Theor Appl Genet (2012) 125:1727–1734
DOI 10.1007/s00122-012-1949-9

ORIGINAL PAPER

A fast expectation-maximum algorithm for fine-scale QTL mapping

Ming Fang

Received: 14 April 2012 / Accepted: 15 July 2012 / Published online: 4 August 2012

Abstract The recent technology of the single-nucleotide-polymorphism (SNP) array makes it possible to genotype millions of SNP markers on genome, which in turn requires to develop fast and efficient method for fine-scale quantitative trait loci (QTL) mapping. The single-marker association (SMA) is the simplest method for fine-scale QTL mapping, but it usually shows many false-positive signals and has low QTL-detection power. Compared with SMA, the haplotype-based method of Meuwissen and Goddard who assume QTL effect to be random and estimate variance components (VC) with identity-by-descent (IBD) matrices that inferred from unknown historic population is more powerful for fine-scale QTL mapping; furthermore, their method also tends to show continuous QTL-detection profile to diminish many false-positive signals. However, as we know, the variance component estimation is usually very time consuming and difficult to converge. Thus, an extremely fast EMF (Expectation-Maximization algorithm under Fixed effect model) is proposed in this research, which assumes a biallelic QTL and uses an expectation-maximization (EM) algorithm to solve model effects. The results of simulation experiments showed that (1) EMF was computationally much faster than VC method; (2) EMF and VC performed similarly in QTL detection power and parameter estimations, and both outperformed the paired-marker analysis and SMA. However, the power of EMF would be lower than that of VC if the QTL was multiallelic.

Communicated by M. Sillanpaa.

M. Fang (✉)
Life Science College, Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China
e-mail: fangming618@126.com

Introduction

Linkage analysis (LA) is an important tool for QTL mapping in which the recombination events within the pedigree provides enough information for localizing QTL. However, LA usually estimates QTL position within a large credible interval (say 10–20 cM) and fails to further fine localize QTL within the credible region even for high density markers, since the recombination event is rarely observed within the pedigree. Compared with LA, linkage disequilibrium (LD) mapping could utilize the recombination information beyond a pedigree and thus can narrow QTL position.

The single-marker analysis (SMA) is a simple method for fine-scale QTL mapping (e.g. Chen and Abecasis 2007; Wang et al. 2005), in which the marker can be directly used for testing the existence of a QTL. However, SMA tends to produce many spurious signals and show low QTL-detection power (Meuwissen and Goddard 2007). Compared with SMA, the haplotype-based methods can not only powerfully localize QTL, but also generate continue QTL-detection profiles to diminish many false-positive signals (Meuwissen and Goddard 2007). The haplotype-based methods were originally developed for fine mapping disease gene loci (Kaplan et al. 1995; Terwilliger 1995; Xiong and Guo 1997; McPeek and Strahs 1999; Morris 2006; Minichiello and Durbin 2006; Wellcome Trust Case Control Consortium 2007; Marchini et al. 2007; Kimmel et al. 2008). Since the haplotype carrying causative mutation will be decayed in the following generations due to recombination, the segment containing causative mutation will be narrowed down within a small region, which provides opportunities for fine localizing the causative mutation.

Fine-scale gene mapping for quantitative trait using haplotypes was investigated by Meuwissen and Goddard (2000, 2001), who used unknown historical recombination

information to calculate identity-by-descent (IBD) probabilities between two individual haplotypes; then construct IBD matrix at putative QTL locus and estimate QTL variance via restricted maximum likelihood method (REML, Patterson and Thompson 1971). Recently, several researchers have applied the Meuwissen and Goddard's method to search QTL for domestic animals (e.g. Druet et al. 2008; Schnabel et al. 2005). However, it is well known that the variance component estimation is usually computationally intensive and difficult to converge. For convenience, the algorithm has been built into the software GridQTL (Hernández-Sánchez et al. 2009) which distributes the analysis in parallel over a large public grid of computers, and thus could enhance computational speed dramatically.

Another strategy for fine-scale QTL mapping was proposed by Pérez-Enciso (2003) who assumed a biallelic QTL and estimated QTL substitution effect rather than QTL variance with a Markov Chain Monte Carlo (MCMC) algorithm. His method focuses on fine mapping along with haplotyping in a unified Bayesian framework. However, his method is still unsuitable for genome-wide QTL mapping due to the computational burden of the MCMC algorithm.

In this study, a fast expectation-maximization (EM) algorithm (Dempster et al. 1977) called EMF is developed, which assumes a biallelic QTL and uses EM algorithm to estimate QTL parameters. Since EMF avoids the IBD-matrices construction and REML estimation, much computational time can be saved. The efficiency of the method is illustrated with substantial simulation experiments.

Method

Model

Let $\mathbf{y}$ be an $n \times 1$ phenotypic vector,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}b + \mathbf{e} \quad (1)$$

where $\boldsymbol{\beta}$ is a vector of covariate effects and $\mathbf{X}$ is an incidence matrix; b is QTL additive effect and $\mathbf{w} = [w_1, \ldots, w_i, \ldots, w_n]^T$ is an $n \times 1$ vector of QTL genotype, where n is the number of phenotypic observations; $\mathbf{e}$ is the vector of random error, which follows normal distribution, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where $\mathbf{I}$ is an $n \times n$ identity matrix. The mutant and wild alleles are denoted by M and m, respectively. If the QTL allele on a haplotype of individual i is IBD to the original mutant allele on ancestral haplotype, the QTL allele is mutant type and denoted by M; otherwise, the QTL allele is wild type and denoted by m. By this definition, the QTL of individual i has three kinds of genotype, MM, Mm and mm, which are indicated by, $w_i = 2$, $w_i = 1$ and $w_i = 0$, respectively. The QTL dominant effect is assumed absent, but also can be easily included in model (1).

Inferring QTL-genotypic probability

EMF needs to calculate π_i^{P} (π_i^{M}), the probabilities that the QTL allele on paternal (maternal) haplotype of individual i is IBD to the original mutant QTL allele M arose T generations ago. In this research, the method of Meuwissen and Goddard (2001) for calculating IBD matrix between unrelated individuals is modified to infer π_i^{P} and π_i^{M}. For clear presentation, the inference of π_i^{P} and π_i^{M} is illustrated using only flanking markers, and it also can be adapted for the multi-marker situation. Given ancestral and flanking-marker haplotypes of individual i, π_i^{P} and π_i^{M} can be obtained according to Bayesian rule (Meuwissen and Goddard 2001),

$$p(\mathrm{IBD}|S_j, S_j) = \frac{p(S_j, S_{j+1}|\mathrm{IBD})p(\mathrm{IBD})}{p(S_j, S_{j+1}|\mathrm{IBD})p(IBD) + p(S_j, S_{j+1}|\mathrm{nonIBD})p(\mathrm{nonIBD})}, \quad (2)$$

where S_j is the indicator of the identity-by-state (IBS) status between individual and ancestral haplotype for jth marker, which equals to 1 or 0 indicating the IBS or nonIBS status of the two haplotypes; p(IBS) is the prior probability that the QTL allele is IBD to the QTL mutation M arose T generation ago. The details of the calculation of Eq. (2) are described in "Appendix 1", which is the same as the calculation of Eq. (2) in Meuwissen and Goddard (2001) except for f(c) and a_j. Assuming that the QTL loci is in Hardy–Weinberg equilibrium, the probability of the three QTL genotypes of individual i can be calculated as $P_{i2} = \pi_i^{\mathrm{P}}\pi_i^{\mathrm{M}}$ for MM, $P_{i1} = (1 - \pi_i^{\mathrm{P}})\pi_i^{\mathrm{M}} + \pi_i^{\mathrm{P}}(1 - \pi_i^{\mathrm{M}})$ for $Mm(mM)$, and $P_{i0} = (1 - \pi_i^{\mathrm{P}})(1 - \pi_i^{\mathrm{M}})$ for mm.

Maximum likelihood estimate via EM algorithm

Reconstruction of ancestral haplotype and estimation of QTL position

The possible QTL position is assumed to locate at the middle of each marker interval, and thus one should only scan the middle points of each marker interval to search QTL. The QTL position λ and the ancestral haplotype carrying QTL mutation M, h_{anc}, can be viewed as fixed parameters. Once one putative QTL is tested, all possible ancestral haplotypes should be tried for finding the most possible ancestral haplotype. For example, if only flanking markers is used and each has two alleles, a total of 4 (2×2) possible ancestral halpotypes should be tried. Then, the ancestral haplotype that generates the maximum log-likelihood ratio (LR) is the maximum likelihood estimate of ancestral haplotype at this position. Scanning all

QTL positions along genome, QTL position also can be estimated at the maximum *LR* score.

Estimation of model effect

Given the putative QTL position λ, one possible ancestral haplotype h_{anc}, the observable phenotypic values $\mathbf{y}$ and individual haplotypes $\mathbf{H}$, model effects $\boldsymbol{\theta} = (\boldsymbol{\beta},\ b,\ \sigma_e^2)^T$ can be estimated via the EM algorithm. Assuming individuals investigated are unrelated (the assumption can be relaxed and will be discussed later), the likelihood function can be expressed as

$$L(\boldsymbol{\theta}|\mathbf{y},\ \mathbf{H},\ \lambda,\ h_{\text{anc}}) = \prod_{i=1}^{n}\left(\sum_{k=0}^{2} P_{ik} f(y_i|w_i = k,\cdot)\right), \tag{4}$$

where $f(y_i|w_i{=}k,\cdot) \propto 1/\sigma_e \cdot \exp(-(y_i - \mathbf{X}_i\boldsymbol{\beta} - w_i b)^2/2{\sigma_e}^2)$, for $k = 0, 1$ and 2, is the likelihood conditional on three genotypes *mm, Mm* and *MM*, respectively. The parameters can be estimated via the EM algorithm (Dempster et al. 1977; Lander and Botstein 1989). The solutions can be expresses as

$$\hat{\boldsymbol{\beta}} = (\mathbf{y} - E(\mathbf{w}|\mathbf{y},\cdot)b)^T(\mathbf{y} - E(\mathbf{w}|\mathbf{y},\cdot)b)/n, \tag{5}$$

$$\hat{b} = \left(E(\mathbf{w}^T\mathbf{w}|\mathbf{y},\cdot)\right)^{-1}E(\mathbf{w}^T|\mathbf{y},\cdot)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{6}$$

$$\hat{\sigma}_e^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - E(\mathbf{w}|\mathbf{y},\cdot)b)^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - E(\mathbf{w}|\mathbf{y},\cdot)b). \tag{7}$$

In these equations, $E(\mathbf{w}|\mathbf{y},\cdot) = [E(w_1|y_1,\cdot)\ E(w_2|y_2,\cdot)\cdots E(w_n|y_n,\cdot)]^T$, $E(\mathbf{w}^T|\mathbf{y},\cdot) = [E(w_1|y_1,\cdot)\ E(w_2|y_2,\cdot)\cdots E(w_n|y_n,\cdot)]$, and $E(\mathbf{w}^T\mathbf{w}|\mathbf{y},\cdot) = \sum_{i=1}^{n} E(w_i^2|y_i,\cdot)$, where $E(w_i|y_i,\cdot) = 2P_{i2}^* + P_{i1}^*$ and $E(w_i^2|y_i,\cdot) = 4P_{i2}^{2*} + P_{i1}^{2*}$, where P_{ik}^* is the posterior probability of QTL genotype for $k = 2$ (*MM*), 1 (*Mm* and *mM*) and 0 (*mm*), respectively. According to Bayesian rule, P_{ik}^* can be expressed as

$$P_{ik}^* = \frac{P_{ik} f(y_i|w_i = k,\cdot)}{\sum_{k'=0}^{2} P_{ik'} f(y_i|w_i = k',\cdot)}. \tag{8}$$

These estimates can be found by iteration of the above equations via the EM algorithm. In each iteration, the algorithm consists of one E-step, Eq. (8), and three M-steps, Eqs. (5–7). The process is repeated until convergence. A statistical test for H_0: $b = 0$ is carried out by $\text{LR} = -2(\log(L_{\text{full}}/L_{\text{reduce}}))$, where $\log(L_{\text{full}})$ and $\log(L_{\text{reduce}})$ are the likelihoods under full model and reduced model ($b = 0$), respectively.

Estimation of QTL variance

Let σ_w^2 be the estimate of the variance of QTL genotype $\{w_i\}_{i=1}^n$, and then the estimate of QTL variance can be expressed as $\sigma_q^2 = b^2\sigma_w^2$. Since $\{w_i\}_{i=1}^n$ are unobservable, σ_w^2 cannot be estimated directly. In practice, one can approximately substitute $\{w_i\}_{i=1}^n$ with their posterior expectation $\{E(w_i|y_i,\cdot)\}_{i=1}^n$, and then σ_q^2 can be estimated with

$$\sigma_q^2 = b^2\left[\sum_{i=1}^{n}(E(w_i|y_i,\cdot))^2 - \left(\sum_{i=1}^{n} E(w_i|y_i,\cdot)\right)^2/n\right]/(n-1). \tag{9}$$

Extension to multilocus mapping

The above method only uses flanking-marker haplotypes to infer IBD probability, which also can be adapted for multilocus analysis where a set of markers surrounding the putative QTL are utilized. The differences between them mainly exist in the inference of IBD probability and the reconstruction of the ancestral haplotype. The IBD inference in multilocus analysis is also similar to that in Meuwissen and Goddard (2001), which also requires some modifications in f(c) and a, and both are the same as those in two-locus analysis (see "Appendix 1"). The ancestral haplotype also can be reconstructed by trying all possible haplotypes, which is similar to two-locus analysis. However, with the number of marker increase, the number of possible ancestral haplotypes will also exponentially increase, so that the program will be quickly forbidden. Therefore, a fast stepwise method, which is suitable for large number of markers, is developed here to solve the problem. The steps of the stepwise method are presented below:

Step 1 Initializing the ancestral haplotype. Scan QTL from the first marker interval to the last marker interval at the middle point using only two-locus EMF analysis to obtain the ancestral haplotype for each adjacent marker pair. Then the primitive ancestral haplotype for all markers can be generated by jointing each pair of haplotypes.

Step 2 Stepwise updating ancestral haplotype for each adjacent marker pair using multilocus information. Repeating two-locus analysis from the first marker interval to the last interval by trying all possible flanking-marker haplotypes, the IBD probability is inferred through the current multilocus haplotype rather than flanking-marker haplotype. When one marker interval is tested, the ancestral haplotype is updated at flanking markers, but unchanged for other loci. When QTL is scanned from the first marker to the last marker, the flanking-marker haplotype is updated pair by pair continuously until the last flanking-marker haplotype. In the end, the ancestral for all markers is also updated in this step.

Step 3 Repeat Step 2 until the ancestral haplotype for all markers is unchanged.

Testing the method

The study investigated a 2-cM region with 21 markers evenly spaced on the small region. One QTL was simulated at position 1.05 cM. The present population was built on a base population created 200 generations ago ($T = 200$) with effective population size $N_e = 200$ and sex ratio 1:1. No pedigree was recorded for this historical population. The frequencies of two alleles of each SNP marker in base population were both 0.5, and the marker alleles were mutated at a rate of 4×10^{-4}/generation. The QTL allele for each individual haplotype in base population was assigned a unique number. The QTL and marker alleles were transmitted to descendants according to Haldane's recombination rule. In the last generation, one of the QTL alleles that still existed with frequency (π) >0.1 and <0.9 was randomly sampled as being mutant allele and assigned effect b, while others were assumed to be the wild type and assigned effect 0. Then the QTL allele effect was determined from $\sigma_q^2 = 2b^2\pi(1-\pi)$. The residual effect was sampled from normal distribution with mean 0 and variance 0.9; the overall mean was set as zero, and no polygenic effect was simulated. The phenotypic value of each individual was simulated by summing the overall mean, QTL effect and residual values.

Six methods were compared, which included VC using 20 markers surrounding the putative QTL (VC20), VC using flanking markers (VC2), EMF using 20 markers (EMF20), EMF using flanking markers (EMF2), paired-marker analysis (MARK2) and SMA. In MARK2, four kinds of flanking-marker haplotypes were treated as fixed effect and solved with the maximum likelihood estimation.

Results

Computational time

The computational time required for the six methods were ~2,600 s (VC20), ~240 s (VC2), ~92 s (EMF20), ~22 s (EMF2), ~2 s (MARK2), and ~1 s (SMA), respectively. These programs were carried out on Pentium IV PC with 1.0-GHz processor and 512 MB RAM. The computational speed of EMF was much faster than that of VC. It was also found that in EMF20, the ancestral haplotype became unchanged after 3–7 rounds of iterations.

Effect of the heritability

The QTL variance was taken as 0.05, 0.1 and 0.2, which led to the QTL heritability equalled to 0.05, 0.1 and 0.18, respectively. VC20 and VC2 were replicated 200 times and the threshold was determined with 200 replications under null model (no QTL model). Other methods were replicated 1,000 times, and the thresholds were determined with 1,000 replications under null model. The $(1 - \alpha)$ 100th percentile of the distribution of the largest *LR* scores for the null model was an approximation of the threshold, where α was taken as 0.01.

The average LR scores are plotted in Fig. 1, which shows that all methods generate a peak at the true position. The statistical power and parameter estimates for each method under three heritabilities are listed in Table 1. It can be seen that the power of each method can be approximately ranked as VC20 ≈ EMF20 > VC2 ≈ EMF2 ≈ MARK2 > SMA; furthermore, the power of each method was decreased with heritability. The position estimates and their standard deviations for each method are summarized in Table 1. Generally, all methods estimated QTL position very close to the true value, but SMA showed the largest standard deviation; furthermore, the standard deviation of QTL position for each method was also increased with heritability. The true rates of the estimates of ancestral haplotype from EMF2 under the three heritabilities were 80.5, 79.17 and 75.17 %, respectively, which showed no great difference among them.

Effect of mutation generation T

Although the data was simulated under $T = 200$, in practice, T was unknown and usually set beforehand (Meuwissen and Goddard 2000, 2001; Lee and van der Werf 2006). To test the effect of T, QTL variance was fixed at 0.18 and T was taken as 10, 100, 200, 400, 600, 800 and 1,000, respectively; and the thresholds were 13.23, 13.37, 11.69, 11.30, 13.19, 12.16 and 11.5, which were obtained from 1,000 replicated experiments under null model. EMF2 was replicated 1,000 times and the power were 94.6, 94.3, 95, 94, 76.8, 77.4 and 74.2 %, respectively, which shows that when T was varied from 10 to 400 (around the true value 200), the power showed no clear difference, but when T was larger than 400 (severely deviated from true value) the difference would be large.

Performance on multallelic QTL

One multiallelic QTL with each allele mutated at different generations was simulated. The effects of these alleles were assumed to be normally distributed with mean 0 and variance 0.2 (see "Appendix 3" for simulation details). The VC2, EMF2 and SMA were compared and the powers of them were 58.3, 51.4 and 43.2 %, respectively, which shows that the power of EMF2 was lower than that of VC2, but both of them were higher than that of SMA.

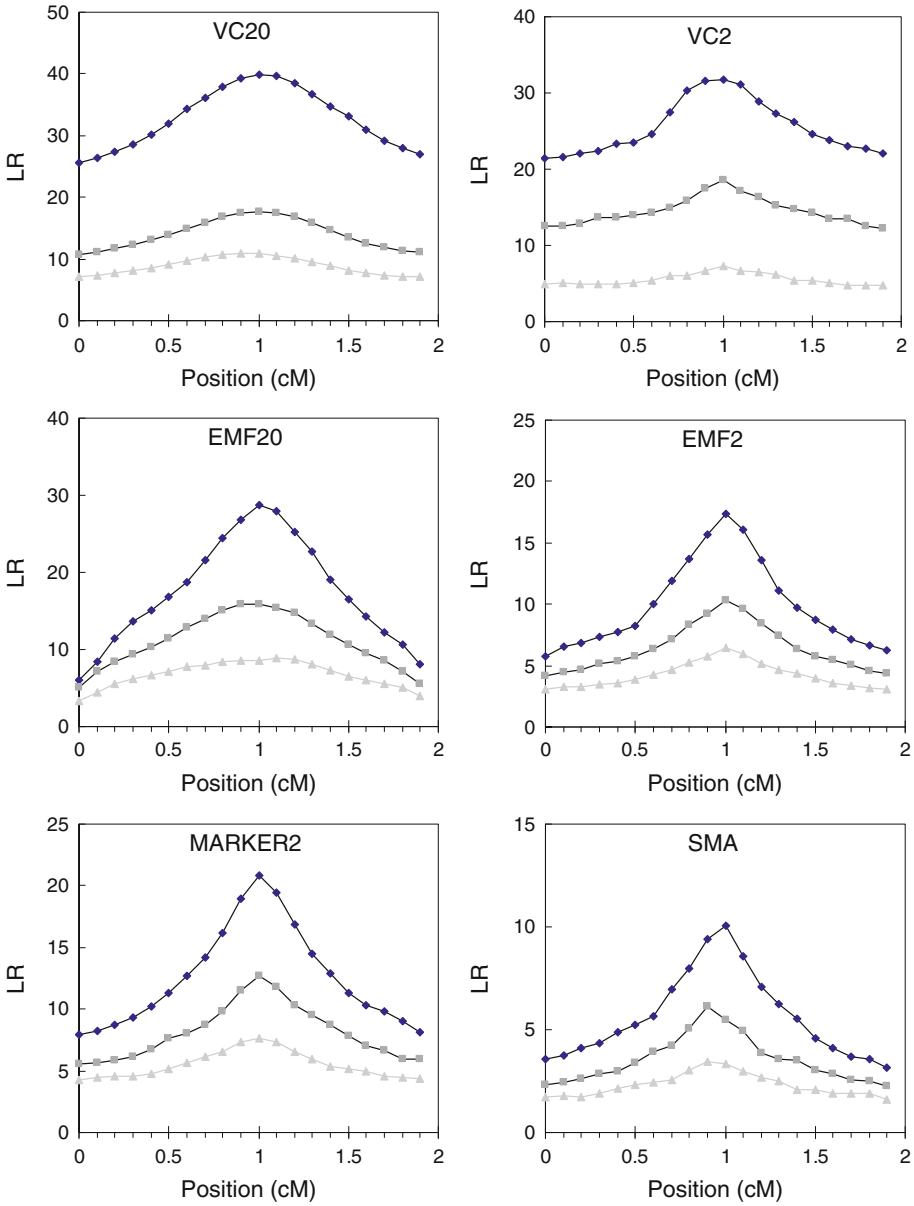

Fig. 1 The profiles of the average *LR* scores of the six methods under three heritabilities 0.18 (*black line*), 0.1 (*dark grey line*) and 0.05 (*light grey line*), respectively. The true position is localized at 1.05 cM

Discussion

A fast EMF was developed for fine-scale QTL mapping. Generally, EMF performed very similar to VC in statistical power and parameter estimate, and both of them outperformed MARK2 and SMA.

The method calculating IBD probability in EMF is the modification of Meuwissen and Goddard's method (2001), but there are two main differences. One is that a_j refers to the probability of the IBS between two individual haplotypes at marker j (i.e. the homozygosity of marker j) in Meuwissen and Goddard (2001), whereas it indicates the probability that an individual haplotype is IBS but nonIBD to the ancestral haplotype at marker j in EMF. If two haplotypes that are IBS at marker j carry marker allele k; then $a_j = q_k^2$ in Meuwissen and Goddard (where q_k is the frequency of kth allele of the marker in the base population), whereas $a_j = q_k$ in EMF. As pointed out by Meuwissen and Goddard (2001), q_k can be estimated with present population. Similar method can be adopted to obtain q_k in EMF, but it was found that both VC and EMF were not very sensitive to q_k (results not shown). The other difference is that EMF introduces a new parameter α in calculation of f(c) (see Eq. 11), which reflects the frequency of the QTL-mutation allele in the present population. Theoretically, α should be estimated with its expectation, $E(\alpha) = 1/(2N_e)$, but it did not give any meaningful results. In fact, the variance of α is very large

Table 1 The parameter estimates under three heritabilities

Heritability	Methods	Position (cM)	QTL variance	Residual variance	Power (%)	Threshold value
0.18	VC20	1.05 (0.284)	0.181 (0.042)	0.908 (0.094)	98.3	15.83
	VC2	1.08 (0.314)	0.182 (0.057)	0.941 (0.106)	94.3	14.48
	EMF20	1.06 (0.245)	0.185 (0.068)	0.904 (0.269)	98.9	13.85
	EMF2	1.06 (0.308)	0.166 (0.066)	0.994 (0.207)	95.1	13.51
	MARK2	1.05 (0.326)	–	1.050 (0.118)	94.4	16.77
	SMA	1.09 (0.376)	–	1.071 (0.115)	84.6	11.56
0.1	VC20	1.07 (0.340)	0.135 (0.038)	0.888 (0.092)	71.3	15.83
	VC2	1.04 (0.336)	0.142 (0.037)	0.876 (0.098)	68.3	14.48
	EMF20	1.03 (0.314)	0.121 (0.039)	0.837 (0.225)	70.1	13.85
	EMF2	1.06 (0.344)	0.116 (0.046)	0.935 (0.174)	67.4	13.68
	MARK2	1.07 (0.329)	–	0.976 (0.092)	67.2	16.77
	SMA	1.10 (0.396)	–	0.980 (0.095)	57.6	11.56
0.05	VC20	1.03 (0.354)	0.110 (0.041)	0.865 (0.090)	30.3	15.83
	VC2	1.05 (0.389)	0.107 (0.024)	0.872 (0.074)	28.6	14.48
	EMF20	1.03 (0.362)	0.109 (0.116)	0.941 (0.116)	29.3	13.85
	EMF2	1.08 (0.361)	0.096 (0.035)	0.912 (0.141)	28.4	13.68
	MARK2	1.01 (0.360)	–	0.934 (0.086)	28.5	16.77
	SMA	1.08 (0.459)	–	0.953 (0.089)	28.2	11.56

The standard deviations of the parameter estimates from replications are given in parenthesis

due to genetic drift, and thus α usually severely deviates from $E(\alpha)$. In this study, α was set as a large number 0.9 and it performed well. α was also varied from 0.5 to 1.0, but the results showed slight difference, which suggested that EMF was not very sensitive to α.

Both VC and EMF introduce a parameter T, the generations since mutation occurred, which is unknown and usually set beforehand. Fortunately, it was found that EMF was not very sensitive to T when T was not severely deviated from true value, which is similar to VC (e.g. Meuwissen and Goddard 2000; Lee and van der Werf 2006).

The EMF modifies Meuwissen and Goddard's method (2001) to infer IBD probability between the individual and ancestral haplotype. Some other methods that infer LD-based IBD matrix between individuals has been studied by many researchers (Hernández-Sánchez et al. 2006; Meuwissen and Goddard 2007; Hill and Hernández-Sánchez 2007). These methods are specifically designed for VC, but with some modifications, they may be suitable for EMF, which are left for further investigation.

This study focuses on unrelated individuals randomly sampled from a population. If a pedigree structure is available, the LA information could also be incorporated along with the LD information. The method that combines LD and LA information is called LDLA. The extension of EMF to LDLA is straightforward, which is described in Appendix 2. The extension still treats QTL effect as fixed effect, and thus the computational advantage of EMF will be held.

In this study, QTL parameters were estimated with the maximum likelihood implemented via EM algorithm, which requires a number of iterations for updating each parameter with calculation of the posterior probabilities P^*_{i0}, P^*_{i1} and P^*_{i2}. Those would costs much CPU time. Another method that approximately substitutes P^*_{i0}, P^*_{i1} and P^*_{i2} with their prior probabilities P_{i0}, P_{i1} and P_{i2} (e.g. Haley and Knott 1992) can avoid the iterative manipulation, and thus would be much faster than EMF.

The EMF assumes a biallelic QTL and it performed well when the assumption is true; however, when QTL was multiallelic the QTL-detection power of EMF would be lower than that of VC. In fact, the EMF could also be modified to accommodate multiple QTL-mutation alleles. In that case, several possible ancestral haplotypes that generate higher *LR* scores should be chosen and the effects of the mutation alleles carried on these ancestral haplotypes are simultaneously included in model, which are not very difficult to implement.

Acknowledgments The three reviewers are thanked for their useful comments. This research was supported by Chinese National Natural Science Foundation grant 31001001.

Appendix 1: Derivation of the probability of an individual haplotype being IBD to the ancestral haplotype at the putative QTL loci

Equation (2) can be written as,

$$p(S_j, S_{j+1}|\mathrm{IBD})p(\mathrm{IBD}) = p(S_j,\ \mathrm{IBD},\ S_{j+1}) = \sum p(\phi)\times p(S|\phi). \quad (10)$$

The second term in Eq. (10) can be factorized as, $p(S|\phi) = \prod_{\text{markerloci j}}^{j+1} p(S(j)|\phi(j))$, where ϕ is the IBD status of a segment including QTL locus, flanking markers (marker j and $j+1$) and the regions in between them; $p(S(j)|\phi(j))$ is the probability of the IBS between individual and ancestral haplotype at marker j conditional on the IBD status of marker j. The calculation of $p(S|\phi)$ for four IBS statuses of flanking markers $(S_j,\ S_{j+1})$, $(1,1)$, $(1,0)$, $(0,1)$ and $(0,0)$ can be easily obtained from a_j and a_{j+1}, where $a_j(a_{j+1})$ denotes the probabilities of the IBS but not the nonIBD between the individual and ancestral haplotype at marker $j(j+1)$ (see Meuwissen and Goddard 2001). Assuming the frequency of each marker allele to be equal in base population, $a_j$can be estimated as 1/(Number of alleles of jth marker). However, this assumption can be relaxed, which will be illustrated in "Discussion".

The first term in Eq. (10), $p(\phi)$ is the probability of the IBD status of the segment including QTL locus, flanking markers (markers j and markers $j+1$) and the regions in between them, which is derived from f(c) (see Meuwissen and Goddard 2001). f(c) is the probability of having an IBD region of size c between two individual haplotypes in Meuwissen and Goddard (2001), but it refers to the probability of having an IBD region of size c between an individual haplotype and the ancestral haplotype in EMF, and thus can be expressed as

$$\mathrm{f}(c) = \exp(-c)^T\alpha, \tag{11}$$

where the first term is the probability that the segment of size c is unbroken for T generations of meiosis; and the second term α is the probability that the intact IBD segment is inherited from the ancestral haplotype carrying the mutant QTL allele. α equals to N_M/N, where N_M is the number of current haplotypes containing the mutation M, and N is the total number of haplotypes. But because N_M is unknown, α also cannot be obtained; therefore, in practice, α should be set beforehand, and the effect of α will be discussed later. The calculation of $p(\phi)$ with f(c) has been explained at length in Meuwissen and Goddard (2001), and thus they are not presented here. Once $p(\phi)$ and $p(S|\phi)$ have been calculated, Eq. (10) can be obtained by summing all possible terms relevant to ϕ that is IBD at QTL (see also Table III of Meuwissen and Goddard 2001 for more details).

The second term in the denominator of Eq. (2) also can be calculated using similar approach, and it can be factorized as

$$\begin{aligned} p(S_j,\ S_{j+1}|\text{nonIBD})p(\text{nonIBD}) &= p(S_j,\ \text{nonIBD},\ S_{j+1}) \\ &= \sum p(\phi) \times p(S|\phi). \end{aligned} \tag{12}$$

which is calculated by summing all possible terms relevant to ϕ that is nonIBD at QTL (see Table III in Meuwissen and Goddard 2001). For clarification, all notations involved in this section are listed in Table 2.

Appendix 2: Extension to the combination of the linkage disequilibrium and linkage analysis

A pedigree with two generations was taken as an example to illustrate the approach to incorporate the linkage information, but the approach can be extended to other more complex pedigrees. Given the linkage phases of the unrelated founders and their offspring, the probabilities that offspring i carries two father's QTL alleles A_1^{P} and A_2^{P} and two mother's QTL alleles A_1^{M} and A_2^{M}, $\mathrm{Prob}(A_1^{\mathrm{P}})$ and $\mathrm{Prob}(A_2^{\mathrm{P}})$, and $\mathrm{Prob}(A_1^{\mathrm{M}})$ and $\mathrm{Prob}(A_2^{\mathrm{M}})$, respectively, can be easily inferred from flanking markers according to Haldane's recombination rule (e.g., using the method of Wang et al. 1995). Given the probability that two QTL alleles (indicated by 1 and 2, respectively) of the father (i_{P}) and mother (i_{M}) of offspring i is IBD to the ancestral mutation allele, denoted by $\pi_{i_{\mathrm{P}}}^1$ and $\pi_{i_{\mathrm{P}}}^2$ (for father), $\pi_{i_{\mathrm{M}}}^1$ and $\pi_{i_{\mathrm{M}}}^2$ (for mother), the probabilities of three QTL genotypes combining LD and LA information can be calculated as, $P_{i1} = (\mathrm{Prob}(A_1^{\mathrm{P}})\pi_{i_{\mathrm{P}}}^1 + \mathrm{Prob}(A_2^{\mathrm{P}})\pi_{i_{\mathrm{P}}}^2) \cdot (\mathrm{Prob}(A_1^{\mathrm{M}}) + \mathrm{Prob}(A_2^{\mathrm{M}})\pi_{i_{\mathrm{M}}}^2)$ for genotype MM, $P_{i3} = (\mathrm{Prob}(A_1^{\mathrm{P}})(1-\pi_{i_{\mathrm{P}}}^{\mathrm{P}}) + \mathrm{Prob}(A_2^{\mathrm{P}})(1-\pi_{i_{\mathrm{M}}}^{\mathrm{M}})) \cdot (\mathrm{Prob}(A_1^{M})(1-\pi_{i_{\mathrm{M}}}^{\mathrm{P}}) + \mathrm{Prob}(A_2^{\mathrm{M}})(1-\pi_{i_{\mathrm{M}}}^{\mathrm{M}}))$ for mm, and $P_{i2} = 1 - P_{i1} - P_{i3}$ for Mm or mM, respectively,

Table 2 List of the notation symbols

S_j	The identity-by-state (IBS) status between an individual haplotype and the ancestral haplotype at marker j $S_j = 1$ ($S_j = 0$) indicates (non)IBS of marker j
p(IBD)	The prior probability of a QTL allele being IBD to the mutational QTL allele M
ϕ	The IBD status of a segment including QTL locus, flanking markers and the regions in between them
$\phi(j)$	The IBD status at marker j
a_j	The probability that an individual haplotype is IBS but nonIBD to the ancestral haplotype at marker j
f(c)	The probability of having an IBD region of size c between individual and ancestral haplotype
α	The probability that an intact IBD segment is inherited from the ancestral haplotype carrying the mutational QTL allele

which assumes the QTL loci is in Hardy–Weinberg equilibrium.

Appendix 3: Simulation of multiple QTL mutations

A chromosome segment with length of 2 cM was simulated. Twelve markers were evenly spaced on the segment and one QTL was localized at 1.05 cM. The population was created 500 generations ago, the effective population size (N_e) was 200, and sex ratio was 1:1. In the base population, two alleles were assigned to each marker with equal frequency, and only one allele was assigned to QTL. The marker alleles were mutated at a rate of 4×10^{-4}/ generation. A new QTL mutation occurred every two generations. One individual haplotype was randomly chosen, and the QTL allele on the haplotype was mutated to a new QTL allele and assigned a new number. The high mutation rate of QTL might result in about 6–12 alleles in the present population. The effects of each QTL allele were randomly sampled from standard normal distribution $N(0, 1)$. At the last generation, the effect of each QTL allele was rescaled so that the mean of QTL effect was zero and the variance was 0.2. The residual effect was sampled from normal distribution with mean 0 and variance 0.9; the overall mean was set as zero, and no polygenic effect was simulated. With these settings, the heritability explained by QTL was 0.18. The phenotypic value for each individual then was generated by summing the overall mean, QTL effect and residual error.

References

Chen WM, Abecasis GR (2007) Family-based association tests for genome wide association scans. Am J Hum Genet 81:913–926

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B 39:1–38

Druet T, Fritz S, Boussaha M, Ben-Jemaa S, Guillaume F, Derbala D, Zelenika D, Lechner D, Charon C, Boichard D, Gut IG, Eggen A, Gautier M (2008) Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178:2227–2235

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Hernández-Sánchez J, Haley CS, Woolliams JA (2006) Prediction of IBD based on population history for fine gene mapping. Genet Sel Evol 38:231–252

Hernández-Sánchez J, Grunchec JA, Knott S (2009) A web application to perform linkage disequilibrium and linkage analyses on a computational grid. Bioinformatics 25:1377–1383

Hill WG, Hernández-Sánchez J (2007) Prediction of multi-locus identity-by-descent. Genetics 176:1–9

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kimmel G, Karp RM, Jordan MI, Halperin E (2008) Association mapping and significance estimation via the coalescent. Am J Hum Genet 83:675–683

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–194

Lee SH, van der Werf JHJ (2006) Simultaneous fine mapping of multiple closely linked quantitative trait loci using combined linkage disequilibrium and linkage with a general pedigree. Genetics 173:2329–2337

Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nat Genet 39:906–913

McPeek MS, Strahs A (1999) Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine scale genetic mapping. Am J Hum Genet 65:858–875

Meuwissen THE, Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155:421–430

Meuwissen THE, Goddard ME (2001) Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol 33:605–634

Meuwissen THE, Goddard ME (2007) Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. Genetics 176:2551–2560

Minichiello M, Durbin R (2006) Mapping trait loci by use of inferred ancestral recombination graphs. Am J Hum Genet 79:910–922

Morris AP (2006) A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. Am J Hum Genet 79:679–694

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554

Pérez-Enciso M (2003) Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163:1497–1510

Schnabel RD, Kim J-J, Ashwell MS, Sonstegard TS, Van Tassell CP, Connor EE, Taylor JF (2005) Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. Proc Natl Acad Sci USA 102:6896–6901

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56:777–787

Wang T, Fernanda RL, van der Beek S, van Arendonk JAM (1995) Covariance between relatives for a marked quantitative trait locus. Genet Sel Evol 27:251–274

Wang WYS, Baratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678

Xiong M, Guo SW (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531