Title

Generalized linear model for interval mapping of quantitative trait loci

Permalink

https://escholarship.org/uc/item/0ws2b0s8

Journal

Theoretical and Applied Genetics: International Journal of Plant Breeding Research, 121(1)

ISSN

1432-2242

Authors

Xu, Shizhong
Hu, Zhiqiu

Publication Date

2010-06-01

DOI

10.1007/s00122-010-1290-0

Peer reviewed

ORIGINAL PAPER

Generalized linear model for interval mapping of quantitative trait loci

Shizhong Xu · Zhiqiu Hu

Abstract We developed a generalized linear model of QTL mapping for discrete traits in line crossing experiments. Parameter estimation was achieved using two different algorithms, a mixture model-based EM (expectation–maximization) algorithm and a GEE (generalized estimating equation) algorithm under a heterogeneous residual variance model. The methods were developed using ordinal data, binary data, binomial data and Poisson data as examples. Applications of the methods to simulated as well as real data are presented. The two different algorithms were compared in the data analyses. In most situations, the two algorithms were indistinguishable, but when large QTL are located in large marker intervals, the mixture model-based EM algorithm can fail to converge to the correct solutions. Both algorithms were coded in C++ and interfaced with SAS as a user-defined SAS procedure called PROC QTL.

Introduction

Interval mapping (Lander and Botstein 1989) is the most commonly used method for mapping quantitative trait loci (QTL). The method usually applies to quantitative traits, i.e., traits that have a continuous distribution. In agricultural crops, the phenotypes of some traits are measured as discrete variables. Ordinal traits, e.g., disease resistance in plants and animals, are typical examples of such discrete

S. Xu (✉) · Z. Hu
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA
e-mail: shizhong.xu@ucr.edu

traits. These traits are usually modeled by a multinomial distribution. Traits measured as counts, such as liter size in pigs and tiller number in rice, are usually modeled by the Poisson distribution. Binomial traits are also common in agricultural experiments, such as the ratio of germinated seeds to the total number of seeds planted. These traits are not normally distributed. Although some transformations, such us the log transformation or the more general Box–Cox transformation (Box and Cox 1964), can be used to improve the normality of the traits, not all traits can be transformed. For example, a binary trait has no appropriate transformation to make it normal.

The generalized linear model approach (McCullagh and Nelder 1989; Wedderburn 1974) is the most appropriate method for analyzing traits with non-normal distribution and it has been widely used in statistics for parameter estimation. Generalized linear model takes advantage of all theory and methods developed in the usual linear model methodology (Searle 1997). It has been applied to QTL mapping for some special traits, e.g., binary traits (Deng et al. 2006; Xu and Atchley 1996; Yi and Xu 1999a, b, 2000), ordinal traits (Hackett and Weller 1995; Rao and Xu 1998) and Poisson traits (Cui et al. 2006; Cui and Yang 2009). Depending on the special characteristics of the traits, distribution-specific generalized linear models have been developed for these traits. These methods are not sufficiently general to extend to all traits that can be modeled by the generalized linear model. For example, the EM algorithm developed by Xu et al. (2003, 2005) only applies to binary and ordinal traits. They treated both the marker genotypes and the latent variable as missing values. Although parameter estimation under the EM algorithm is simple, the information matrix of the estimated parameters is difficult to calculate. A more comprehensive analysis of the generalized linear model applied to QTL mapping is

the seminal paper by Lange and Whittaker (2001). They adopted the generalized estimating equations (GEE) approach to analyze multiple traits with arbitrary combination of continuous and discrete trait components. The method replaces the unobserved QTL genotypes by the conditional expectations of the genotype indicator variable given flanking marker information. The uncertainties of the genotype indicator variables are ignored. In addition, detailed formulas for the partial derivatives of the expectation of the data with respect to the parameters are not given.

When there are no missing values, commercial software packages are available to estimate parameters under a wide range of distribution of the traits, e.g., SAS (SAS Institute 1999) and GENESTAT (PHOEBE Biostatistics Group 2007, http://www.genestat.org). Although these programs may handle missing values using the imputation algorithm or the EM algorithm (Dempster et al. 1977), the missing patterns handled by these commercial programs are usually different from that of QTL mapping. In interval mapping, genotypes are missing for every individual at a putative QTL position unless the QTL overlaps with a fully informative marker. Special mixture models are required in QTL mapping (Lander and Botstein 1989). Hackett and Weller (1995) and Xu and Atchley (1996) were the first group of people to use the EM algorithm to estimate QTL parameters for ordinal traits, but they did not investigate the variance–covariance matrix of the estimated QTL effects. Hackett and Weller (1995) took advantage of an existing commercial software (GENESTAT) for generalized linear model analysis by iteratively calling the subroutine for generalized linear model with non-missing genotypes and calculating the weight (posterior probability of QTL genotype). The attractive property of that method is that users do not have to write their own code for the maximization step, which is conducted by the commercial software. They only incorporated the expectation step into the existing program for parameter updating. As a result, no variance–covariance matrix of the estimated parameters was provided. Xu et al. (2003) developed an EM algorithm for binary data and used the Monte Carlo simulation approach to obtain the Louis' (1982) information matrix. The variance–covariance matrix of the estimated parameters can be approximated by the inverse of the information matrix. This method is computationally intensive due to the use of Monte Carlo simulation for approximate integration.

In the statistics literature, generalized linear model with missing covariates is often handled with the EM algorithm (Horton and Laird 1998). However, other methods are also available, as summarized by Ibrahim et al. (2005), who reviewed four general approaches: maximum likelihood method implemented via the EM algorithm by the method of weights (Horton and Laird 1998), multiple imputation

(Rubin 1987), fully Bayesian (Ibrahim et al. 2002) and weighted estimation equation (Robins and Ritov 2001). Ibrahim et al. (2005) concluded that the most accurate method is the fully Bayesian method, although the method is associated with a high cost in terms of computing time. The second best method is the EM algorithm via the method of weights. Application of this method to interval mapping has not been analyzed.

The mixture model-based maximum likelihood estimation of parameters has a slow convergence speed. The computational intensity is high within each round of the iteration. Therefore, an approximate method that improves the computational efficacy with little loss in power may be desirable. Xu (1998) developed a weighted least square approximation for QTL mapping of normally distributed traits. Recently, Han and Xu (2008) made a further improvement of the weighted least square estimation. Their idea can be applied to the generalized linear model. Therefore, we will present an approximate method to improve the computational efficiency over the mixture model maximum likelihood method.

Model and methods

Generalized linear model

We will use ordinal trait QTL mapping as an example to introduce the generalized linear model. Extension to other traits will be given later. Suppose that a disease phenotype of individual j $(j = 1, \ldots, n)$ is measured by an ordinal variable denoted by $T_j = 1, \ldots, p + 1$, where $p + 1$ is the total number of disease classes and n is the sample size. Let $Y_j = \{Y_{jk}\}$, $\forall k = 1, \ldots, p + 1$, be a $(p + 1) \times 1$ vector to indicate the disease status of individual j. The kth element of Y_j is defined as

$$Y_{jk} = \begin{cases} 1 & \text{if} \quad T_j = k \\ 0 & \text{if} \quad T_j \neq k \end{cases} \tag{1}$$

Note that the phenotype of the ordinal trait has been formulated as a multivariate Bernoulli variable (or multinomial variable with sample size one). Different link functions can be used to describe the relationship of the observed ordinal phenotype and the genetic effects of QTL. The most commonly used link functions are the logit and probit link functions (McCullagh and Nelder 1989). Here, we described the probit link function and leave the logit link function in Appendix D for interested readers. Under the probit link function, the expectation of Y_{jk} is

$$\mu_{jk} = E(Y_{jk}) = \Phi(\alpha_k + X_j\beta + Z_j\gamma) - \Phi(\alpha_{k-1} + X_j\beta + Z_j\gamma) \tag{2}$$

where $\alpha_k (\alpha_0 = -\infty \text{ and } \alpha_{p+1} = +\infty)$ is the intercept, β is a $q \times 1$ vector for some systematic effects (not related to the

effects of QTL), and γ is an $r \times 1$ vector for the effects of a quantitative trait locus. The symbol $\Phi(\cdot)$ is the standardized cumulative normal function. The design matrix X_j is assumed to be known, but Z_j may not be fully observed because it is determined by the genotype of individual j for the locus of interest. We will defer the definition of Z_j to the next section. Because the link function is probit, this type of analysis is called probit analysis. Let $\mu_j = \{\mu_{jk}\}$ be a $(p+1) \times 1$ vector. The expectation for vector Y_j is $E(Y_j) = \mu_j$ and the variance–covariance matrix of Y_j is

$$V_j = \text{var}(Y_j) = \text{diag}(\mu_j) - \mu_j \mu_j^T \qquad (3)$$

The method to be developed requires the inverse of matrix V_j. However, V_j is not of full rank (see Appendix C for proof). We can use a generalized inverse of V_j, such as $V_j^- = \text{diag}^{-1}(\mu_j)$, in place of V_j^{-1} (see Appendix C for the generalized inverse). The weight matrix is

$$W_j = \text{diag}^{-1}(\mu_j) \qquad (4)$$

The parameter vector is denoted by $\theta = \{\alpha, \beta, \gamma\}$ with a dimensionality of $(p + q + r) \times 1$. Once μ_j and W_j and are defined, the reweighted least square method of Wedderburn (1974) can be used to estimate the parameters.

Mixture model maximum likelihood estimation

When the design matrix Z_j is fully observed, the maximum likelihood solution of parameters can be solved in a straightforward manner using the reweighted least squares approach (Wedderburn 1974). For the paper to be self contained, the iteration equation and the information matrix in the situation of no missing value is briefly described. The Newton–Raphson iteration is

$$\theta^{(t+1)} = \theta^{(t)} + \Delta\theta \qquad (5)$$

where

$$\Delta\theta = \left[\sum_{j=1}^{n} D_j^T W_j D_j\right]^{-1} \left[\sum_{j=1}^{n} D_j^T W_j (Y_j - \mu_j)\right] \qquad (6)$$

is the iteratively reweighted least squares formula for parameter updating (increment of the parameter from iteration t to $t + 1$). Matrix D_j is the partial derivative matrix of μ_j with respect to θ,

$$D_j = \begin{bmatrix} \frac{\partial \mu_j}{\partial \alpha^T} & \frac{\partial \mu_j}{\partial \beta^T} & \frac{\partial \mu_j}{\partial \gamma^T} \end{bmatrix} \qquad (7)$$

with a dimensionality of $(p + 1) \times (p + q + r)$. Once the iteration process converges, the information matrix is automatically given, as it is the coefficient matrix in the last step of the iteration (Wedderburn 1974),

$$I(\theta) = \sum_{j=1}^{n} D_j^T W_j D_j \qquad (8)$$

From this information matrix, the variance–covariance matrix of estimated parameters is calculated because the inverse of the information matrix approximately equals the variance–covariance matrix of the estimated parameters.

In QTL mapping for ordinal traits, the generalized linear model in its original form can be applied if one is interested in individual marker analysis because, at markers, the genotypes are observed and thus matrix Z_j is known. In interval mapping, however, effects of QTL that are located between markers should also be estimated. In this case, the genotypes of QTL are not observed and must be inferred using flanking marker genotypes. This is a typical missing value problem. The missing value Z_j can be inferred from linked markers. We now use an F_2 population as an example to show how to handle the missing value of Z_j. Let

$$p_j(g) = \Pr(Z_j = H_g | \text{marker}), \quad \forall g = 1, 2, 3 \qquad (9)$$

be the conditional probability of QTL genotype given marker information, where the marker information can be either drawn from two flanking markers (interval mapping, Lander and Botstein 1989) or multiple markers (multipoint analysis, Jiang and Zeng 1997). For an F_2 population, matrix H_g is defined as the gth row of matrix H, where

$$H = \begin{bmatrix} +1 & 0 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad (10)$$

Corresponding to the definition of H, the QTL effect vector γ is defined as $\gamma = \begin{bmatrix} a & d \end{bmatrix}^T$ where a and d represents the additive and dominance effects, respectively. When Z_j is missing, the generalized linear model becomes a generalized linear mixture model. Under the mixture model approach, we need to define the genotype-specific expectation, the genotype-specific variance matrix and the genotype-specific derivatives for each individual. Let

$$\mu_{jk}(g) = E(Y_{jk}|g) \\ = \Phi(\alpha_k + X_j\beta + H_g\gamma) - \Phi(\alpha_{k-1} + X_j\beta + H_g\gamma) \qquad (11)$$

be the genotype-specific expectation of Y_{jk} when j takes the gth genotype for $g = 1, 2, 3$. The corresponding genotype specific weight matrix is

$$W_j(g) = \text{diag}^{-1}[\mu_j(g)] \qquad (12)$$

Let $D_j(g)$ be the genotype-specific partial derivatives of the expectation with respect to the parameters,

$$D_j(g) = \begin{bmatrix} \frac{\partial \mu_j(g)}{\partial \alpha^T} & \frac{\partial \mu_j(g)}{\partial \beta^T} & \frac{\partial \mu_j(g)}{\partial \gamma^T} \end{bmatrix} \qquad (13)$$

The closed form of matrix $D_j(g)$ is given in Appendix A. The increment of parameters in the iteration is

$$\Delta\theta = \left[\sum_{j=1}^{n}\sum_{g=1}^{3}p_j^*(g)D_j^T(g)W_j(g)D_j(g)\right]^{-1}$$
$$\times \left[\sum_{j=1}^{n}\sum_{g=1}^{3}p_j^*(g)D_j^T(g)W_j(g)\big(Y_j-\mu_j(g)\big)\right] \quad (14)$$

where $p_j^*(g)$ is the posterior probability of QTL genotype after the phenotype information is incorporated and is given by

$$p_j^*(g) = \frac{p_j(g)p(Y_j|g)}{\sum_{g'=1}^{3}p_j(g')p(Y_j|g')} \quad (15)$$

where

$$p(Y_j|g) = \prod_{k=1}^{p+1}\mu_{jk}^{Y_{jk}}(g) = Y_j^T\mu_j(g) \quad (16)$$

is the multinomial probability. Derivation of the EM algorithm (Eq. 14) is given in Appendix B.

Unfortunately, the information matrix under the EM algorithm is not identical to the coefficient matrix of the reweighted least squares equation; rather, it has to be adjusted for the information loss due to the uncertainty of QTL genotypes. The Louis' (1982) adjustment of the information matrix is

$$I(\theta) = \sum_{j=1}^{n}E\big[B_j\big(\theta|Z_j\big)\big] - \sum_{j=1}^{n}\text{var}\Big[S_j(\theta|Z)_j\Big] \quad (17)$$

The first term in the above expression (Eq. 17) is

$$E\big[B_j(\theta|Z_j)\big] = \sum_{g=1}^{3}p_j^*(g)D_j^T(g)W_j(g)D_j(g) \quad (18)$$

which is the expected value of the negative Hessian matrix. The second term of Eq. (17) is

$$\text{var}\big[S_j(\theta|Z_j)\big] = \sum_{g=1}^{3}p_j^*(g)\big[S_j(\theta|g) - \bar{S}_j(\theta)\big]\big[S_j(\theta|g) - \bar{S}_j(\theta)\big]^T \quad (19)$$

which is the variance matrix of the score vector, where

$$S_j(\theta|g) = D_j^T(g)W_j(g)\big(Y_j-\mu_j(g)\big) \quad (20)$$

and

$$\bar{S}_j(\theta) = \sum_{g=1}^{3}p_j^*(g)S_j(\theta|g) \quad (21)$$

Approximation under the heterogeneous variance model

The EM implemented mixture model approach described above is computationally intensive due to (1) the mixture model itself and (2) the extra step in calculating the

information matrix of parameters. Here, we introduce an approximate method that replaces the mixture model by a heterogeneous residual variance model. Let

$$U_j = E(Z_j) = \sum_{g=1}^{3}p_j(g)H_g \quad (22)$$

be the conditional expectation of Z_j given marker information and

$$\Sigma_j = \text{var}\big(Z_j\big) = \sum_{g=1}^{3}p_j(g)\big(H_g - U_j\big)^T\big(H_g - U_j\big) \quad (23)$$

be the corresponding conditional variance–covariance matrix of Z_j. If Z_j were observed, we would have

$$\mu_{jk} = E\big(Y_{jk}\big) = \Phi\big(\alpha_k + X_j\beta + Z_j\gamma\big) - \Phi\big(\alpha_{k-1} + X_j\beta + Z_j\gamma\big) \quad (24)$$

When Z_j is missing, we can replace Z_j by U_j and adjust for the over dispersion caused by the substitution,

$$\mu_{jk} = E\big(Y_{jk}\big)$$
$$\approx \Phi\left[\frac{1}{\sigma_j}\big(\alpha_k + X_j\beta + U_j\gamma\big)\right] - \Phi\left[\frac{1}{\sigma_j}\big(\alpha_{k-1} + X_j\beta + U_j\gamma\big)\right] \quad (25)$$

where

$$\sigma_j^2 = \gamma^T\Sigma_j\gamma + 1 \quad (26)$$

is the heterogeneous dispersion (see Xu 1998).

We are now in a position to explain the over dispersion. It makes more sense to introduce the liability model before we explain the overdispersion. Let

$$l_j = X_j\beta + Z_j\gamma + \varepsilon_j \quad (27)$$

be the liability for individual j, where $\varepsilon_j \sim N(0, \sigma^2)$ is the residual error of the liability. The liability is a latent variable that controls the observed ordinal phenotype through a series of thresholds $\alpha = [\alpha_1, \ldots, \alpha_p]^T$ with $T_j = k$ if $a_{k-1} \le l_j < \alpha_k$. The residual error variance is not estimable and thus we set $\sigma^2 = 1$. When Z_j is replaced by U_j, the variance of the liability becomes

$$\sigma_j^2 = \text{var}\big(l_j\big) = \gamma^T\text{var}\big(Z_j\big)\gamma + \text{var}\big(\varepsilon_j\big) = \gamma^T\Sigma_j\gamma + 1 \quad (28)$$

Because $\sigma_j^2 \ge 1$, the model is called overdispersion. Since σ_j^2 varies from one individual to another, the model is also called the heterogeneous residual variance model. To adjust for the heterogeneous overdispersion, we replace $\alpha_k + X_j\beta + U_j\gamma$ by $\frac{1}{\sigma_j}\big(\alpha_k + X_j\beta + U_j\gamma\big)$. This adjustment serves as a way to standardize the liability so that the adjusted liability has a mean $\frac{1}{\sigma_j}\big(\alpha_k + X_j\beta + U_j\gamma\big)$ and a unity variance.

Unlike the mixture model, the expectation μ_j is no longer a function of g. Similarly, the weight matrix is

$$W_j = \text{diag}^{-1}(\mu_j) \tag{29}$$

This modification leads to a change in matrix D_j, the partial derivatives of μ_j with respect to the parameters, which is given in Appendix A. The same iteration equation given in Eq. (6) is used here for the heterogeneous residual variance model. The attractive property of this approximation is that as the information matrix is given in Eq. (8) no adjustment is required because EM algorithm is not used here for the approximation.

Extension to other traits

Ordinal traits are the most commonly observed discrete traits in QTL mapping experiments. Other discrete traits also commonly seen in QTL mapping experiments are binary traits, binomial traits and Poisson traits. This section is dedicated to these commonly observed discrete traits. The mixture model algorithm and the heterogeneous variance approximation apply to all traits as long as the traits can be analyzed under the generalized linear model. To apply the algorithms to any specific trait, we only need to find: (1) the distribution of the trait (probability density of the data point), (2) the expectation of the data point, (3) the weight (inverse of the variance) of the data point and (4) the partial derivative of the expectation with respect to the parameters. We now introduce these discrete traits and leave the details of the formulas in Appendix A for interested readers.

Binary traits

Binary traits can be treated as a special case of ordinal traits with $p = 1$. Without any modification, the method developed for ordinal traits can be applied to binary traits with $Y_j = \begin{bmatrix} Y_{j1} & Y_{j2} \end{bmatrix}^T$ defined as a 2×1 vector. Each of the two components is defined as a binary variable and the two components are perfectly correlated. Here, we simplify the problem by defining Y_j as a univariate binary trait. This univariate treatment not only saves computing time but also simplifies the notation. We now use the univariate definition to define the binary phenotype,

$$Y_j = \begin{cases} 1 & \text{for presence of the trait} \\ 0 & \text{for absence of the trait} \end{cases} \tag{30}$$

The expectation and variance of the phenotype given the parameter value are

$$E(Y_j) = \mu_j = \Phi(X_j\beta + Z_j\gamma) \tag{31}$$

and

$$\text{var}(Y_j) = V_j = \mu_j(1 - \mu_j) \tag{32}$$

respectively. The probability distribution is

$$p(Y_j) = \mu_j^{Y_j}(1 - \mu_j)^{1-Y_j} \tag{33}$$

Details for the mixture model and the heterogeneous variance model are given in Appendix A.

Binomial traits

Let n_j be the number of trials observed from individual j and m_j be the number of events happened to individual j. The binomial phenotype for individual j is defined as $Y_j = m_j/n_j$ (expressed as a fraction so that $0 \leq Y_j \leq 1$). Under the probit model, the expectation and the variance of the phenotype are

$$E(Y_j) = \mu_j = \Phi(X_j\beta + Z_j\gamma) \tag{34}$$

and

$$\text{var}(Y_j) = V_j = \frac{1}{n_j}\mu_j(1 - \mu_j) \tag{35}$$

respectively. The weight is

$$W_j = V_j^{-1} = \frac{n_j}{\mu_j(1 - \mu_j)} \tag{36}$$

The probability distribution is

$$p(Y_j) = \frac{(n_j)!}{(n_j)!(n_j - n_jY_j)!}\mu_j^{n_jY_j}(1 - \mu_j)^{n_j(1-Y_j)} \tag{37}$$

Details regarding the partial derivatives of the expectation with respect to the parameters both in the mixture model and in the heterogeneous variance model are given in Appendix A.

Poisson traits

Let $Y_j = 0, 1, \ldots, \infty$ be the phenotype of a Poisson trait. The expectation and the variance of the phenotype are equivalent, $E(Y_j) = \text{var}(Y_j) = \mu_j$, where

$$\mu_j = V_j = \exp(X_j\beta + Z_j\gamma) \tag{38}$$

The weight is

$$W_j = V_j^{-1} = \frac{1}{\exp(X_j\beta + Z_j\gamma)} \tag{39}$$

The probability density is

$$p(Y_j) = \frac{\mu_j^{Y_j}}{(Y_j)!}\exp(-\mu_j) \tag{40}$$

Details regarding the partial derivatives of the expectation with respect to the parameters both in the mixture model and in the heterogeneous variance model are given in Appendix A.

Hypothesis tests

Two different hypothesis tests are provided, the likelihood ratio test and the Wald test (1949). The likelihood ratio test

requires evaluation of different likelihood functions (full model and reduced model). The Wald test is much simpler than the likelihood ratio test statistic because it requires only $\text{var}(\hat{\theta}) \approx I^{-1}(\hat{\theta})$ and the estimated parameters to form the test statistics. Let $V_\gamma = \text{var}(\hat{\gamma})$ be the subset of matrix $\text{var}(\hat{\theta})$, the Wald test statistics for the null hypothesis, $H_0 : \gamma = 0$, is

$$\text{Wald} = \hat{\gamma}^T V_\gamma^{-1} \hat{\gamma} \tag{41}$$

Recall that $\gamma = [a \; d]^T$ represents both the additive and dominance effects. Under the null hypothesis, the Wald test follows approximately the Chi-square distribution with two degrees of freedom. Therefore, the Wald test is comparable to the likelihood ratio test statistics (McCullagh and Nelder 1989). When testing a single parameter (either a or d), the Wald test statistics is equivalent to the Chi-square test with one degree of freedom. Therefore, some people used the Wald test and the Chi-square test interchangeably (e.g., Han and Xu 2008).

Applications

Binary traits

This example demonstrates the application of the generalized linear model to QTL mapping for binary traits in wheat. The experiment was conducted by Dou et al. (2009) who made the data available to us for this analysis. A female sterile line XND126 and an elite cultivar Gaocheng 8901 with normal fertility were crossed for genetic analysis of female sterility measured as a binary trait. The parents and their F_1 and F_2 progeny were planted at Huaian experimental station in China for the 2006–2007 growing season under the normal autumn sowing condition. The mapping population was an F_2 family consisting of 234 individual plants. The binary trait was the presence of seed setting of the female plants. About five-sixth of the F_2 progeny had seeded splikelets (phenotype 1) and the remaining one-sixth plants did not have seeded splikelet (phenotype 0). This is a typical binary trait regarding the presence of seeds. Among the plants that had seeded spikelets, the number of seeded spikelets varied and can be treated as a binomial trait for further QTL analysis (see Sect. "Binomial traits" for binomial trait QTL mapping). A total of 28 SSR markers were used in this experiment. These markers covered 5 chromosomes of the wheat genome with an average genome marker density of 15.5 cM per marker interval. The 5 chromosomes are only part of the wheat genome.

These chromosomes were scanned for QTL of the binary trait using both the mixture model and the heterogeneous variance model. The LOD profiles of the two

methods are shown in Fig. 1. Two QTL on chromosome 2 and one QTL on chromosome 5 have been detected with high LOD score. The chromosome-wise empirical threshold values are lower than LOD 3. With the chromosome-wise threshold values, we detected one more QTL on chromosome 1. The estimated QTL parameters are listed in Table 1. The two models (mixture model and heterogeneous variance model) produced very similar results.

Binomial traits

The same experiment conducted by Dou et al. (2009) also recorded the number of seeded spikelets and the total number spikelets for each plant. The ratio of the two records is a binomial trait. The same mapping population and the same linkage map were used also for the binomial trait QTL mapping. Again, both the mixture model and the heterogeneous variance model were used for the binomial trait analysis. Unfortunately, the mixture model failed to generate meaningful result. Therefore, the result of the mixture model analysis was not reported. In chromosome regions, where there were no QTL, the mixture model generated result similar to that of the heterogeneous variance model. For regions with large QTL, the mixture model approach failed to converge. The possible reason for the failure will be presented in Sect. "Discussion". We now focus on the result of the heterogeneous variance model. The LOD profile is shown in Fig. 2. First, the pattern of the LOD profile is similar to that of the binary trait analysis. There is strong evidence that there is one QTL on chromosome 1 and two QTL on chromosome 2. The LOD score here for the binomial trait is not in the same scale as that of the binary trait. The highest LOD is
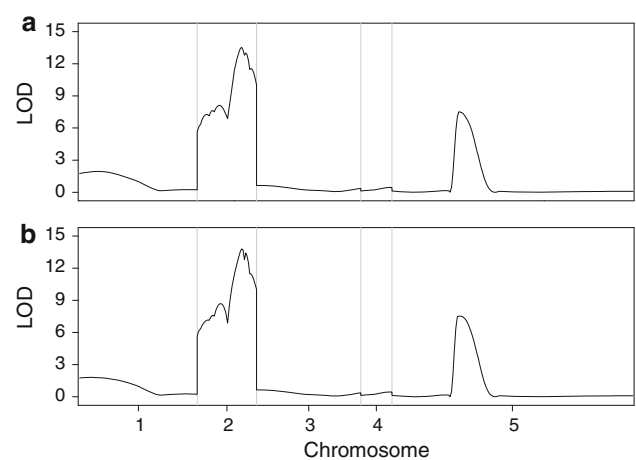

Fig. 1 The LOD profiles for the female sterility (binary) trait of wheat. **a** The *upper panel* shows the LOD profile for the heterogeneous variance model; **b** the *lower panel* shows the LOD profile for the mixture model. The chromosomes are separated by the *vertical gray reference lines*

Table 1 The estimated QTL parameters of the wheat female sterility (binary) trait

Chromosome	Position (cM)	Support interval	LOD score	Critical value	Effect
Heterogeneous					
1	11.88	0.00–38.13	1.94	1.43	0.4489 (3.1510E-2)
2	13.85	4.45–19.15	8.12	1.67	1.1145 (4.1370E-2)
2	28.71	25.60–32.61	13.54	1.67	1.3577 (4.5660E-2)
5	43.15	41.75–48.94	7.52	2.10	0.9414 (3.2080E-2)
Mixture					
1	8.91	0.00–40.12	1.80	1.41	0.4375 (2.5957E-2)
2	14.81	11.44–18.18	8.68	1.72	1.1339 (4.4504E-2)
2	28.71	26.51–30.01	13.80	1.72	1.3674 (4.6637E-2)
5	44.12	41.75–50.38	7.53	2.10	1.0129 (4.2112E-2)

The empirical chromosome-wise critical values are also given. The standard errors of estimated QTL effects are given in parentheses

about 1,000, almost a hundred times more than the binary trait. This inflation of the LOD reflects the increased power of the binomial data analysis than the binary data analysis. Regions of other chromosomes also show LOD score higher than the empirical chromosome-wise threshold values. These include chromosomes four and five. The estimated QTL parameters are listed in Table 2.

Poisson traits

This example demonstrates the application of the generalized linear model QTL mapping for traits with a Poisson distribution. The data were provided by Dr. Gangqiang Cao at Zengzhou University, China. The result has not been published in any form. The mapping population was a double haploid family of the rice initiated from the cross between IR64 and Azucena. The trait analyzed was the tiller number with an assumed Poisson distribution. The sample size was $n = 110$ and the number of markers was 175. These markers covered 12 chromosomes (2,031 cM in total length) of the rice genome with an average marker interval of 11.6 cM. This dataset was different from that used by Cui et al. (2006). Both experiments were initialized from the same line cross with the same linkage map, but the experiments were conducted in different times and different locations by different investigators. Interval mappings under both the mixture model and the heterogeneous variance model were applied to the data. The LOD score profiles obtained from the two different methods are depicted in Fig. 3. The LOD score from the heterogeneous variance model is slightly higher than that of the mixture model, but the difference is almost negligible. If LOD = 3 was used as the criterion for significance, two QTL would have been detected on chromosomes 1 and 4. We used the quick method of Piepho (2001) to calculate the empirical threshold for each chromosome and used these thresholds to declare significance
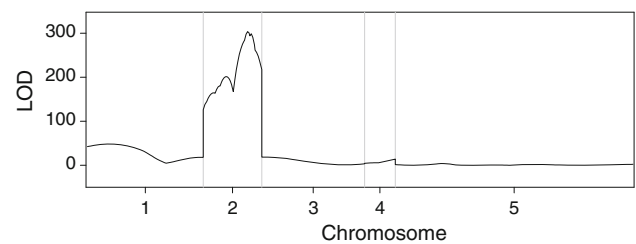

Fig. 2 The LOD score profile of the female sterility (binomial) trait in wheat under the heterogeneous variance model. The mixture model failed to converge to the correct solution and thus are not presented. The chromosomes are separated by the *vertical gray reference lines*

of QTL. The LOD thresholds are substantially less than LOD 3. Using the chromosome-specific thresholds, we detected one more QTL at the end of chromosome 12. The estimated QTL parameters are listed in Table 3. The supporting interval for each estimated QTL position was determined by the one-LOD drop approach (Ooijen 1992). The two methods differ slightly in the estimated QTL positions and the supporting intervals. The supporting intervals of QTL positions for the heterogeneous variance model were consistently shorter than those of the mixture model. Overall, the difference between the two methods is very small and can be safely ignored for this data analysis.

Simulation studies

Binomial traits

This simulation experiment was to demonstrate the difference between the mixture model and the heterogeneous variance model for binomial trait QTL mapping. For the female sterility trait of wheat, the mixture model approach failed to converge for the binomial data analysis but succeeded for the binary data. Binomial data were supposed to be more informative than the binary data, but

Table 2 Estimated QTL parameters for the wheat female sterility (binomial) trait from the heterogeneous variance model analysis

Chromosome	Position (cM)	Support interval	LOD score	Critical value	Effect
1	15.00	9.50–20.50	48.40	2.11	0.4803 (1.040E-03)
1	74.53	68.53–75.89	18.19	2.11	0.2401 (7.063E-04)
2	14.96	14.46–15.46	201.45	2.31	0.8922 (1.033E-03)
2	28.97	28.47–29.47	303.53	2.31	1.0484 (1.030E-03)
4	20.15	18.42–20.15	14.02	1.53	−0.2117 (6.890E-04)
5	30.31	27.50–34.81	4.03	2.27	−0.1203 (7.746E-04)

The empirical chromosome-wise critical values are also given. The standard errors of estimated QTL effects are provide in parentheses

more information turned out to do more harm than good to the mixture model, a typical problem of inconsistency has occurred. The problem happened in the calculation of the posterior probability of QTL genotype. The binomial density is equivalent to the product of multiple independent Bernoulli trials. This density is extremely sensitive to the change of genotype from one form to another, especially when the number of trials is large. The super sensitiveness of binomial density to the QTL genotype change led to degeneracy of the posterior probability of QTL genotype. We conducted a small-scale simulation experiment to demonstrate the problem of the mixture model. We simulated a QTL at position 55 cM of a chromosome with 100 cM in length. Five markers were placed evenly on the chromosome with 20 cM per marker interval. The sample size was 500 for an F_2 population. The binomial trait phenotype was simulated for each plant with a constant number of trials for all plants. We simulated the following number of trials in four experiments: 20, 40, 80 and 160. The LOD score profiles obtained from the mixture model and the heterogeneous variance model are shown in Fig. 4 for all the four experiments. We can see that when the number of trials were 20 and 40, the LOD profiles of the two methods are very similar. As the number of trials increased to 80 and 160, the differences between the two models are dramatic, with LOD profile of the mixture model drastically deviated from that of the heterogeneous variance model when the putative QTL position off the marker position. The strange leaf-like pattern of the mixture model LOD profile reflects the instability and failure of the mixture method.

Poisson traits

This simulation study was to evaluate the differences between the EM implemented mixture model and the heterogeneous residual variance model of interval mapping for Poisson traits. The simulation experiments were much more comprehensive than the previous one. The criteria of evaluation include the statistical powers, the test statistic profiles, the estimation errors of QTL parameters, the
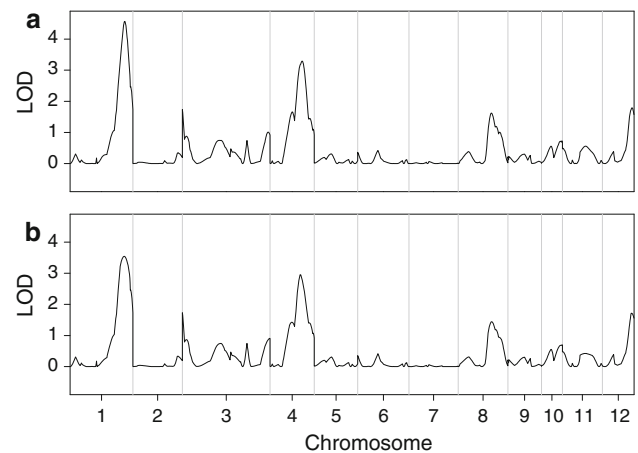

Fig. 3 LOD profiles for the tiller number of rice (Poisson trait) QTL mapping. **a** The *upper panel* shows the LOD profile for the heterogeneous variance model; **b** the *lower panel* shows LOD profile for the mixture model. The chromosomes are separated by the *vertical gray reference lines*

biases of the estimation and the computational times. The factors considered are the marker density (A), size of the QTL (B), mean of the trait (C), the sample size (D) and the QTL position (E).

A single chromosome with 100 cM in length was simulated for an F_2 population derived from the cross of two inbred lines. The chromosome was covered evenly by the following number of markers: 101, 51, 21, 11 and 6. These correspond to 1, 2, 5, 10 and 20 cM per marker interval. The size of the QTL (additive effect a) was investigated at the following five levels: 0.142, 0.324, 0.471, 0.707 and 0.926. These five levels of the additive effects correspond to the following levels of the heritability (h^2): 0.01, 0.05, 0.10, 0.20 and 0.30. For simplicity, dominance effect was not simulated and also not included in the model for data analysis. The mean of the Poisson trait was determined by the non-QTL effect β (intercept, a scalar because no other non-QTL effects were simulated). The five levels of the intercept (β) were −1.0, −0.5, 0.0, 0.5 and 1.0. The sample size was investigated at the following five levels: 100, 200, 300, 500 and 600. The simulated QTL was located at the

Table 3 Estimated QTL parameters using the mixture model and the heterogeneous variance model for the tiller number (Poisson) trait QTL mapping experiment

Chromosome	Position (cM)	Support interval	LOD score	Critical value	Effect
Heterogeneous					
1	196.41	183.09–209.72	4.569	1.975	0.2014 (1.848E-3)
4	115.16	96.82–130.45	3.290	1.878	0.1570 (1.553E-3)
12	106.62	89.88–113.58	1.791	1.663	0.1032 (1.271E-3)
Mixture					
1	194.44	173.48–216.30	3.543	1.933	0.1549 (1.419 E-3)
4	109.23	94.84–130.45	2.955	1.841	0.1350 (1.345E-3)
12	104.63	88.93–113.58	1.715203	1.656	0.0979 (1.207E-3)

The empirical chromosome-wise critical values are also given. The standard errors of estimated QTL effects are given in parentheses
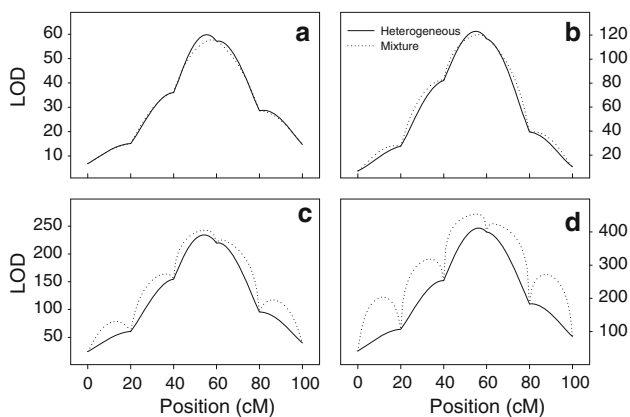

Fig. 4 Comparison of the LOD profiles for the mixture model (*dotted line*) and the heterogeneous variance model (*solid line*) for a binomial trait of a simulation experiment. The sample size was 500 and the chromosome size was 100 cM long covered by 6 evenly distributed markers (20 cM per marker interval). The binomial trait is measured as the ratio of the number of events to the number of trails. **a** The number of trials per plant was 20; **b** the number of trials per plant was 40; **c** the number of trials per plant was 80; **d** the number of trials per plant was 160

Table 4 Annotation of the simulation experiments

Label	Treatment	Level				
		1	2	3	4	5
A	Marker interval (cM)	1	2	5	10	20
B	Heritability (%)	1	5	10	20	30
C	Intercept	−1.0	−0.5	0	0.5	1.0
D	Sample size	100	200	300	500	600
E	QTL position	10	25	50	75	90

to $5 \times 4 + 1 = 21$. Each of the 21 treatment combinations was replicated 1,000 times to examine the empirical statistical powers and estimation errors of parameters for comparisons of the two models. The critical value for QTL detection was determined by simulating additional 1,000 samples under the null QTL model ($a = 0.0$) for each of the treatment combinations. The empirical power for each experiment was the proportion of the samples in which the QTL was detected out of the 1,000 replicates.

Before we discuss the biases and errors of the estimated QTL parameters, let us look at Fig. 5 which shows the average LOD test statistic profiles from the 1,000 replicates for all the experiments. The solid and dotted curves represent the LOD score profiles for the heterogeneous variance model and the mixture model, respectively. The straight horizontal lines are the critical values used to declare QTL significance for power studies. These critical values were drawn from 1,000 additional simulations under the null model ($a = 0$). Figure 5 provides a rich source of information regarding the behaviors of the two models. We only highlight a few important points here. First, at marker positions, the LOD scores for the heterogeneous variance model and the mixture model are identical, as expected, but off the markers the mixture model consistently produced higher LOD scores than the heterogeneous variance model. The higher LOD scores of the mixture model did not translate into higher powers because the critical values for the mixture model were also higher than the heterogeneous
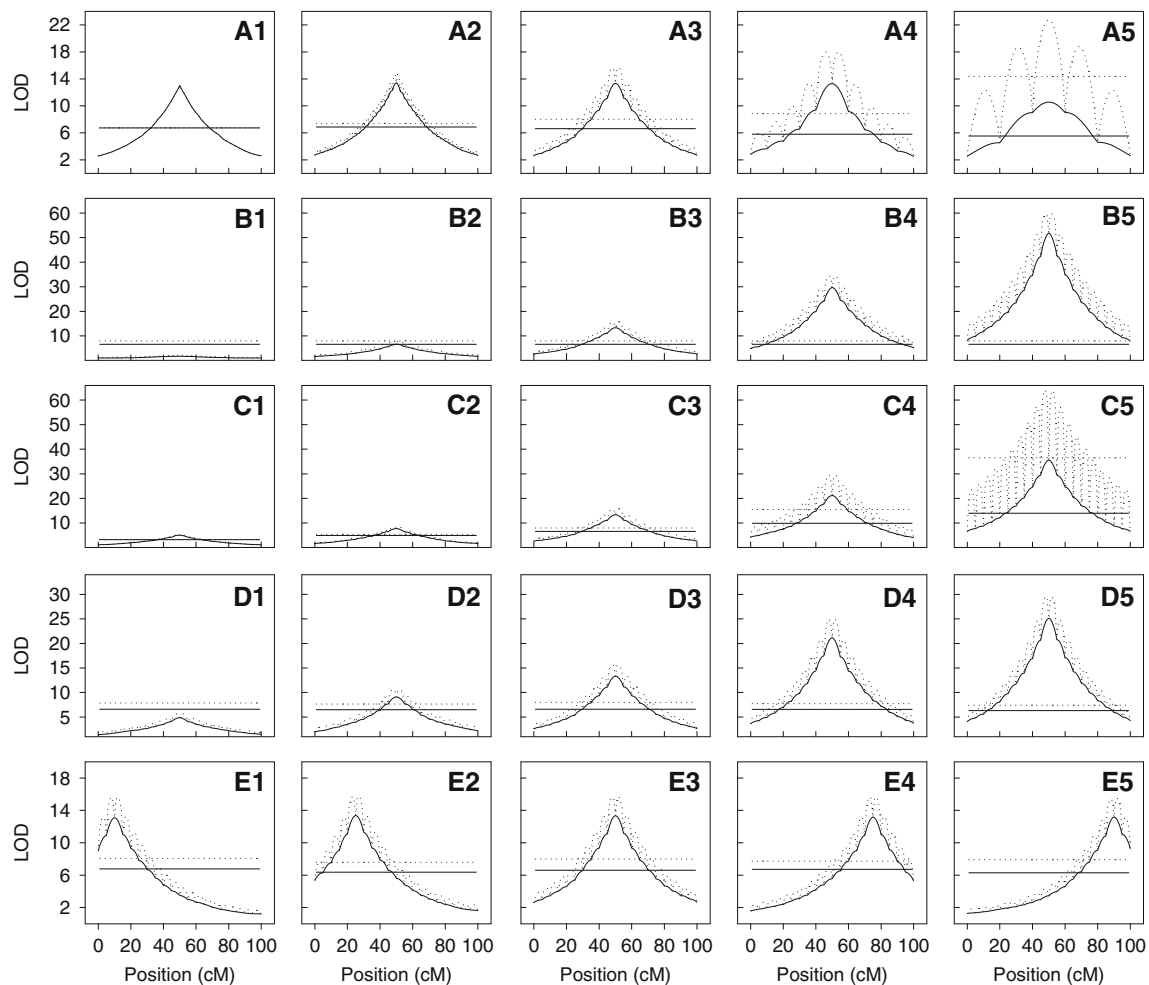
following different positions: 10, 25, 50, 75 and 90 cM. The true QTL parameter values and the experimental parameters for this comprehensive simulation experiment are summarized in Table 4.

The total number of combinations for all the five factors is $5^5 = 3{,}125$. If each combination was simulated 1,000 times, the work load would be very intensive. Therefore, we decided to evaluate a small subset of the treatment combinations to draw conclusions. We chose sample size $n = 300$, marker interval 5 cM, QTL size $a = 0.471$ ($h^2 = 0.10$), non-QTL effect $\beta = 0.0$ and QTL position at 50 cM as the basic experimental setup (central level for each of the five factors). Under the basic setup of the experiment, we then evaluated one factor at a time by expanding to all the five levels for the factor of interest with the remaining factors held at the basic levels. This reduced the total number of experiments from 3,125

Fig. 5 Average LOD score profiles of the simulation experiments of QTL mapping for a Poisson trait. The *solid curve* represents the average LOD score profile of the heterogeneous variance model while the *dotted line* represents the average LOD score profile of the mixture model. The *solid* and *dotted horizontal lines* denote the empirical critical values under a Type I error of 0.05 used to declare the significance of the QTL for the heterogeneous variance model and the mixture model, respectively. Labels *A*, *B*, *C*, *D* and *E* represent five different factors (treatments) that affect results of QTL mapping. These factors are marker density (**a**), QTL size measured in heritability (**b**), intercept (**c**), sample size (**d**) and QTL position (**e**). Numbers 1, 2, 3, 4 and 5 stand for five levels within each factor. Detailed information about the factors and levels within each factor can be found in Table 4

variance model. Second, the leaf-like patterns of the LOD score profiles of the mixture model became more severe when the marker interval was increased. This reflected an intrinsic flaw of the mixture model. In contrast, the heterogeneous variance model produced much more smoothed LOD profiles.

We now compare the empirical statistical powers for the two models (see the top panels in Fig. 6). The solid and open symbols represent the powers for the heterogeneous variance model and the mixture model, respectively. In all cases, the heterogeneous variance model had a slightly higher power than the mixture model, although the difference was almost negligible. Some factors had large effects on the powers and others had small effects on the powers. In general, the power declined as the marker interval increased. Sample size and the size of QTL are

most influencing factors on the powers. Intercept and position of the true QTL had little influence on the powers.

Let us turn into the panels in the second row of Fig. 6 to examine the factors that affect the estimated QTL positions. The solid and open symbols indicate the average estimated QTL positions for the two models obtained from the 1,000 replications. The vertical bars represent the standard deviations of the estimated QTL position calculated from the 1,000 replicated experiments. The red horizontal lines indicate the true positions of the simulated QTL. Both models are unbiased and the standard deviations of the estimated positions are very similar for the two models. The size of the QTL and the sample size had the largest influence on the accuracy of the estimated QTL position.

Panels of the third row of Fig. 6 show the average estimated intercepts and the standard deviations obtained

Fig. 6 Estimated QTL parameters and empirical powers of the simulation experiments of QTL mapping for a Poisson trait. Labels *A*, *B*, *C*, *D* and *E* represent five treatments (factors). Within each treatment, there are five different levels represented by levels 1 (*up triangle*), 2 (*upside down triangle*), 3 (*circle*), 4 (*square*) and 5 (*diamond*). The *solid (filled) symbol* represents result of the heterogeneous variance model while the *open symbol* represents the result of the mixture model. The *vertical bars* represent the standard deviations calculated from the 1,000 replicated experiments. The *horizontal lines* represent the true values of the QTL parameters. See Table 4 for the detailed information about the factors and levels within factors
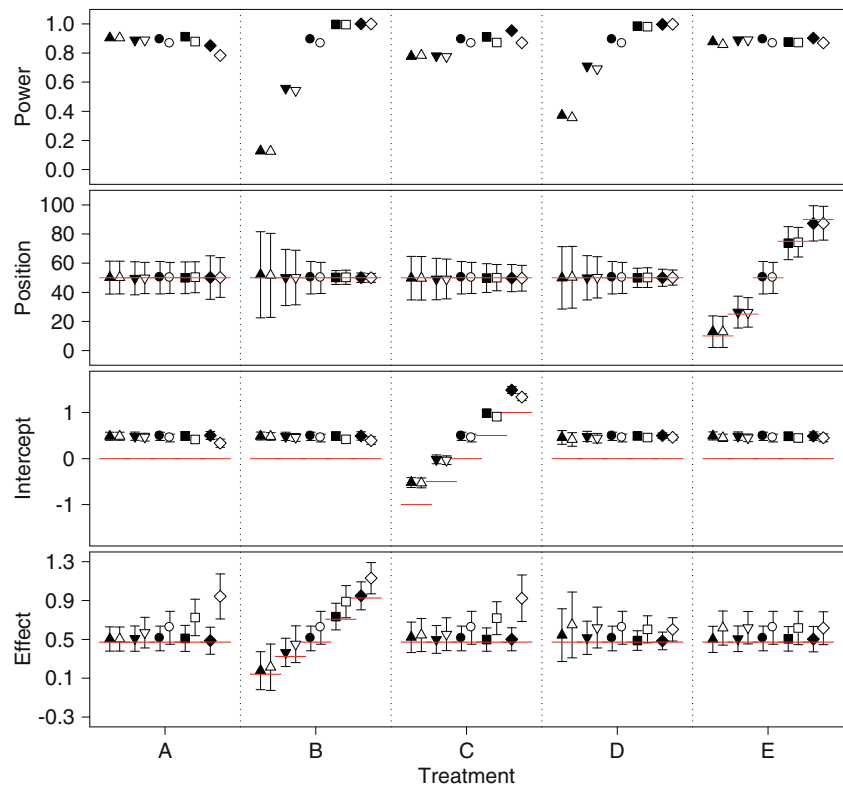
from the 1,000 replicated simulations. The two models had very similar estimated intercepts, but both models were biased upwardly.

The panels at the bottom of Fig. 6 show the results of the estimated QTL effect (*a*) for the two models. The heterogeneous model was unbiased but the mixture model was consistently biased upwardly. The bias for the mixture model was more serious as the marker interval increased.

Overall, the heterogeneous variance model performed consistently better than the mixture model. This observation was not expected. Our original purpose of proposing the heterogeneous variance model was to improve the computational efficiency. We did not anticipate that the mixture model was not consistent and performed poorly. The simulation experiments did show that the heterogeneous variance model took, on average, about one-third of the computing time of the mixture model. To our surprise, the heterogeneous variance model outperformed the mixture model in all cases considered in the experiments.

Discussion

The most commonly used algorithm for missing values in generalized linear model is the EM algorithm (Horton and Laird 1998). We extended the EM algorithm to interval mapping, where the independent variable (genotype indicator variable) is missing for all individuals. We also

developed a heterogeneous variance model to approximate the mixture model. Both the binary data and the Poisson data analyses showed that the two methods generated similar results, but the heterogeneous variance model is computationally faster than the mixture model-based EM algorithm. To our surprise, the mixture model approach is not always better than the heterogeneous variance model in terms of better estimation of QTL parameters. The binomial data analysis showed that the mixture model approach failed to converge to the correct values for some marker intervals while the heterogeneous variance model worked very well. The failure is due to several possible reasons: (1) low information content for large marker intervals, (2) large QTL effects and (3) inconsistency of the EM algorithm. For the first reason, a large marker interval means that QTL position in the middle of the interval has little information regarding the genotype of the putative QTL. The posterior probability of QTL genotype is largely determined by the data and the current parameter values. In the end, the posterior probabilities become degenerated (probability equals unity for one genotype and zero for all other genotypes) for all individuals. This phenomenon was observed only for large intervals. The second reason of the failure is due to large QTL effects. We noticed that the failure of the mixture model approach did not happen in intervals that do not have QTL, even though some of those intervals are very large. It is the combination of large interval with large QTL effects that caused the failure. The

third reason of the failure is due the inconsistency (intrinsic flaw) of the mixture model. More extensive simulation experiments using the Poisson data further justified the heterogeneous variance model for interval mapping.

Most QTL mapping experiments in the future will be done with a high marker density. In that case, the mixture model and heterogeneous variance model will have negligible difference in efficiency, with the latter slightly more preferable than the former due to its light computational load. Theory and methods of interval mapping under the mixture model are well developed both for normally distributed traits and for discrete traits. Variance–covariance matrix for the estimated parameters is also available, but only for normally distributed trait interval mapping (Kao and Zeng 1997). This research is the first attempt to develop the covariance matrix of estimated QTL parameters under the generalized linear mixture model. With the availability of the variance–covariance matrix of estimated QTL effects, we have a choice to perform either the Wald test or the likelihood ratio test. The two test statistics are asymptotically the same, with the latter slightly more preferable when the sample size is small (Frank 2001).

The approximate heterogeneous variance model was originally developed by one of us for normal trait QTL mapping (Xu 1998). We demonstrated here that it worked equally well for the generalized linear model. Simulation experiments and real data analysis all demonstrated that the approximation is very close to, and sometimes better than, the mixture model. The advantage of the approximate method is the avoidance of mixture model and thus avoidance of the EM algorithm. As a result, computing the variance–covariance matrix of the estimated parameters becomes straightforward, a by-product of the iteration process. In addition, the heterogeneous variance model appears to be more robust and stable compared with the mixture model-based EM algorithm according to our binomial data analysis and the simulation study.

Both the mixture model maximum likelihood and the heterogeneous variance approximation have been coded in an existing QTL mapping program called PROC QTL (Hu and Xu 2009). This program is a user-defined SAS procedure. Users need to specify the distribution of the data. The default distribution is normal. The current version of PROC QTL can handle binary, binomial, ordinal and Poisson distributions. More distributions will be added in the future, depending on the availability of the data. The program is downloadable from our Web site (http://www.statgen.ucr.edu).

Appendix A: partial derivatives

(1) Ordinal data with observed genotypes

Let $g = 1, 2, 3$ index the three genotypes for an F_2 population derived from the cross of two inbred lines. The expectation of Y_{jk} conditional on the parameters for genotype g is

$$\mu_{jk}(g) = E(Y_{jk}|g) = \Phi(\alpha_k + X_j\beta + H_g\gamma) - \Phi(\alpha_{k-1} + X_j\beta + H_g\gamma) \tag{42}$$

Define $\mu_j(g) = \left[\mu_{j1}(g)\ \mu_{j2}(g), \ldots, \mu_{j(p+1)}(g)\right]^T$ as a $(p+1) \times 1$ vector for the expectation of Y_j. The D matrix for genotype g is

$$D_j(g) = \left[\frac{\partial \mu_j(g)}{\partial \alpha^T}\quad \frac{\partial \mu_j(g)}{\partial \beta^T}\quad \frac{\partial \mu_j(g)}{\partial \gamma^T}\right] \tag{43}$$

where

$$\frac{\partial \mu_{jk}(g)}{\partial \alpha_{k-1}} = -\phi(\alpha_{k-1} + X_j\beta + H_g\gamma)$$
$$\frac{\partial \mu_{jk}(g)}{\partial \alpha_k} = \phi(\alpha_{k-1} + X_j\beta + H_g\gamma) \tag{44}$$
$$\frac{\partial \mu_{jk}(g)}{\partial \alpha_l} = 0, \forall\, l \neq k-1, k$$
$$\frac{\partial \mu_{jk}(g)}{\partial \beta} = X_j^T\left[\phi(\alpha_k + X_j\beta + H_g\gamma) - \phi(\alpha_{k-1} + X_j\beta + H_g\gamma)\right] \tag{45}$$

and

$$\frac{\partial \mu_{jk}(g)}{\partial \gamma} = H_g^T\left[\phi(\alpha_k + X_j\beta + H_g\gamma) - \phi(\alpha_{k-1} + X_j\beta + H_g\gamma)\right] \tag{46}$$

(2) Ordinal data under the heterogeneous variance model (approximation)

The expectation of Y_{jk} conditional on parameters is

$$\mu_{jk} = E(Y_{jk}) \approx \Phi\left[\frac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma)\right] - \Phi\left[\frac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma)\right] \tag{47}$$

Let $\mu_j = \left[\mu_{j1}\ \mu_{j2}, \ldots, \mu_{j(p+1)}\right]^T$ be a $(p+1) \times 1$ vector for the expectation of Y_j. The D matrix is defined as

$$D_j = \begin{bmatrix} \frac{\partial \mu_j}{\partial \alpha^T} & \frac{\partial \mu_j}{\partial \beta^T} & \frac{\partial \mu_j}{\partial \gamma^T} \end{bmatrix} \quad (48)$$

where

$$\frac{\partial \mu_{jk}}{\partial \alpha_{k-1}} = -\frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma) \right]$$

$$\frac{\partial \mu_{jk}}{\partial \alpha_k} = \frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma) \right] \quad (49)$$

$$\frac{\partial \mu_{jk}}{\partial \alpha_l} = 0, \forall\, l \neq \{k-1, k\}$$

$$\frac{\partial \mu_{jk}}{\partial \beta} = \frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma) \right] X_j^T$$

$$\qquad - \frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma) \right] X_j^T \quad (50)$$

and

$$\frac{\partial \mu_{jk}}{\partial \gamma} = \frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(\alpha_k + X_j\beta + U_j\gamma) \right]$$

$$\qquad \times \left[U_j^T - \frac{1}{\sigma_j^2}(\alpha_k + X_j\beta + U_j\gamma)\Sigma_j\gamma \right]$$

$$\qquad - \frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma) \right]$$

$$\qquad \times \left[U_j^T - \frac{1}{\sigma_j}(\alpha_{k-1} + X_j\beta + U_j\gamma)\Sigma_j\gamma \right] \quad (51)$$

(3) Binary data with observed genotypes

Let $g = 1, 2, 3$ index the three genotypes and define the genotype-specific expectation of Y_j by

$$\mu_j(g) = E(Y_j|g) = \Phi(X_j\beta + H_g\gamma) \quad (52)$$

The D matrix for genotype g is

$$D_j(g) = \begin{bmatrix} \frac{\partial \mu_j(g)}{\partial \beta^T} & \frac{\partial \mu_j(g)}{\partial \gamma^T} \end{bmatrix} \quad (53)$$

where

$$\frac{\partial \mu_j(g)}{\partial \beta} = X_j^T \phi(X_j\beta + H_g\gamma) \quad (54)$$

and

$$\frac{\partial \mu_j(g)}{\partial \gamma} = H_g^T \phi(X_j\beta + H_g\gamma) \quad (55)$$

(4) Binary data under the heterogeneous variance model

Define the expectation of Y_j by

$$\mu_j = E(Y_j) \approx \Phi \left[\frac{1}{\sigma_j}(X_j\beta + U_j\gamma) \right] \quad (56)$$

The D matrix is

$$D_j = \begin{bmatrix} \frac{\partial \mu_j}{\partial \beta^T} & \frac{\partial \mu_j}{\partial \gamma^T} \end{bmatrix} \quad (57)$$

where

$$\frac{\partial \mu_j}{\partial \beta} = \frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(X_j\beta + U_j\gamma) \right] X_j^T \quad (58)$$

and

$$\frac{\partial \mu_j}{\partial \gamma} = \frac{1}{\sigma_j} \phi \left[\frac{1}{\sigma_j}(X_j\beta + U_j\gamma) \right] \left[U_j^T - \frac{1}{\sigma_j^2}(X_j\beta + U_j\gamma)\Sigma_j\gamma \right] \quad (59)$$

(5) Poisson data with observed genotypes

The expectation of Y_j given genotype g is

$$\mu_j(g) = E(Y_j|g) = \exp(X_j\beta + H_g\gamma), \forall\, g = 1, 2, 3 \quad (60)$$

The D matrix is

$$D_j(g) = \begin{bmatrix} \frac{\partial \mu_j(g)}{\partial \beta^T} & \frac{\partial \mu_j(g)}{\partial \gamma^T} \end{bmatrix} \quad (61)$$

where

$$\frac{\partial \mu_j(g)}{\partial \beta} = \exp(X_j\beta + H_g\gamma)X_j^T$$

$$\frac{\partial \mu_j(g)}{\partial \gamma} = \exp(X_j\beta + H_g\gamma)H_g^T \quad (62)$$

(6) Poisson data under the overdispersion model (approximation)

Denote the expectation of Y_j by

$$\mu_j = E(Y_j) \approx \exp \left[\frac{1}{\sigma_j}(X_j\beta + U_j\gamma) \right] \quad (63)$$

The D matrix is

$$D_j = \begin{bmatrix} \frac{\partial \mu_j}{\partial \beta^T} & \frac{\partial \mu_j}{\partial \gamma^T} \end{bmatrix} \quad (64)$$

where

$$\frac{\partial \mu_j}{\partial \beta} = \frac{\mu_j}{\sigma_j} X_j^T$$

$$\frac{\partial \mu_j}{\partial \gamma} = \frac{\mu_j}{\sigma_j} U_j^T - \frac{\mu_j \ln(\mu_j)}{\sigma_j^2}\Sigma_j\gamma \quad (65)$$

Appendix B: derivation of the EM algorithm

Since the ordinal phenotype has been presented as multi-variable Bernoulli variable and the parameter vector has been partitioned into three blocks, it is a little bit tedious to derive the EM algorithm for parameter estimation. Therefore, we only used the Poisson data as an example for the derivation. The derivation applies to generalized linear model for any distributions. Let $G_j = 1, 2, 3$ be a discrete variable for the genotype of individual j. Let $\delta_j = \begin{bmatrix} \delta(G_j, 1) & \delta(G_j, 2) & \delta(G_j, 3) \end{bmatrix}^T$ be a vector of three indicator variables of the genotypes for individual j following a

multinomial distribution with sample size one. The relationship between G_j and δ_j is

$$\delta(G_j, g) = \begin{cases} 1 & G_j = g \\ 0 & G_j \neq g \end{cases} \tag{66}$$

for $g = 1, 2, 3$. When the genotype of individual j is known, the complete-data log likelihood function for the parameters is

$$L(\theta, \delta) = \sum_{j=1}^{n} \sum_{g=1}^{3} \delta(G_j, g) \ln p(Y_j|g) \tag{67}$$

where

$$p(Y_j|g) = \frac{\left[\exp(X_j\beta + H_g\gamma)\right]^{Y_j}}{(Y_j)!} \exp\left[-\exp(X_j\beta + H_g\gamma)\right] \tag{68}$$

is the Poisson probability. The expectation of the complete-data log likelihood function is

$$E[L(\theta, \delta)] = \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) \left[Y_j(X_j\beta + H_g\gamma) - \exp(X_j\beta + H_g\gamma)\right] \tag{69}$$

where

$$p_j^*(g) = E\left[\delta(G_j, g)|\theta^{(t)}, Y_j\right] = \frac{p_j(g)p(Y_j|g)}{\sum_{g'}^{3} p_j(g')p(Y_j|g')} \tag{70}$$

is the posterior probability of $\delta(G_j, g) = 1$ given $\theta = \theta^{(t)}$. Now $p_j^*(g)$ is treated as a constant (not a function of the parameters because the unknown parameter involved in $p_j^*(g)$ has been replaced by the parameter value at iteration t). The EM algorithm actually maximizes the expectation of the complete-data log likelihood function, not the original observed log likelihood function. The maximum likelihood estimation in the neighborhood of $\theta = \theta^{(t)}$ can be obtained through Taylor expansion around $\theta = \theta^{(t)}$ (the Newton–Raphson method),

$$\theta^{(t+1)} = \theta^{(t)} + \Delta\theta$$
$$= \theta^{(t)} + \left[-\frac{\partial^2 E[L(\theta, \delta)]}{\partial\theta\partial\theta^T}\right]^{-1} \left[\frac{\partial E[L(\theta, \delta)]}{\partial\theta}\right] \tag{71}$$

Now we only need to prove that

$$\frac{\partial E[L(\theta, \delta)]}{\partial\theta} = \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) D_j^T(g) W_j(g) \left[Y_j - \mu_j(g)\right] \tag{72}$$

and

$$-\frac{\partial^2 E[L(\theta, \delta)]}{\partial\theta\partial\theta^T} = \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) D_j^T(g) W_j(g) D_j(g) \tag{73}$$

The partial derivative of the complete-data log likelihood function with respect to the unknown parameters is

$$\frac{\partial E[L(\theta, \delta)]}{\partial\theta} = \begin{bmatrix} \frac{\partial E[L(\theta, \delta)]}{\partial\beta} \\ \frac{\partial E[L(\theta, \delta)]}{\partial\gamma} \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) X_j^T \left[Y_j - \exp(X_j\beta + H_g\gamma)\right] \\ \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) H_g^T \left[Y_j - \exp(X_j\beta + H_g\gamma)\right] \end{bmatrix} \tag{74}$$

The second partial derivative is

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{\partial^2 E[L(\theta, \delta)]}{\partial\beta\partial\beta^T} & \frac{\partial^2 E[L(\theta, \delta)]}{\partial\beta\partial\gamma^T} \\ \frac{\partial^2 E[L(\theta, \delta)]}{\partial\gamma\partial\beta^T} & \frac{\partial^2 E[L(\theta, \delta)]}{\partial\gamma\partial\gamma^T} \end{bmatrix} \tag{75}$$

where

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial\beta\partial\beta^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) X_j^T \exp(X_j\beta + H_g\gamma) X_j$$

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial\beta\partial\gamma^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) X_j^T \exp(X_j\beta + H_g\gamma) H_g$$

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial\gamma\partial\beta^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) H_g^T \exp(X_j\beta + H_g\gamma) X_j$$

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial\gamma\partial\gamma^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) H_g^T \exp(X_j\beta + H_g\gamma) H_g \tag{76}$$

Given that

$$D_j^T(g) = \begin{bmatrix} \frac{\partial\mu_j(g)}{\partial\beta} \\ \frac{\partial\mu_j(g)}{\partial\gamma} \end{bmatrix} = \begin{bmatrix} \exp(X_j\beta + H_g\gamma) X_j^T \\ \exp(X_j\beta + H_g\gamma) H_g^T \end{bmatrix} \tag{77}$$

(see Eqs. 61 and 62 in Appendix A) and

$$W_j(g) = \frac{1}{\exp(X_j\beta + H_g\gamma)} = \frac{1}{\mu_j(g)} \tag{78}$$

(see Eq. 39 in the main text), we have

$$\begin{bmatrix} X_j^T \\ H_g^T \end{bmatrix} = D_j(g) W_j(g) \tag{79}$$

Substituting X_j and H_g in Eqs. (74) and (76) by Eq. (79), we have

$$\frac{\partial E[L(\theta, \delta)]}{\partial\theta} = \begin{bmatrix} \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) \frac{\partial\mu_j(g)}{\partial\beta} W_j(g) \left[Y_j - \exp(X_j\beta + H_g\gamma)\right] \\ \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) \frac{\partial\mu_j(g)}{\partial\gamma} W_j(g) \left[Y_j - \exp(X_j\beta + H_g\gamma)\right] \end{bmatrix} \tag{80}$$

and

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial \beta \partial \beta^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) \frac{\partial \mu_j(g)}{\partial \beta} W_j(g) \frac{\partial \mu_j(g)}{\partial \beta^T}$$

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial \beta \partial \gamma^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) \frac{\partial \mu_j(g)}{\partial \beta} W_j(g) \frac{\partial \mu_j(g)}{\partial \gamma^T}$$

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial \gamma \partial \beta^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) \frac{\partial \mu_j(g)}{\partial \gamma} W_j(g) \frac{\partial \mu_j(g)}{\partial \beta^T}$$ (81)

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial \gamma \partial \gamma^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) \frac{\partial \mu_j(g)}{\partial \gamma} W_j(g) \frac{\partial \mu_j(g)}{\partial \gamma^T}$$

Therefore,

$$\frac{\partial E[L(\theta, \delta)]}{\partial \theta} = \sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) D_j^T(g) W_j(g) \left[Y_j - \mu_j(g) \right]$$ (82)

and

$$\frac{\partial^2 E[L(\theta, \delta)]}{\partial \theta \partial \theta^T} = -\sum_{j=1}^{n} \sum_{g=1}^{3} p_j^*(g) D_j^T(g) W_j(g) D_j(g)$$ (83)

This concludes the derivation of the EM algorithm.

Appendix C: generalized inverse of variance matrix

For ordinal traits, the observed data point Y_j for individual j is multinomial with sample size one. Therefore, the variance–covariance matrix is

$$V_j = \mathrm{var}(Y_j) = \mathrm{diag}(\mu_j) - \mu_j \mu_j^T = \psi_j - \mu_j \mu_j^T$$ (84)

where $\psi_j = \mathrm{diag}(\mu_j)$. This variance matrix can be rewritten in a general form as

$$V_j = \psi_j + c \mu_j \mu_j^T$$ (85)

where $c = -1$. The inverse of this general form has an explicit expression (Giri 1996),

$$V_j^{-1} = \psi_j^{-1} - \frac{1}{\mu_j^T \psi_j^{-1} \mu_j + c} \psi_j^{-1} \mu_j \mu_j^T \psi_j^{-1}$$ (86)

The fact that $\psi_j = \mathrm{diag}(\mu_j)$ leads to

$$\mu_j^T \psi_j^{-1} \mu_j = \sum_{k=1}^{p+1} \mu_{jk} = 1$$ (87)

Therefore, the inverse matrix can be written as

$$V_j^{-1} = \psi_j^{-1} - \frac{1}{1+c} \psi_j^{-1} \mu_j \mu_j^T \psi_j^{-1}$$ (88)

Since $1/(1 + c)$ is not defined when $c = -1$, the inverse matrix V_j^{-1} does not exist.

We now prove that ψ_j^{-1} is a generalized inverse of matrix V_j. To prove this, we only need to show that $V_j \psi_j^{-1} V_j = V_j$ because a generalize inverse V_j^- is defined as a matrix such that $V_j V_j^- V_j = V_j$. Substituting V_j^- by ψ_j^{-1}, we get

$$
\begin{aligned}
V_j \psi_j^{-1} V_j &= (\psi_j - \mu_j \mu_j^T) \psi_j^{-1} (\psi_j - \mu_j \mu_j^T) \\
&= \psi_j - \mu_j \mu_j^T - \mu_j \mu_j^T + \mu_j \mu_j^T \psi_j^{-1} \mu_j \mu_j^T \\
&= \psi_j - \mu_j \mu_j^T \\
&= V_j
\end{aligned}
$$ (89)

Keep in mind that the above derivation requires $\psi_j^{-1} \psi_j = I$ and $\mu_j^T \psi_j^{-1} \mu_j = 1$ for simplification. This concludes the proof that ψ_j^{-1} is a generalized inverse of matrix V_j.

In fact, ψ_j^{-1} is just one of an infinite number of generalized inverses of matrix V_j. A general form of the generalized inverse can be expressed by

$$V_j^- = \psi_j^{-1} - d \psi_j^{-1} \mu_j \mu_j^T \psi_j^{-1}$$ (90)

where d is a real number (a scalar not a matrix). The generalized inverse ψ_j^{-1} is simply obtained by setting $d = 0$. The following equation serves as a proof of this general form of the generalized inverse,

$$
\begin{aligned}
&(\psi_j - \mu_j \mu_j^T)(\psi_j^{-1} - d \psi_j^{-1} \mu_j \mu_j^T \psi_j^{-1})(\psi_j - \mu_j \mu_j^T) \\
&= (\psi_j - \mu_j \mu_j^T) \left[\psi_j^{-1}(\psi_j - \mu_j \mu_j^T) - d \psi_j^{-1} \mu_j \mu_j^T \psi_j^{-1}(\psi_j - \mu_j \mu_j^T) \right] \\
&= (\psi_j - \mu_j \mu_j^T) \left(I - \psi_j^{-1} \mu_j \mu_j^T - d \psi_j^{-1} \mu_j \mu_j^T + d \psi_j^{-1} \mu_j \mu_j^T \psi_j^{-1} \mu_j \mu_j^T \right) \\
&= (\psi_j - \mu_j \mu_j^T) \left(I - \psi_j^{-1} \mu_j \mu_j^T - d \psi_j^{-1} \mu_j \mu_j^T + d \psi_j^{-1} \mu_j \mu_j^T \right) \\
&= (\psi_j - \mu_j \mu_j^T) \left(I - \psi_j^{-1} \mu_j \mu_j^T \right) \\
&= \psi_j - \mu_j \mu_j^T - \psi_j \psi_j^{-1} \mu_j \mu_j^T + \mu_j \mu_j^T \psi_j^{-1} \mu_j \mu_j^T \\
&= \psi_j - \mu_j \mu_j^T - \mu_j \mu_j^T + \mu_j \mu_j^T \\
&= \psi_j - \mu_j \mu_j^T
\end{aligned}
$$ (91)

Appendix D: logistic regression

We now use binary trait QTL mapping as an example to demonstrate the logit link function of the generalized linear model. Recall that the univariate definition of the binary phenotype is

$$Y_j = \begin{cases} 1 & \text{for presence of the trait} \\ 0 & \text{for absence of the trait} \end{cases}$$ (92)

Under the logit link function, the expectation and variance of the phenotype given the parameter values are

$$E(Y_j) = \mu_j = \frac{\exp(X_j \beta + Z_j \gamma)}{1 + \exp(X_j \beta + Z_j \gamma)}$$ (93)

and

$$\mathrm{var}(Y_j) = V_j = \mu_j(1 - \mu_j)$$ (94)

respectively. The probability density is

$$p(Y_j) = \mu_j^{Y_j}(1-\mu_j)^{1-Y_j} \tag{95}$$

The link function is logit because

$$\eta_j = \text{logit}(\mu_j) = \ln\frac{\mu_j}{1-\mu_j} = X_j\beta + Z_j\gamma \tag{96}$$

Genotype observed

Let $g = 1, 2, 3$ index the three genotypes and

$$\mu_j(g) = E(Y_j|g) = \frac{\exp(X_j\beta + H_g\gamma)}{1 + \exp(X_j\beta + H_g\gamma)} \tag{97}$$

be the genotype specific expectation. Let

$$\eta_j(g) = X_j\beta + H_g\gamma \tag{98}$$

and

$$\xi(\eta_j) = \frac{\partial\mu_j(g)}{\partial\eta_j(g)} = \frac{\exp[\eta_j(g)]}{\{1 + \exp[\eta_j(g)]\}^2} \tag{99}$$

The D matrix for genotype g is defined as

$$D_j(g) = \begin{bmatrix} \frac{\partial\mu_j(g)}{\partial\beta^T} & \frac{\partial\mu_j(g)}{\partial\gamma^T} \end{bmatrix} \tag{100}$$

where

$$\frac{\partial\mu_j(g)}{\partial\beta} = \xi(\eta_j)X_j^T \tag{101}$$

and

$$\frac{\partial\mu_j(g)}{\partial\gamma} = \xi(\eta_j)H_g^T \tag{102}$$

Heterogeneous variance model

Let us define

$$\eta_j = \frac{1}{\sigma_j}(X_j\beta + U_j\gamma) \tag{103}$$

The expectation of Y_j is

$$\mu_j = E(Y_j) \approx \frac{\exp(\eta_j)}{1 + \exp(\eta_j)} \tag{104}$$

Define

$$\xi(\eta_j) = \frac{\partial\mu_j}{\partial\eta_j} = \frac{\exp(\eta_j)}{[1 + \exp(\eta_j)]^2} \tag{105}$$

The D matrix is

$$D_j = \begin{bmatrix} \frac{\partial\mu_j}{\partial\beta^T} & \frac{\partial\mu_j}{\partial\gamma^T} \end{bmatrix} \tag{106}$$

where

$$\frac{\partial\mu_j}{\partial\beta} = \frac{1}{\sigma_j}\xi(\eta_j)X_j^T \tag{107}$$

and

$$\frac{\partial\mu_j}{\partial\gamma} = \frac{1}{\sigma_j}\xi(\eta_j)\left[U_j^T - \frac{1}{\sigma_j^2}(X_j\beta + U_j\gamma)\Sigma_j\gamma\right] \tag{108}$$

References

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc B 26:211–252

Cui Y, Yang W (2009) Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros. J Theor Biol 256:276–285

Cui Y, Kim DY, Zhu J (2006) On the generalized poisson regression mixture model for mapping quantitative trait loci with count data. Genetics 174:2159–2172

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

Deng W, Chen H, Li Z (2006) A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. Genetics 172:1349–1358

Dou B, Hou B, Xu H, Lou X, Chi X, Yang J, Wang F, Ni Z, Sun Q (2009) Efficient mapping of a female sterile gene in wheat (*Triticum aestivum* L.). Genet Res 91:337–343

Frank EH (2001) Regression modeling strategies. Springer, New York

Giri NC (1996) Multivariate statistical analysis. Marcel Dekker, Inc, New York

Hackett CA, Weller JI (1995) Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics 51:1252–1263

Han L, Xu S (2008) A Fisher scoring algorithm for the weighted regression method of QTL mapping. Heredity 101:453–464

Horton NJ, Laird NM (1998) Maximum likelihood analysis of generalized linear models with missing covariates. Stat Methods Med Res 8:37–50

Hu Z, Xu S (2009) PROC QTL—A SAS procedure for mapping quantitative trait loci. Int J Plant Genomics 2009:3. doi:10.1155/2009/141234

Ibrahim JG, Chen M-H, Lipsitz SR (2002) Bayesian methods for generalized linear models with missing covariates. Can J Stat 30:55–78

Ibrahim JG, Chen M-H, Lipsitz SR, Herring AH (2005) Missing-data methods for generalized linear models: a comparative review. J Am Stat Assoc 100:332–346

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

Kao CH, Zeng ZB (1997) General formulas for obtaining the MLEs and the asymptotic variance–covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics 53:653–665

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

Lange C, Whittaker JC (2001) Mapping quantitative trait loci using generalized estimating equations. Genetics 159:1325–1337

Louis T (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc B 44:226–233

McCullagh P, Nelder JA (1989) Generalized linear models. Chapman & Hall, New York

Ooijen JW (1992) Accuracy of mapping quantitative trait loci in autogamous species. Theor Appl Genet 84:803–811

Piepho HP (2001) A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics 157:425–432

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

Robins JM, Ritov Y (2001) On double robustness. Stat Sin 11:920–936

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

SAS Institute (1999) SAS/STAT users' guide, vol 8. SAS Publishing, Cary

Searle SR (1997) Linear models. Wiley, New York

Wald A (1949) Note on the consistency of the maximum likelihood estimate. Ann Math Stat 20:595–601

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

Xu S (1998) Iteratively reweighted least squares mapping of quantitative trait loci. Behav Genetics 28:341–355

Xu S, Atchley WR (1996) Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics 143:1417–1424

Xu S, Yi N, Burke D, Galecki A, Miller RA (2003) An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. Genet Res 82:127–138

Xu C, Zhang Y-M, Xu S (2005) An EM algorithm for mapping quantitative resistance loci. Heredity 94:119–128

Yi N, Xu S (1999a) Mapping quantitative trait loci for complex binary traits in outbred populations. Heredity 82:668–676

Yi N, Xu S (1999b) A random model approach to mapping quantitative trait loci for complex binary traits in outbred populations. Genetics 153:1029–1040

Yi N, Xu S (2000) Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155:1391–1403