Heredity

Heredity (2012) 109, 41–49

www.nature.com/hdy

ORIGINAL ARTICLE

Generalized linear mixed models for mapping multiple quantitative trait loci

X Che[1] and S Xu[2]

Many biological traits are discretely distributed in phenotype but continuously distributed in genetics because they are controlled by multiple genes and environmental variants. Due to the quantitative nature of the genetic background, these multiple genes are called quantitative trait loci (QTL). When the QTL effects are treated as random, they can be estimated in a single generalized linear mixed model (GLMM), even if the number of QTL may be larger than the sample size. The GLMM in its original form cannot be applied to QTL mapping for discrete traits if there are missing genotypes. We examined two alternative missing genotype-handling methods: the expectation method and the overdispersion method. Simulation studies show that the two methods are efficient for multiple QTL mapping (MQM) under the GLMM framework. The overdispersion method showed slight advantages over the expectation method in terms of smaller mean-squared errors of the estimated QTL effects. The two methods of GLMM were applied to MQM for the female fertility trait of wheat. Multiple QTL were detected to control the variation of the number of seeded spikelets.
Heredity (2012) **109,** 41–49; doi:10.1038/hdy.2012.10; published online 14 March 2012

Keywords: binary trait; binomial trait; mixed model; overdispersion; QTL

INTRODUCTION

Linear mixed model (LMM) methodology is a powerful technology to analyze models containing both the fixed and random effects. The model was first proposed to estimate genetic parameters for unbalanced data (Henderson, 1950). This technique has been used to map genes controlling the variation of quantitative traits (Xu and Yi, 2000; Boer *et al*, 2007). The LMM methodology cannot be directly applied to traits with discrete distributions. Wedderburn (1974) proposed a linear predictor and a link function to handle discrete traits. The linear predictor is simply a linear model combining information from the independent variables. The link function is used to describe the relationship between the linear predictor and the expectation of the response variable. This approach eventually leads to a special area of statistics called the generalized linear model (GLM) (McCullagh and Nelder, 1989).

Xu and Hu (2010) recently developed a GLM approach to interval mapping (IM) for traits with discrete distribution. The purpose of that study was to investigate the efficiencies of two different methods for handling missing genotypes: (1) the heterogeneous residual variance method and (2) the mixture model method. In the first method (heterogeneous residual variance method), we replaced the missing genotypes of quantitative trait loci (QTL) by the conditional expectations of the genotype indicator variables and then took into account the heterogeneous residual variances of different genotypes due to heterogeneous information contents. In the second method (the mixture model method), we fully utilized the conditional distributions of the missing genotypes. Theoretically, the mixture model approach should be optimal. In practice, the heterogeneous residual variance method is more efficient because it is robust to departure from the assumed normal distribution of the residuals. On the contrary, the mixture model is very sensitive to the departure of an assumed distribution and the choice of the initial values of the parameters. These missing-genotype-handling methods have not been applied to multiple QTL mapping (MQM).

When the number of QTL included in a model reaches a certain level, for example, the number of QTL is larger than the sample size, the model is oversaturated. In this case, some kind of penalty is required to shrink the superfluous QTL down to zero. The penalty is accomplished by treating each QTL effect, say QTL k, as a random effect with a $N(0, \sigma_k^2)$ distribution. When the linear predictor contains both fixed and random effects, the model is then called the generalized LMM (GLMM) (Breslow and Clayton, 1993; McCulloch and Neuhaus, 2005). Special algorithms have been developed to estimate variance components and predict the random effects, for example, the pseudo likelihood algorithm (Wolfinger and O'Connell, 1993). However, existing GLMM have not fully considered the missing genotype problem.

In this study, we extended the GLM for IM of QTL (Xu and Hu, 2010) to GLMM for MQM. The difference between IM and MQM is that IM uses a model that contains only one QTL effect at a time (the entire genome analysis requires multiple analyses of many single-effect models), whereas MQM estimates all QTL simultaneously in a single model. Although Xu and Hu (2010) developed two methods for GLM analysis, we only examined the heterogeneous residual variance method. The mixture model did not offer any advantages over the heterogeneous residual variance method (Xu and Hu, 2010), and thus will not be examined here in this study. In addition, we evaluated a simple method called the expectation method, in which the missing genotypes of QTL are simply replaced by the conditional expectation of the genotype indicator variables. The heterogeneous residual

[1]Department of Statistics, University of California, Riverside, CA, USA and [2]Department of Botany and Plant Sciences, University of California, Riverside, CA, USA
Correspondence: Dr S Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA.
E-mail: shizhong.xu@ucr.edu
Received 11 November 2011; revised 5 January 2012; accepted 9 January 2012; published online 14 March 2012

variance method called by Xu and Hu (2010) is now rephrased as the overdispersion method. We believe that overdispersion is a more appropriate term in the context of GLMM.

METHODS

Generalized linear mixed model

We use a binomial trait as an example to demonstrate the new methodology, although the method can be applied to other discrete traits. Let y_j be the number of events and t_j be the number of trials for individual j from a population of n individuals. Let $E(y_j/t_j)=\mu_j$ be the expectation of the binomial trait. Define $\eta_j=\Phi^{-1}(\mu_j)$ as a linear predictor with the probit link function. The linear predictor is a function of marker genotypes, as described below,

$$\eta_j = \beta + \sum_{k=1}^{m} Z_{jk}\gamma_k \tag{1}$$

where β is the intercept, γ_k is the effect for marker k, Z_{jk} is an independent variable determined by the genotype of marker k of individual j and m is the total number of markers included in the model. In a later section, markers are replaced by pseudo markers. Each marker is then considered as a putative QTL. Therefore, we may call m the number of putative QTL. Details about Z_{jk} will be described later.

When m is large, say $m>n$, the model is oversaturated and solutions of the parameters will not be unique. To overcome this problem, a penalty should be placed on the QTL effects. Ridge regression (Hoerl and Kennard, 1970) is often used as a penalized regression analysis. It corresponds to the L_2 penalty (Zou, 2006; Friedman *et al*, 2010), in which γ_k is treated as a random effect and further described by a $N(0, \sigma_k^2)$ distribution. Once γ_k is treated as a random effect, it becomes a random variable and thus does not reduce the degree of freedom of the residual. In addition, the zero mean distribution serves as a 'prior' belief of no effect from the Bayesian point of view. These are the very reasons why a mixed model can handle a very large number of regression coefficients once the coefficients are treated as random effects. The intercept β is treated as a fixed effect (no distribution is assigned) because we do not want to penalize a model based on the size of the intercept. The linear predictor includes both the fixed effect (β) and the random effects (γ), and thus is called the mixed model. The least absolute shrinkage and selection operator (Lasso) method developed by Tibshirani (1996) is another penalized regression analysis, called the L_1 penalty. We will not pursue the Lasso approach because it is beyond the scope of the GLMM.

Let us denote all QTL effects by an $m\times 1$ vector $\gamma = \{\gamma_k\}, \forall k = 1, \cdots, m$ and denote the multivariate normal density of γ by $p(\gamma|G)=N(\gamma|0,G)$ where $G = \mathrm{diag}\{\sigma_k^2\}$ is a diagonal matrix for the variance components. This special notation for probability density $p(\gamma|G)=N(\gamma|0,G)$ is adopted from Gelman *et al.* (2004). It represents both the distribution and the density, that is,

$$N(\gamma|0,G) = \frac{1}{(2\pi)^{\frac{m}{2}}|G|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\gamma^T G^{-1}\gamma\right) \tag{2}$$

Conditional on $\eta_j=\beta+Z_j\gamma$, the binomial distribution for y_j is

$$p(y_j|\eta_j) = \mathrm{const}\times\left[\Phi(\eta_j)\right]^{y_j}\left[1-\Phi(\eta_j)\right]^{t_j-y_j} \tag{3}$$

When γ are treated as random effects, they are no longer considered as parameters in the GLMM, although they remain to be important genetic parameters in terms of QTL mapping. The parameters are now formed by $\theta=\{\beta,G\}$. Conditional on $\eta=\beta+Z\gamma$, we have the joint probability for the binomial trait of the entire sample

$$p(y|\eta) = \prod_{j=1}^{n} p(y_j|\eta_j) \tag{4}$$

The likelihood function for the parameter vector $\theta=\{\beta,G\}$ is proportional to the following probability

$$p(y|\beta,G) = \int p(y|\eta)p(\gamma|G)d\gamma \tag{5}$$

where the integration is taken with respect to γ. The integral is multivariate and no explicit expression exists. The log likelihood function for parameter $\theta=\{\beta,G\}$ is defined as

$$L(\beta,G) = \ln p(y|\beta,G) = \ln\left[\int p(y|\eta)p(\gamma|G)d\gamma\right] \tag{6}$$

and thus also does not have an analytical expression. The maximum likelihood estimate of $\theta=\{\beta,G\}$ would be obtained by solving $\frac{\partial}{\partial\theta}L(\beta,G)=0$ if $L\{\beta,G\}$ were explicitly expressed. A pseudo likelihood algorithm was developed to solve for the parameters (Wolfinger and O'Connell, 1993). Laplace approximation has also been used to replace the integral (Vonesh, 1996). In this study, we adopted a simple method that does not involve numerical integration. This method is called the MAP estimation, as described below.

MAP estimation

The word MAP stands for maximum a posteriori (DeGroot, 2004), which is a terminology related to Bayesian analysis. Our GLMM is a frequentist approach if we treat $\{\beta,G\}$ as parameters. However, if we consider $\{\beta,\gamma\}$ as parameters and treat G as a prior variance matrix for γ, the problem becomes a Bayesian problem and parameter estimation can be achieved under the Bayesian framework. In a typical Bayesian problem, the parameters in the prior should be provided by the investigator before the data analysis. It is hard to provide a prior value for G, and thus we must estimate G from the data. Once G is estimated from the data, the problem is more like a mixed model problem. Therefore, the difference between the Bayesian model and the GLMM becomes blurred. We may consider the MAP algorithm as a simplified approach to estimating parameters under the GLMM framework (see McGilchrist 1994). We will first provide the MAP estimation and then show the difference between the MAP estimates and the ML estimates.

Unlike the ML estimation in which the target function for maximization is $L(\beta,G)$, in the MAP estimation, we maximize the log posterior function defined as

$$L(\beta,\gamma,G) = L(\beta,\gamma)+L(G) \tag{7}$$

where

$$L(\beta,\gamma) = \sum_{j=1}^{n}\left[y_j\ln\mu_j+(t_j-y_j)\ln(1-\mu_j)\right] \tag{8}$$

and

$$L(G) = -\frac{1}{2}\ln|G| - \frac{1}{2}\gamma^T G^{-1}\gamma = -\frac{1}{2}\sum_{k=1}^{m}\ln(\sigma_k^2) - \frac{1}{2}\sum_{k=1}^{m}\frac{\gamma_k^2}{\sigma_k^2} \tag{9}$$

The MAP estimation for $\xi=\{\beta,\gamma,G\}$ is obtained by setting $\frac{\partial}{\partial\xi}L(\beta,\gamma,G)=0$ and solving for ξ. The iteration process is summarized in the following sequences.

Step (1): Set $t=0$ and initialize all parameters $\xi^{(t)} = \left\{\beta^{(t)},\gamma^{(t)},G^{(t)}\right\}$.

Step (2): Update β using

$$\beta^{(t+1)} = \beta^{(t)} - \left[\frac{\partial^2 L(\beta,\gamma)}{\partial\beta\partial\beta^T}\right]^{-1}\left[\frac{\partial L(\beta,\gamma)}{\partial\beta}\right] \tag{10}$$

Step (3): Update γ_k for $k=1,\ldots,m$ using

$$\gamma_k^{(t+1)} = \gamma_k^{(t)} - \left[\frac{\partial^2 L(\beta,\gamma,G)}{\partial\gamma_k\partial\gamma_k^T}\right]^{-1}\left[\frac{\partial L(\beta,\gamma,G)}{\partial\gamma_k}\right] \tag{11}$$

Step (4): Update σ_k^2 for $k=1,\ldots,m$ using

$$\sigma_k^{2(t+1)} = (\gamma_k^{(t+1)})^2 \tag{12}$$

Step (5): Repeat Steps (2) to Step (4) until the sequence converges.

Note that Steps (2) and (3) are the first step iteration of the Newton–Raphson algorithm (Ypma, 1995). The MAP approach for GLMM was first proposed by McGilchrist (1994). It is a much simplified algorithm that has avoided multiple integration. The original MAP of McGilchrist (1994) did not

address the missing value problem, which will be dealt with in the next section of this study.

Let us now compare the MAP with the EM algorithm. The target function to be maximized with the EM algorithm is

$$\mathrm{E}[L(\beta,\gamma,G)] = \mathrm{E}[L(\beta,\gamma)]+\mathrm{E}[L(G)] \tag{13}$$

where the expectation is taken with respect to γ. The MLE of $\theta=\{\beta,G\}$ is obtained by solving

$$\frac{\partial}{\partial\theta}\mathrm{E}[L(\beta,\gamma,G)] = \frac{\partial}{\partial\theta}\mathrm{E}[L(\beta,\gamma)]+\frac{\partial}{\partial\theta}\mathrm{E}[L(G)] = 0 \tag{14}$$

The corresponding EM steps are modifications of the MAP steps as shown below. The updating step for β in the EM is

$$\beta^{(t+1)} = \beta^{(t)} - \left[\frac{\partial^2\mathrm{E}[L(\beta,\gamma)]}{\partial\beta\partial\beta^T}\right]^{-1}\left[\frac{\partial\mathrm{E}[L(\beta,\gamma)]}{\partial\beta}\right] \tag{15}$$

which is a maximization (M) step. The updating step for $\gamma_k\ \forall k=1,\cdots,m$ is

$$\gamma_k^{(t+1)} = \gamma_k^{(t)} - \left[\frac{\partial^2\mathrm{E}[L(\beta,\gamma,G)]}{\partial\gamma_k\partial\gamma_k^T}\right]^{-1}\left[\frac{\partial\mathrm{E}[L(\beta,\gamma,G)]}{\partial\gamma_k}\right] \tag{16}$$

This is the expectation (E) step. Another maximization (M) step is to update σ_k^2 for k=1,…,m using

$$\sigma_k^{2(t+1)} = \mathrm{E}(\gamma_k^2) = \mathrm{E}^2(\gamma_k)+\mathrm{var}(\gamma_k) = \gamma_k^{2(t+1)}+s_k^2 \tag{17}$$

where $\mathrm{E}(\gamma_k) = \gamma_k^{(t+1)}$ is the expectation of γ_k and

$$s_k^2 = -\left[\frac{\partial^2\mathrm{E}[L(\beta,\gamma,G)]}{\partial\gamma_k\partial\gamma_k^T}\right]^{-1} \tag{18}$$

is the variance of γ_k. The EM algorithm requires calculation of the expectation of the first- and second-order partial derivations of the target function, which is by no means a simple task. This is the very reason why McGilchrist (1994) proposed the MAP for GLMM. Note that the updating step for σ_k^2 is explicit and obtained by setting $\frac{\partial}{\partial\sigma_k^2}L(G) = 0$ for the MAP and $\frac{\partial}{\partial\sigma_k^2}\mathrm{E}[L(G)] = 0$ for the EM algorithm. Therefore, the MAP estimation does not exactly lead to the EM estimation in the frequentist framework. However, the results are very close and this is why McGilchrist (1994) developed the MAP estimation for variance component analysis in the GLMM framework.

LOD (log of odds) score test

The estimated QTL effect (after MAP iteration converges) is denoted by $\hat{\gamma}_k$. We can now perform statistical tests. The test statistic for H_0:γ_k=0 may be the t-test,

$$t_k = \hat{\gamma}_k/s_k \tag{19}$$

It is called the t-test because it is expressed as the ratio of the estimated effect to the s.e. However, under the null model, this test statistic may not follow the t-distribution because of the penalty placed on the estimation. This test statistic is negative if the estimated QTL effect is negative. The Wald test (Wald, 1943) is simply the square of the t-test

$$W_k = \frac{\hat{\gamma}_k^2}{s_k^2} \tag{20}$$

which is similar to the likelihood ratio test statistic. The best presentation of the test statistic is the LOD score defined as

$$\mathrm{LOD}_k = \frac{W_k}{2\ln(10)} \tag{21}$$

A nice property of the LOD score test is that an empirical critical value of

$$\mathrm{LOD} = 3+\log_{10}(m) \tag{22}$$

may be used to declare statistical significance at the 0.05 type I error rate (Kidd and Ott, 1984; Risch, 1991). The number m occurred in $\log_{10}(m)$ is the number of putative QTL included in the model. The special case of m=1 corresponds to the LOD 3 criterion.

Missing genotypes

In QTL mapping, the genotype indicator variable (Z_{jk}) is missing if the QTL position does not overlap with a fully informative marker. However, partial information is available due to linkage disequilibrium. We examined two methods for handling missing genotypes.

Expectation method. The linkage disequilibrium allows us to infer the conditional distribution of Z_{jk} given information from linked markers. Let A_1A_1, A_1A_2 and A_2A_2 be the three genotypes of a QTL for an individual in an F_2 population. The Z variable is determined by the genotype of locus k,

$$Z_{jk} = \begin{cases} +1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \\ -1 & \text{for } A_2A_2 \end{cases} \tag{23}$$

In the context of GLMM, γ_k=a_k, where a_k is called the additive effect of locus k. When Z_{jk} is missing, the expectation and variance of it are inferred from the genotypes of flanking markers (Jiang and Zeng, 1997). Let $p_j(+1)$, $p_j(0)$ and $p_j(-1)$ be the conditional probabilities of the three genotypes inferred from neighboring markers using the multipoint method (Jiang and Zeng, 1997). The expectation and variance of Z_{jk} are (Xu and Hu, 2010)

$$\mathrm{E}(Z_{jk}) = U_{jk} = p_j(+1) - p_j(-1) \tag{24}$$

and

$$\mathrm{var}(Z_{jk}) = \Sigma_{jk} = \left[p_j(+1)+p_j(-1)\right] - \left[p_j(+1) - p_j(-1)\right]^2 \tag{25}$$

With the expectation method, we simply replace Z_{jk} by U_{jk}. Therefore, the linear predictor is defined as

$$\eta_j = \beta+\sum_{k=1}^{m} U_{jk}\gamma_k \tag{26}$$

Everything else remains the same as the situation with complete genotypic information.

Overdispersion method. The expectation method only takes advantage of the first moment of the distribution of Z_{jk}. The second moment information has been ignored, which will generate a situation called overdispersion. For locus k, the overdispersion is defined as

$$o_{jk} = \gamma_k^T\Sigma_{jk}\gamma_k+1 \tag{27}$$

Incorporating this overdispersion, we redefine the linear predictor as

$$\eta_{jk} = \frac{1}{\sqrt{o_{jk}}}\left(\beta+U_{jk}\gamma_k+\xi_{jk}\right) \tag{28}$$

where

$$\xi_{jk} = \sum_{k'\neq k}^{p} U_{jk'}\gamma_{k'} \tag{29}$$

is an offset of the linear predictor contributed by other loci. We now have a locus-specific $\mu_{jk} = \Phi(\eta_{jk})$ to define various log functions for maximization.

RESULTS

Simulation study

Binomial data. We simulated a single large chromosome of 2400-cM long evenly covered by 241 co-dominance markers (10 cM per marker interval). The simulated population was an F_2 family derived from the cross of two inbred lines with a sample size n=500. The genotype indicator variable for individual j at locus k was defined as Z_{jk}={+1, 0, −1} for the three genotypes (A_1A_1, A_1A_2, A_2A_2). Dominance effects were not simulated and also not included in the model for this simulation experiment. A total of 20 QTL were simulated with the true sizes and locations of the QTL depicted in Figure 1 (the top panel). Most QTL were placed in the left part of the genome. Some QTL were far apart from each other, whereas others were clustered in some

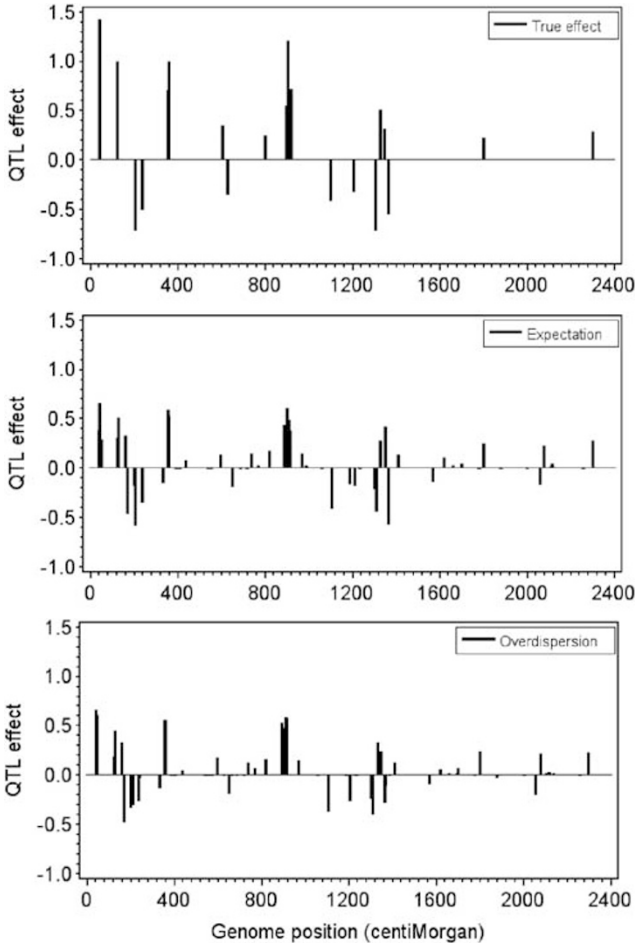

Figure 1 True QTL effects (top panel) and their estimated values for the simulated binomial trait (BINOMIAL) using the expectation method (panel in the middle) and the overdispersion method (bottom panel). The estimate QTL effects are drawn from a single simulated sample. The positions of 20 simulated QTL are indicated by the inward ticks on the horizontal axis.

narrow regions. About half of the simulated QTL overlapped with true markers (known genotypes) and the remaining QTL were located between markers (having missing genotypes). We first generated a linear predictor η_j for each individual using the genotypes of the 20 simulated QTL and the true effects of these QTL. The linear predictor was then converted into the probability of a binomial variable using $\mu_j = \Phi(\eta_j)$. We then simulated a zero-truncated Poisson variable with mean 4 as the number of trials for individual j, denoted by t_j (the number of trial must be $>$zero). We then simulated the number of events y_j from the corresponding binomial distribution defined by μ_j and t_j, that is, $y_j \sim$ Binomial (μ_j, t_j). The simulation experiment was replicated 100 times.

In the simulated data analysis, we added a pseudo marker in every 2.5 cM of the genome, which is equivalent to adding three pseudo markers per marker interval (10 cM is the length of each interval). Genotypic probabilities of the pseudo markers were inferred from information of flanking markers (Jiang and Zeng, 1997). These probabilities were used to calculate U_{jk} and Σ_{jk}. The total number of putative loci analyzed was $m = 241+3\times240 = 241+720 = 961$ with 241 true markers and 720 pseudo markers. This m is almost twice the size of the sample (n=500). We wrote a SAS/IML program to analyze

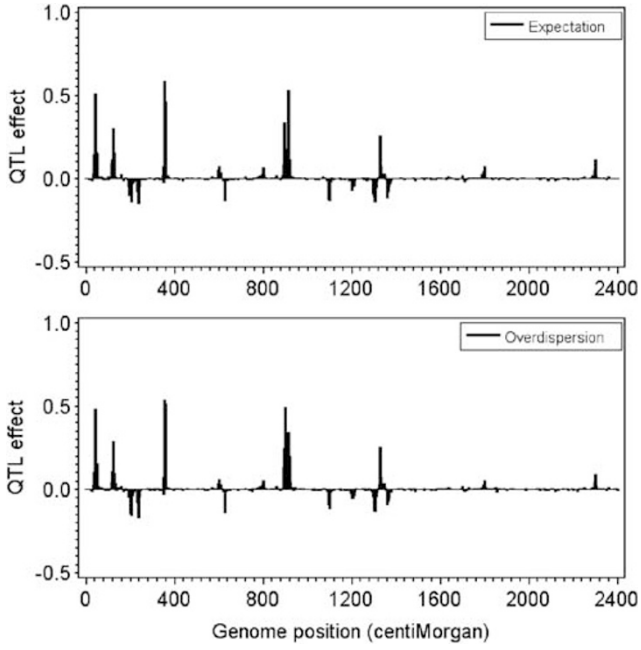

Figure 2 Estimated QTL effects for the simulated binomial trait (BINOMIAL) using the expectation method (panel in the middle) and the overdispersion method (bottom panel). The estimated QTL effects are the averages of 100 replicated samples.

the data. The IML code is available from the corresponding author on request.

The estimated QTL effects from one random sample are presented in Figure 1 for the two methods (expectation and overdispersion) along with the true simulated effects. The two methods produced very similar results, which mimic the true QTL effects closely in terms of locations and the sizes of the effects. The general observations from this figure are (1) a large QTL effect may be split into two or more small effects in the neighborhood of the true QTL and (2) the estimated effects are generally smaller than the true effects due to penalty. Without the penalty, however, we cannot estimate all the 961 putative QTL simultaneously. In any real data analysis with a single sample, the pattern shown in Figure 1 is what an investigator expects to see.

Figure 2 shows the plot of the average estimated QTL effects (across 100 replicated samples) against the genome location. This time, the positions and the patterns of the QTL are extremely close to the true QTL shown in Figure 1 (the top panel). However, the average estimates of the QTL effects are severely biased downwards (towards zero). The differences between the two methods were barely noticed from the visual plots. The simulation experiments allow us to evaluate the bias and estimation error of each QTL and eventually the mean-squared error (MSE) for all the QTL. Let $\bar{\gamma}_k$ be the average estimate of γ_k for the 100 replicates and s_k^2 be the variance of the estimated γ_k across the 100 samples, the MSE for γ_k is defined as

$$\mathrm{MSE}_k = (\bar{\gamma}_k - \gamma_k)^2 + s_k^2 \tag{30}$$

The sum of MSE's for all QTL is

$$\mathrm{MSE} = \sum_{k=1}^{m} (\bar{\gamma}_k - \gamma_k)^2 + \sum_{k=1}^{m} s_k^2 = \mathrm{Bias+Error} \tag{31}$$

The MSEs for the two methods (expectation and overdispersion) are shown in Table 1. The overdispersion method has a slightly larger bias but with a smaller error compared with the expectation method.

Table 1 Comparison of the MSE for the two methods in the 100 replicated simulation experiments

Data type	*Model*	*Bias*[a]	*Error*[b]	*MSE*[c]
Binomial	Expectation	6.00	3.52	9.52
	Overdispersion	6.01	3.14	9.15
Binary	Expected	5.97	8.06	14.03
	Overdispersion	6.63	5.04	11.67

Abbreviation: MSE, mean-squared errors.
[a]Bias is defined as the sum of squared differences between the true QTL effects and the average estimated QTL effects.
[b]Error is defined as the sum of the variances of the estimated effects obtained from all replicates.
[c]MSE is the sum of Bias and the Error. Please see equation (31) in the text for details.

Overall, the overdispersion method has a smaller MSE than the expectation method. The bias defined here from the replicated simulation experiments may be overstated for the following reasons: (1) with the high density of the putative QTL in the model, a true QTL is often detected by a nearby marker close to the true QTL. The exact locations vary from one sample to another, but all in the neighborhood of the true QTL. When an average is taken across the samples, the effect of the true location is diluted by those samples in which the estimated QTL is a few cM away from the true QTL. For example, if a true QTL is estimated in the true location (A) from one sample and it is estimated in a position 2.5 cM away from the true location (B) in the second sample, the average effect of the two samples is then halved for the true location. This problem will be corrected in any real data analysis because a QTL detected in experiment A (denoted by QTL_A) will be treated as the same QTL detected in experiment B (denoted by QTL_B) as long as QTL_A and QTL_B are not too far away from each other. Therefore, the smaller estimation error of the overdispersion method is perhaps more important than the large bias.

Binary data. The experimental design is exactly the same as that of the binomial experiment. The only difference in the simulation is that the trial was a fixed number of one for every individual in the binary data simulation experiment. The estimated QTL effects from one random sample are presented in Figure 3 for the two methods (expectation and overdispersion) along with the true simulated effects. Again, the two methods produced very similar results. However, they differ from the true QTL effects more than what we observed for the binomial trait analysis. Some QTL with small effects have been missed here, for example, the last simulated QTL in the genome. This indicates lower efficiency of QTL mapping for binary trait analysis than for binomial trait analysis.

Figure 4 shows the plot of the average estimated QTL effects (across 100 replicated samples) against the genome location. Again, the positions and the patterns of the QTL are close to the true QTL shown in Figure 3 (the top panel). The MSE's for the two methods (expectation and overdispersion) are shown in Table 1. The overdispersion method has a much larger bias but with much smaller error compared with the expectation method. Overall, the overdispersion method has a smaller MSE than the expectation method. The advantages of the overdispersion method are well supported in the simulation experiments.

Mapping wheat fertility QTL

The experiment was conducted by Dou *et al* (2009). The mapping population contained 243 F_2 individuals derived from the cross of two inbred lines. The trait of interest is the female fertility measured as a binomial trait. The event is the number of seeded spikelets per plant (average 19.13 seeded spikelets) and the trial is the total number of spikelets per plant (average 25.15 spikelets). A total of 28 markers were genotyped in this experiment. These markers covered five chromosomes of the wheat genome with an average marker interval of 15.5 cM. The five chromosomes are only part of the wheat genome.

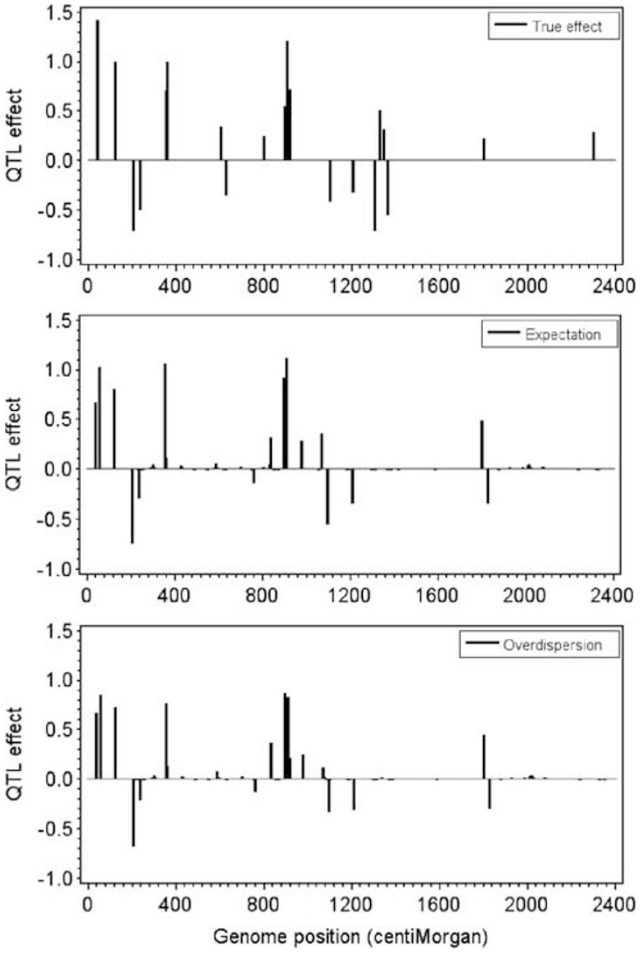

Figure 3 True QTL effects (top panel) and their estimated values for the simulated binary trait (BINARY) using the expectation method (panel in the middle) and the overdispersion method (bottom panel). The estimated QTL effects are drawn from a single simulated sample.

Binomial trait. As the marker map is sparse, we inserted one pseudo marker in every 2 cM, generating a total of 197 loci (28 true markers and 169 pseudo markers). The pseudo markers have missing genotypes and the probability distributions of these pseudo markers were inferred from linked markers using the multipoint methods (Jiang and Zeng, 1997). The sample size was n=243 and the size of the model was m=197. Both the expectation and overdispersion methods were used for the binomial data analysis.

For the real data analysis, we need to calculate the LOD score for each putative locus. The estimated QTL effects for the two methods are depicted in Figure 5 (the top panel). The LOD score profiles for the two methods are depicted in Figure 5 (the bottom panel). The two methods show some similarity and differences. Using the $LOD=3+\log_{10}(197)=5.2944$ as the threshold (Kidd and Ott, 1984), the expectation method detected 17 QTL, whereas the overdispersion method detected 15 QTL. Among these detected QTL, eight of them

were detected by both methods. The estimated effects along with the s.e. and the LOD scores obtained from the overdispersion method are listed in Table 2. Most detected QTL were located on chromosome II, IV and V. The QTL with the largest effect and LOD score occurred on the second chromosome at position 28.71 cM (cumulative position of 104.60 cM). This QTL was split into a few smaller ones in the neighborhood of the major peak by the expectation method. Unlike the simulation study where the true effects of QTL were known, for the wheat data, the true QTL were not known. Therefore, we were not

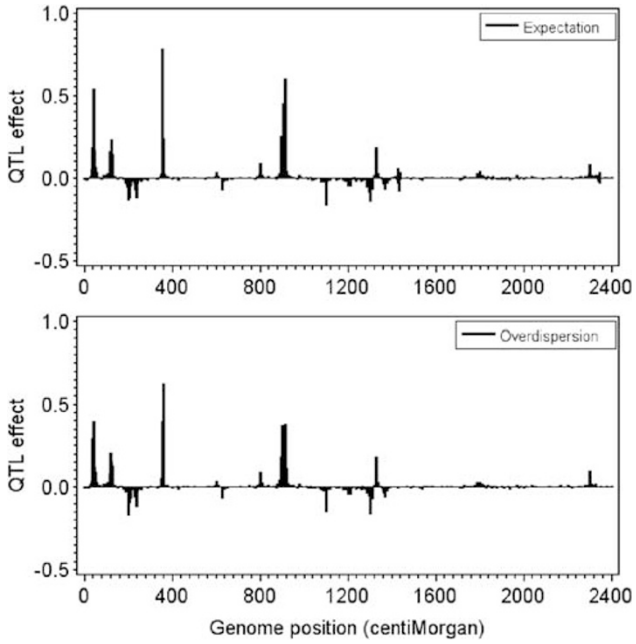

Figure 4 Estimated QTL effects for the simulated binary trait (BINARY) using the expectation method (panel in the middle) and the overdispersion method (bottom panel). The estimated QTL effects are the averages of 100 replicated samples.

Table 2 QTL detected for the binomial trait of wheat fertility using the overdispersion method

QTL	*Chromosome*	*Position (cM)*	*Marker*[a]	*Estimate*[b]	*StdErr*[c]	*LOD*
1	1	0.00	1	0.2171	0.0266	14.40
2	2	0.00	1	−1.0517	0.0278	308.75
3	2	2.12	1	0.9841	0.0283	260.97
4	2	10.96	1	0.3985	0.0303	37.36
5	2	16.16	0	−0.7670	0.0311	131.24
6	2	28.70	0	1.6423	0.0306	621.89
7	2	38.29	1	−0.1356	0.0272	5.37
8	3	67.32	1	0.1635	0.0273	7.78
9	4	9.20	1	0.5755	0.0293	83.30
10	4	14.92	1	−0.6445	0.0300	99.78
11	5	0.00	1	0.3878	0.0281	41.17
12	5	20.83	0	−0.8852	0.0336	150.14
13	5	39.87	0	0.5121	0.0292	66.67
14	5	63.60	0	−0.6898	0.0321	100.17
15	5	82.68	0	0.6414	0.0301	98.17

Abbreviation: QTL, quantitative trait loci.
[a]This column indicates whether the QTL overlaps with a true marker (1) or a pseudo marker (0).
[b]This column gives the estimated QTL effect.
[c]This column shows the s.e. of the estimated QTL effect.

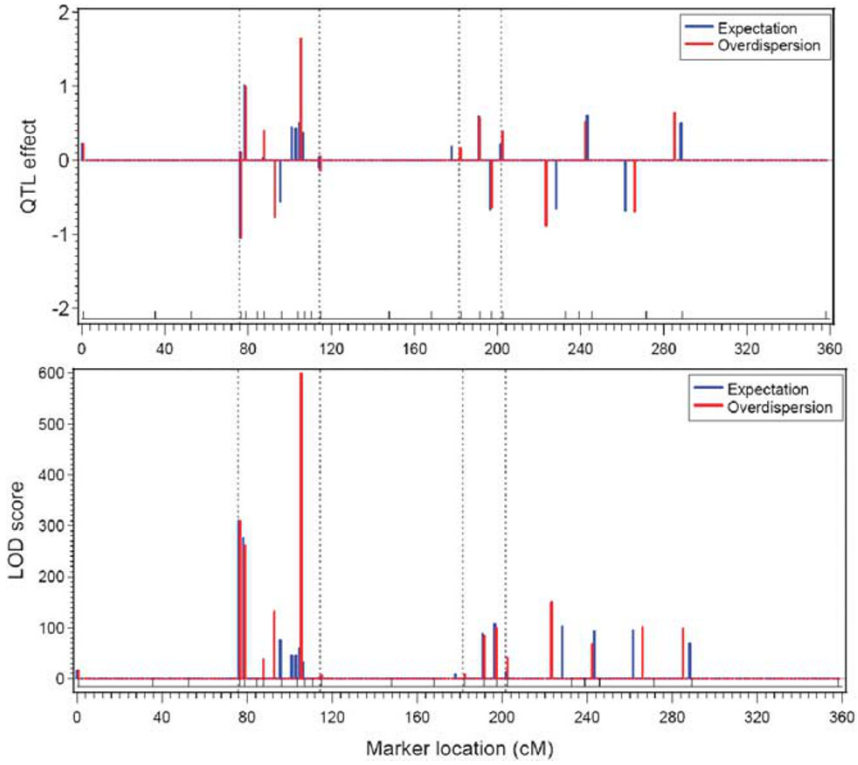

Figure 5 Binomial trait analysis of the wheat experiment using the expectation method (blue) and the overdispersion method (red). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis.

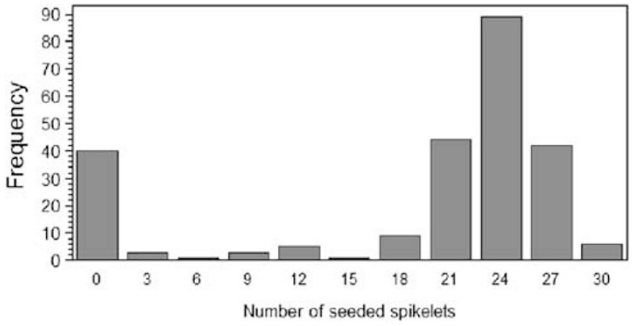

Figure 6 Frequency distribution of the number of seeded spikelets of the F_2 wheat population. Among the 243 plants, 39 of them had no seeds (zero category).

able to compare the biases and the MSE of the estimated QTL effects. We chose an alternative approach for evaluating the two methods, that is the leave-one-out cross validation (Picard and Cook, 1984). The cross validation approach only evaluates the predictabilities of the models. For the purpose of molecular breeding and marker assistant selection, higher predictability is more preferable. For the purpose of gene cloning, the biases of QTL effect and location estimates are of major concern. We used the Pearson correlation coefficient ($r_{y\hat{y}}$) between the observed (y) and predicted ($\hat{y}$) trait values as a measurement of the predictability. The Pearson correlation coefficients for the expectation and overdispersion methods were 0.5166 and 0.5290, respectively. The overdispersion method showed a slight advantage over the expectation method. We also examined the prediction errors defined by

$$\mathrm{PE} = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \tag{32}$$

for the two methods. The results of PEs were 0.100563 and 0.098658, respectively, for the expectation and the overdispersion methods. The PE comparison is consistent with the Pearson correlation comparison. We concluded that incorporation of overdispersion does show the expected benefit (increase in predictability) in QTL mapping over the simple expectation method.

Binary trait. Among the 243 plants, 39 of them did not have seeds at all. The frequency distribution of the number of seeded spikelets is shown in Figure 6. It appears that the zero category was inflated. The binomial data analysis did not differentiate QTL responsible for seed presence and absence. We now defined a binary trait as seed presence/absence and used the two methods (expectation and overdispersion) to analyze the binary trait. The estimated QTL effect profiles are shown in the top panel of Figure 7 and the LOD score profiles are depicted in the bottom panel of the same figure. The two methods appeared to generate much the same result. Using the LOD 5.29 criterion, we only detected a single QTL at position 28.71 cM of chromosome II (cumulative position 104.60 cM). This QTL is the same one as that detected for the binomial trait (the largest QTL for the binomial trait) detected by the overdispersion method. Our conclusion was that, except for this particular QTL, the multiple QTL detected for the binomial trait reported early were all responsible for the variation of the number of seeded spikelets, not the seed presence/absence trait. The leave-one-out cross validation analysis did not show much difference for the two methods. The Pearson correlation coefficients between the observed and predicted trait values were 0.4715 and 0.4729, respectively, for the expectation and overdispersion methods. The corresponding PE's were 0.104914 and 0.104721, respectively. Both criteria indicate that the overdispersion method is better than the expectation method.

DISCUSSION

The overdispersion method for handling missing genotypes was proposed by Xu and Hu (2010) for IM under the GLM framework. We examined this method and an additional one (expectation method) under the GLMM framework for mapping multiple QTL. The GLMM and GLM are different and thus the extension is not a trivial task. The overdispersion method consistently showed advantages over the expectation method in both the simulated data and real data analyses and in both the binomial and binary trait analyses. Based on the visual plots of the estimated QTL effects, the advantages appeared to be marginal. Then why should we bother to develop such a method, given the observed marginal advantage? First, the overdispersion method does not require much more computational load than the expectation method. The computational times of the two methods are pretty much the same when we used the numerical differentiation packages to evaluate the first and second partial derivatives. Therefore, we should take any opportunity to extract maximum information from the data; even a slight advantage is worth the effort. Secondly, the simulation experiments are always limited. It is hard to simulate all possible scenarios so that the advantages of the overdispersion method are fully exposed. In some situations, the advantage may be obvious and we may simply fail to identify those situations. Thirdly, the two methods for the wheat data analysis of the binomial trait already demonstrated some interesting differences that are worth of discussion. The largest QTL detected by the overdispersion method was split into several smaller QTL by the expectation method. The cross validation analysis showed that the overdispersion method gave a better prediction, implying that the single large QTL may most likely represent the truth. The binary data analysis of the wheat experiment showed that the same locus also had a large effect on the binary trait. This time both methods showed a single large QTL. This observation further supports the single large QTL hypothesis. Without the overdispersion method, we would not have such a confidence of this single large QTL.

The advantage of the overdispersion method will diminish as the marker density increases. In the situation where the entire genome is sequenced, the two methods would converge to the same result because genotypes of all markers will be observed. However, full genome sequences for most species are not expected to happen soon. In addition, missing genotypes may still exist due to human and technical errors in experiments. Therefore, the missing genotype handling methods remain useful in the foreseeable future.

The GLMM is sufficiently general so that it can handle traits with any distributions as long as the likelihood function is programmable. The normal distribution for the QTL effects may be substituted by other distributions. Explicit expressions of the derivatives are not required to implement the Newton–Raphson updates. Recently, Yi and Banerjee (2009) developed a hierarchical GLM for mapping discrete trait QTL. They used the pseudo likelihood approach to approximate the observed log likelihood function. The authors used an EM algorithm to estimate the QTL effects but they treated $\{\beta,\gamma\}$ as parameters and G as missing values. In addition, Yi and Banerjee (2009) only considered marker analysis with missing marker genotypes replaced by the conditional expectation, which is equivalent to the expectation method of this study. However, they only considered missing marker genotypes in the sense that majority of the individuals are genotyped. The missing genotypes in their study were solely

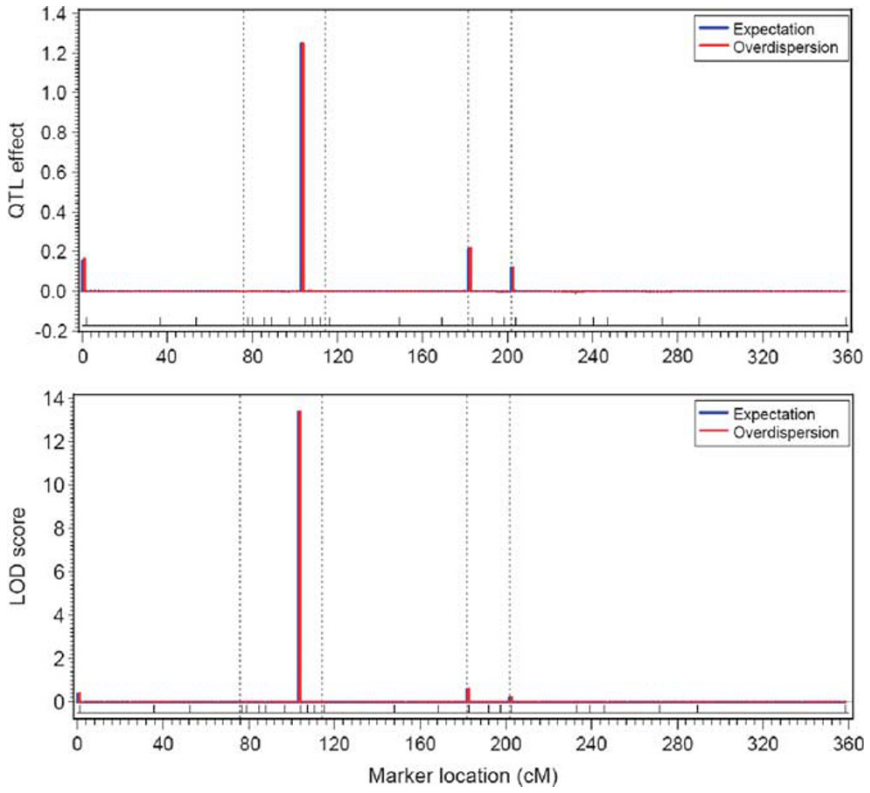

Figure 7 Binary trait (seed presence/absence) analysis using the expectation method (blue) and the overdispersion method (red). The top panel shows the estimated QTL effects and the bottom panel shows the LOD scores. Chromosomes are separated by the dotted vertical lines. Positions of true markers are indicated by the inward ticks on the horizontal axis.

caused by technical or human errors. They did not insert pseudo marker in every few centiMorgan to saturate the genome.

Responding to a reviewer's suggestion, we analyzed both the binomial and binary traits of the wheat experiment using the LMM by ignoring the discrete nature of the traits. The correct model should be the GLMM, but we used the LMM as an *ad hoc* model to analyze the discrete traits. The results are depicted in Supplementary Figure S1 for the binomial trait and Supplementary Figure S2 for the binary trait. Supplementary Figure S1 shows the estimated QTL effects and LOD scores of the LMM analysis for the binomial trait. Only one large QTL was detected using this *ad hoc* model. Comparing Supplementary Figure S1 here with Figure 5 of the main text, we can see that many small- to median-sized QTL detected by the GLMM were missed. Results of leave-one-out cross validation are shown in Supplementary Table S1. The Pearson correlation coefficients between the observed and predicted trait values were dropped from 0.517 (expectation) and 0.529 (overdispersion) to 0.495 (expectation) and 0.497 (overdispersion). This means that the median-sized QTL detected by GLMM do contribute to the binomial trait variation, and ignoring the discrete nature of the trait has decreased the predictability of the model. The binary trait comparison between GLMM and LMM favors even more for the GLMM (see Supplementary Figure S2 and Table S2 of the Supplementary material).

GLM or GLMM represents an important area of statistics. It was particularly designed to deal with discrete traits or other traits deviating from a normal distribution. In statistics, people rarely argue the suitability GLMM given that LMM is already available for normally distributed traits. In case–control studies for human diseases, logistical regression (belongs to GLM) is often used to detect disease QTL (Hunter *et al*, 2007) because case (designated by 1) and control (designated by 0) consist of the two binary states of the disease outcome. People rarely analyze the 0-1 binary trait using the simple regression analysis by ignoring the discrete nature of the trait. The situation is different for QTL mapping in plants and animals. Every time a new method is developed for discrete traits, the investigator must face challenges from peers about how much improvement can be achieved if the discrete nature of the trait is ignored. These challenges repeatedly occurred and may largely credit (or blame) to the works by Visscher *et al* (1996) and Rebai (1997) who showed marginal improvement of GLM over LM for binary trait QTL mapping when the binary trait is treated as if it were continuous. Rao and Xu's (1998) conclusion about the *ad hoc* treatment of categorical trait analysis was slightly different. They found that if a categorical trait is analyzed using simple linear models, the power and accuracy of QTL parameter estimation can be reduced substantially if the categorical nature of the trait is ignored. Although from practical point of view, it is true that the loss of power and accuracy may be marginal when discrete traits are treated as continuous ones, the GLM or GLMM is built based on a rigorous statistical foundation and thus its suitability should not be argued. Especially, in the era of high power computing, one should not use a suboptimal algorithm on knowing the availability of the optimal algorithm. On the other hand, if an investigator presents result of a binary trait analysis using simple method by treating the binary phenotype as a continuous trait, then the investigator will often face criticism from the peers for not using the correct model, given the availability of GLM or GLMM. For the benefit of these investigators, the new GLMM approach provides a useful tool for correctly analyzing their data to avoid rejection of their fine works.

DATA ARCHIVING

Simulated data, the test dataset from Dou *et al.* (2009) and SAS scripts for analyzing these datasets have been deposited at Dryad: doi:10.5061/dryad.mn159hq6.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Boer MP, Wright D, Feng L, Podlich DW, Luo L, Cooper M *et al.* (2007). A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. *Genetics* **177**: 1801–1813.

Breslow NE, Clayton DG (1993). Approximate inference in generalized linear mixed models. *J Am Stat Assoc* **88**: 9–25.

DeGroot MH (2004). *Optimal Statistical Decision*. John Wiley & Sons: Hoboken, New Jersey.

Dou B, Hou B, Xu H, Lou X, Chi X, Yang J *et al.* (2009). Efficient mapping of a female sterile gene in wheat (Triticum aestivum L). *Genet Res Camb* **91**: 337–343.

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* **33**: 1–22.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis*. Chapman & Hall, New York.

Henderson CR (1950). Estimation of genetic parameters (Abstract). *Ann Math Statist* **21**: 309–310.

Hoerl AE, Kennard RW (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**: 55–67.

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE *et al.* (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**: 870–874.

Jiang C, Zeng Z-B (1997). Mapping quantitative trait loci with dominance and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.

Kidd KK, Ott J (1984). Power and sample size in linkage studies. Human Gene Mapping 7(1984): Seventh International Workshop on Human Gene Mapping. *Cytogenet Cell Genet* **37**: 510–511.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman & Hall: New York.

McCulloch CE, Neuhaus JM (2005). *Generalized Linear Mixed Model. Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd: San Francisco.

McGilchrist CA (1994). Estimation in generalized mixed models. *JR Stat Soc B* **56**: 61–69.

Picard R, Cook D (1984). Cross-validation of regression models. *J Am Stat Assoc* **79**: 575–583.

Rao SQ, Xu S (1998). Mapping quantitative trait loci for ordered categorical traits in four—way crosses. *Heredity* **81**: 214–224.

Rebai A (1997). Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. *Genet Res, Camb* **69**: 69–74.

Risch N (1991). A note on multiple testing procedures in linkage analysis. *Am J Hum Genet* **48**: 1058–1064.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J R Stat Soc B* **58**: 267–288.

Visscher PM, Haley CS, Knott SA (1996). Mapping QTLs for binary traits in backcross and F2 populations. *Genet Res Camb* **68**: 55–63.

Vonesh EF (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika* **83**: 447–452.

Wald A (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans Amer Math Soc* **54**: 426–482.

Wedderburn RWM (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**: 439–447.

Wolfinger R, O'Connell M (1993). Generalized linear mixed models: A pseudo-likelihood approach. *J Statist Comput Simul* **48**: 233–243.

Xu S, Hu Z (2010). Generalized linear model for interval mapping of quantitative trait loci. *Theor Appl Genet* **121**: 47–63.

Xu S, Yi N (2000). Mixed model analysis of quantitative trait loci. *Proc Nat Acad Sci USA* **97**: 14542–14547.

Yi N, Banerjee S (2009). Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics* **181**: 1101–1113.

Ypma TJ (1995). Historical development of the Newton-Raphson method. *SIAM Review* **37**: 531–551.

Zou H (2006). The adaptive Lasso and its oracle properties. *J Am Stat Assoc* **101**: 1418–1429.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)