to interval mapping of quantitative trait loci. *American Journal of Human Genetics* **54**, 1092–1103.

Garey, M. R. & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA: W. H. Freeman.

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press

George, A. W., Visscher, P. & Haley, C. (2000). Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics* **156**, 2081–2092.

Gessler, D. D. G. & Xu, S. (1996). Using the expectation or the distribution of the identity by descent for mapping quantitative trait loci under the random model. *American Journal of Human Genetics* **59**, 1382–1390.

Goddard, M. (1992). A mixed model for analyses of data on multiple genetic markers. *Theory and Applied Genetics* **83**, 878–886.

Goldgar, D. (1990). Multipoint analysis of human quantitative genetic variation. *American Journal of Human Genetics* **47**, 957–967.

Haley, C. & Knott, S. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338.

Knott, S., Marklund, L., Haley, C., Andersson, K., Davies, D., Ellegren, H., Fredholm, M., Hansson, I., Hoyheim, B., Lundström, K., Moller, M. & Andersson, L. (1998). Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* **149**, 1069–1080.

Kruglyak, L. & Lander, E. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.

Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., Waterworth, D., Beckmann, J. S. & Bergmann, S. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics* **12**, 1–17.

Lander, E. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lundström, K., Karlsson, A., Håkansson, J., Hansson, I., Johansson, M., Andersson, L. & Andersson, K. (1995). Production, carcass and meat quality traits of f2-crosses between european wild pigs and domestic pigs including halothane gene carriers. *Animal Science* **61**, 325–331.

Lynch, M. & Walsh, B. (1997). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Mao, Y. & Xu, S. (2005). A Monte Carlo algorithm for computing the IBD matrices using incomplete marker information. *Heredity* **94**(3), 305–315. doi: 10.1038/sj.hdy.6800564.

Marchini, J. & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews, Genetics* **11**, 499–511. doi: 10.1038/nrg2796.

Morton, N. & Maclean, C. (1974). Analysis of family resemblance. III. complex segregation of quantitative traits. *American Journal of Human Genetics* **26**, 489–503.

Olson, J. (1995). Robust multipoint linkage analysis: an extension of the Haseman-Elston method. *Genetics Epidemiology* **12**, 177–193.

Pérez-Enciso, M., Varona, L. & Rothschild, M. (2000). Computation of identity by descent probabilities conditional on DNA markers via a Monte Carlo Markov Chain method. *Genetics, Selection, Evolution* **32**, 467–482.

Pong-Wong, R., George, A., Woolliams, J. & Haley, C. (2001). A simple and rapid method for calculating identity-by-descent matrices using multiple markers. *Genetics, Selection, Evolution* **33**, 453–471.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org. ISBN 3-900051-07-0.

Rönnegård, L. & Carlborg, Ö. (2007). Separation of base allele and sampling term effects gives new insights in variance component QTL analysis. *BMC Genetics* **8**, 1. doi: 10.1186/1471-2156-8-1.

Rönnegård, L., Mischenko, K., Holmgren, S. & Carlborg, Ö. (2007). Increasing the efficiency of variance component quantitative trait loci analysis by using reduced-rank identity-by-descent matrices. *Genetics* **176**, 1935–1938. doi: 10.1534/genetics.107.071977.

Rönnegård, L., Pong-Wong, R. & Carlborg, Ö. (2008). Defining the assumptions underlying modeling of epistatic QTL using variance component methods. *The Journal of Heredity* **99**, 421–425. doi: 10.1093/jhered/esn017.

Schork, N. J. (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: Efficiency, power, and modeling considerations. *American Journal of Human Genetics* **53**, 1306–1319.

Sorensen, D. & Gianola, D. (2002). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Berlin: Springer.

Stern, M., Duggirala, R., Mitchell, B., Reinhart, L., Sivakumar, S., Shipman, P., Uresandi, O., Benavides, E., Blangero, J. & O'Connell, P. (1996). Evidence for linkage of regions on chromosome 6 and 11 to plasma glucose concentrations in Mexican Americans. *Genome Research* **6**, 724–734.

Thompson, E. A. & Heath, S. C. (1999). Estimation of conditional multilocus gene identity among relatives. *Lecture Notes-Monograph Series* **33**, 95–113.

Wang, T., Fernando, R., van der Beek, S., Grossman, M. & van Arendonk, J. (1995). Covariance between relatives for a marked quantitative trait locus. *Genetics, Selection, Evolution* **27**, 251–274.

Xu, S. (1996). Computation of the full likelihood function for estimating variance at a quantitative trait locus. *Genetics* **144**, 1951–1960.

Xu, S. & Atchley, W. R. (1995). A random model approach to interval mapping of quantitative trait loci. *Genetics* **141**, 1189–1197.

https://doi.org/10.1017/S0016672311000152 Published online by Cambridge University Press