Journal of Theoretical Biology 256 (2009) 276–285

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros

Yuehua Cui *, Wenzhao Yang

Department of Statistics and Probability, Michigan State University, A-432 Wells Hall, East Lansing, MI 48824, USA

ARTICLE INFO

Article history:
Received 1 July 2008
Received in revised form
18 September 2008
Accepted 1 October 2008
Available online 15 October 2008

Keywords:
EM algorithm
Quantitative trait loci
Zero-inflated count data
Zero-inflated generalized Poisson regression model

ABSTRACT

Phenotypes measured in counts are commonly observed in nature. Statistical methods for mapping quantitative trait loci (QTL) underlying count traits are documented in the literature. The majority of them assume that the count phenotype follows a Poisson distribution with appropriate techniques being applied to handle data dispersion. When a count trait has a genetic basis, "naturally occurring" zero status also reflects the underlying gene effects. Simply ignoring or miss-handling the zero data may lead to wrong QTL inference. In this article, we propose an interval mapping approach for mapping QTL underlying count phenotypes containing many zeros. The effects of QTLs on the zero-inflated count trait are modelled through the zero-inflated generalized Poisson regression mixture model, which can handle the zero inflation and Poisson dispersion in the same distribution. We implement the approach using the EM algorithm with the Newton–Raphson algorithm embedded in the M-step, and provide a genome-wide scan for testing and estimating the QTL effects. The performance of the proposed method is evaluated through extensive simulation studies. Extensions to composite and multiple interval mapping are discussed. The utility of the developed approach is illustrated through a mouse F_2 intercross data set. Significant QTLs are detected to control mouse cholesterol gallstone formation.

Published by Elsevier Ltd.

1. Introduction

Quantitative trait loci (QTL) mapping has been proven to be a powerful approach for elucidating the genetic architecture of a quantitative trait (Mackay, 2001). In the past decades, statistical methods for QTL mapping have been flourished in the literature developed under different frameworks (Lander and Botstein, 1989; Haley and Knott, 1992; Kruglyak and Lander, 1995; Sen and Churchill, 2001; Wu et al., 2004). Along the line, successful examples from QTL mapping have been well documented in the literature (e.g., Frary et al., 2000; Li et al., 2006). With the development of biotechnology and advanced statistical methods, QTL mapping would still serve as a powerful tool for targeting genetic regions harboring potential genes underlying phenotypic variations.

In nature, phenotypic variation can be displayed in a continuous or discrete scale. For example, the measurement of body weight/height displays in a continuous scale, while measurements such as the number of flowers or the number of new roots generated display in a countable discrete scale. Most statistical methods developed for QTL mapping assume normal distribution for continuous phenotypes which is valid in most cases. For discrete phenotypes such as count data, normal assumption fails in most cases. QTL mapping assuming Poisson regression models provide a standard procedure for the analysis of count data (Shepel et al., 1998; Rebaï, 1997; Sen and Churchill, 2001). In practice, count data are often over- or under-dispersed relative to the Poisson distribution. To take account of data dispersion issue risen naturally from count data, a generalized estimating equation (GEE) approach was applied in QTL mapping count trait (Lange and Whittaker, 2001; Thomson, 2003). More recently Cui et al. (2006) developed a new approach based on the generalized Poisson (GP) regression mixture model to deal with over- or under-dispersion issue. This approach shows relative merits over the GEE type approach in QTL mapping dispersed count data.

Another type of over-dispersion relative to Poisson distribution is that often the number of zero counts are much greater than expected for the Poisson distribution. There are many examples in nature showing this type of variation. For example, in counting tumor lesions on chicken exposed to Merek's disease virus, a chick may have no tumor developed either because it is resistant to the virus, or simply because no disease virus has touched the chick. Consequently, there are two sets of zeros produced. One set of zeros reflects the nature of true zero status and is called *structural zeros*. These zeros indicate that a chick may carry certain genes

* Corresponding author. Tel.: +1 517 432 7098; fax: +1 517 432 1405.
E-mail address: cui@stt.msu.edu (Y. Cui).

0022-5193/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.jtbi.2008.10.003

whose function makes a chick less susceptible to the disease. Other zeros may occur by chance and do not reflect the underlying gene function, and are called *sampling zeros*. When excess zeros exist, regular approaches for modelling count data cannot be applied directly. Statistical approaches for modelling count data with excess zeros than expected have been widely studied. Lambert (1992) derived the zero-inflated Poisson (ZIP) regression model. Mullahy (1986) described the Poisson *hurdle* model which is also termed a *two-part* model by Heilbron (1994). The *hurdle* model is a reparameterization of the ZIP model, but they differ in a regression context. Famoye and Singh (2006) proposed a zero-inflated generalized Poisson (ZIGP) regression model in which zero inflation and Poisson dispersion can be handled in one distribution. However, none of these approaches have been applied in QTL mapping study.

In this article, we propose a rigorous extension of the interval mapping approach to count trait with excess zeros. The effects of QTL on the zero-inflated count trait are modelled through the ZIGP regression mixture model, which subsumes the ZIP model and can handle zero inflation as well as over- or under-dispersion in one distribution setup. We implement the approach using the EM algorithm with the Newton–Raphson procedure embedded in the M-step, and provide a genome-wide scan for testing and estimating the QTL effects. The performance of the proposed method is evaluated through extensive simulation studies. The utility of the developed approach is illustrated through a mouse F_2 intercross data set with the number of cholesterol gallstones as phenotype.

2. Methods

2.1. ZIGP regression model

Let $y_i,\ i=1,\ldots,n$ be the response variables measured in count. The probability density function of y_i assuming a GP distribution is given by

$$p(y_i;\lambda_i,\phi)=\left(\frac{\lambda_i}{1+\phi\lambda_i}\right)^{y_i}\frac{(1+\phi y_i)^{y_i-1}}{y_i!}\exp\left\{\frac{-\lambda_i(1+\phi y_i)}{1+\phi\lambda_i}\right\}$$
$$y_i=0,1,\ldots \qquad (1)$$

where λ_i is the mean of the GP function and can be expressed as a function of genetic and non-genetic factors, i.e., $\lambda_i=\lambda_i(x_i)=\exp(x_i'\boldsymbol{\beta})$, where x_i is a p-dimensional vector of covariates including genetic and non-genetic factors, $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameters, and ϕ is a dispersion parameter. The mean of the GP distribution is given by $E(y_i|\lambda_i,\phi)=\lambda_i$ and the variance is given by $Var(y_i|\lambda_i,\phi)=\lambda_i(1+\phi\lambda_i)^2$. When $\phi=0$, the GP model reduces to the Poisson model and positive or negative value of ϕ corresponds to over- or under-dispersion of data (Famoye, 1993; Cui et al., 2006).

When there are many zero observations than expected, the generalized Poisson regression (GPR) model will not provide good fit in general. The *sampling* zeros can be fitted into the GPR model, but not the *structural* zeros. A good alternative to fit zero-inflated count data would be a ZIGP regression model. The ZIGP model is defined as

$$f(y_i;\lambda_i,\phi)=\begin{cases}\omega_i+(1-\omega_i)p(y_i=0;\lambda_i,\phi) & \text{if } y_i=0\\ (1-\omega_i)p(y_i;\lambda_i,\phi) & \text{if } y_i>0\end{cases} \qquad (2)$$

where $p(y_i;\lambda_i,\phi)$ is the GP density function given in model (1); $0<\omega_i=P(y_i=0)<1$ specifies the probability of zero status including both *sampling* and *structural* zeros; λ_i and ϕ are defined similarly as in model (1); ω_i specifies the probability of *structural* zero status and can be modelled using a logit link function in which $\text{logit}(\omega_i)=\log\omega_i/(1-\omega_i)=z_i'\gamma$, where z_i is the ith row vector of the covariate matrix and γ is the parameter vector. In general, the covariates of X and Z may or may not coincide. When they do coincide, a more parsimonious model can be fitted by supposing that the two linear predictors are related in a certain way. For example, if the same covariates affect ω_i and λ_i, we can write ω_i as a function of λ_i (Lambert, 1992), i.e., $\log\omega_i/(1-\omega_i)=-\tau x_i'\beta$ which leads to $\omega_i=1/(1+\lambda_i^{\tau})$.

For the ZIGP model given in model (2), its mean and variance can be expressed as

$$E(y_i|\lambda_i,\phi,\tau)=(1-\omega_i)\lambda_i$$

$$\begin{aligned}Var(y_i|\lambda_i,\phi,\tau)&=(1-\omega_i)[\lambda_i^2+\lambda_i(1+\phi\lambda_i)^2]-(1-\omega_i)^2\lambda_i^2\\&=E(y_i|\lambda_i,\phi,\tau)[(1+\phi\lambda_i)^2+\omega_i\lambda_i]\end{aligned}$$

The ZIGP regression model with logit link for ω is denoted as $ZIGP(\tau)$ (Famoye and Singh, 2006). In addition to modeling zero inflation by parameter τ, the parameter ϕ also provides a measure of Poisson dispersion. Thus, the $ZIGP(\tau)$ model can take zero inflation and Poisson dispersion into account in the expression of one distribution. When $\phi=0$, the $ZIGP(\tau)$ model reduces to the $ZIP(\tau)$ model defined by Lambert (1992). Therefore, the $ZIGP(\tau)$ model is also a generalization of the $ZIP(\tau)$ model. Large negative or positive value of τ indicates that the zero state becomes more or less likely.

2.2. The finite $ZIGP(\tau)$ mixture model

We have described the $ZIGP(\tau)$ model in general. In the following we will describe how it can be fitted into a QTL mapping framework to map QTLs underlying count trait with many zeros. Statistical methods for QTL mapping dates back to Lander and Botstein's (1989) seminal work in interval mapping. For simplicity, we start with an interval mapping method for zero-inflated count traits assuming an experimental F_2 cross design. The model can be easily extended to other genetic designs, such as backcross or RIL population. An extension of the model to composite interval mapping (CIM) (Zeng, 1994) and multiple interval mapping (MIM) (Kao et al., 1999) is discussed in the following section. Consider an F_2 intercross, initiated with two contrasting homozygous inbred lines with distinct phenotypes. The genetic information inherited by F_2 individuals represents random perturbations of the parental lines. Significant difference of the genotypic means of the three genotypes carried by F_2 individuals at a particular genomic position implies that there is a QTL underlying the quantitative variation of the studied trait at that position, and a QTL can be claimed at that position.

In general, a genetic linkage map can be constructed with molecular markers based on an F_2 segregation population. Assume that a sample of size n is randomly collected from this F_2 population. The observed molecular markers are normally neutral markers which do not show linkage with the studied trait. The real QTL linked with the trait could be located anywhere on the genome and may not be observed. Suppose there is a putative segregating QTL, with alleles Q and q, that show linkage with a zero-inflated count trait. The purpose of a QTL mapping study is to infer the QTL effects as well as their locations on chromosomes. The actual QTL genotypes and positions are unobservable, but can be inferred through the observed molecular markers. The statistical foundation of QTL mapping lies in a mixture model in which each observation y is assumed to have arisen from one of a known or unknown number of components (Lander and Botstein, 1989). Assuming that there are J QTL genotypes contributing to the variation of a count trait, the mixture model is

expressed as

$$y \sim f(y;\boldsymbol{\varphi}) = \pi_1 f_1(y;\varphi_1) + \cdots + \pi_J f(y;\varphi_J) \tag{3}$$

where $\pi = (\pi_1, \ldots, \pi_J)'$ are the mixture proportions (i.e., QTL genotype frequencies) which are constrained to be non-negative and $\sum_{j=1}^{J} \pi_j = 1$; $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_J)'$ are component specific parameters, with φ_j being specific to component j. For an F_2 cross, there are three possible genotypes in offsprings. Therefore, the distribution of y_i for each individual i is modelled through three mixture components. Each mixture proportion represents the conditional probability of the corresponding QTL genotype given on the flanking marker genotype. In our current setting with an F_2 design, the conditional probability of QTL genotype j for individual i given on the observed flanking marker M_i, defined as π_{ij}, can be easily derived as shown in general QTL mapping literature (Wu et al., 2007), where j takes value 2, 1 or 0 depending on whether the QTL genotype is QQ, Qq or qq. The mixture model, therefore, has the form

$$f(y_i;\lambda_i,\phi,\tau) = \pi_{i2} f_2(y_i;\lambda_2,\phi,\tau) + \pi_{i1} f_1(y_i;\lambda_1,\phi,\tau) + \pi_{i0} f_0(y_i;\lambda_0,\phi,\tau) \tag{4}$$

From the above mixture distribution, we can easily compute the unconditional mean and variance of y_i which are expressed as

$$\mu_i = E(y_i) = E[E(y_i|\lambda_i,\phi,\tau)] = \sum_{j=0}^{2} \pi_{ij}(1-\omega_i)\lambda_{ij}$$

and

$$\begin{aligned} Var(y_i) &= E(Var(y_i|\lambda_i,\phi,\tau)) + Var(E(y_i|\lambda_i,\phi,\tau)) \\ &= \sum_{j=0}^{2}(1-\omega_i)[\lambda_{ij}^2 + \lambda_{ij}(1+\phi\lambda_{ij})^2] - [E(y_i)]^2 \end{aligned}$$

Let $x_i = (1, x_{i1}, x_{i2})'$ be a vector for the ith individual, where

$$x_{i1} = \begin{cases} +1 & \text{for } QQ \\ 0 & \text{for } Qq \\ -1 & \text{for } qq \end{cases}$$

and

$$x_{i2} = \begin{cases} 1 & \text{for } Qq \\ 0 & \text{for } QQ \text{ or } qq \end{cases}$$

Then the mean of the ZIGP model for each mixture component, conditional on the QTL genotype G_i, can be expressed as $\lambda_i|G_i = \exp(x_i'\boldsymbol{\beta})$ which leads to

$$\lambda_i|G_i = \begin{cases} \lambda_2 = \exp(\mu + a) & \text{for } QQ \\ \lambda_1 = \exp(\mu + d) & \text{for } Qq \\ \lambda_0 = \exp(\mu - a) & \text{for } qq \end{cases} \tag{5}$$

where $\boldsymbol{\beta} = (\mu, a, d)'$ in which μ is the overall genetic effect, a is the additive genetic effect and d is the dominant genetic effect (Lynch and Walsh, 1998).

2.3. Parameter estimation

Assuming independent observations, the log-likelihood function given the phenotype $\mathbf{y}$ and marker data $\mathbf{M}$ can be expressed as

$$\begin{aligned} \ell_n(\beta,\phi,\tau|\mathbf{y},\mathbf{M}) = \sum_{i=1}^{n} \log\{&\pi_{i2} f_2(y_i;\lambda_{i|2},\phi,\tau) + \pi_{i1} f_1(y_i;\lambda_{i|1},\phi,\tau) \\ &+ \pi_{i0} f_0(y_i;\lambda_{i|0},\phi,\tau)\} \end{aligned} \tag{6}$$

Define $\Omega = (\boldsymbol{\beta},\phi,\tau)' = (\mu,a,d,\phi,\tau)'$ which contains the quantitative genetic parameters, the dispersion parameter and the zero-inflation parameter. The maximum-likelihood estimate (MLE) $\hat{\Omega}$ for Ω can be calculated by solving the partial-derivative equation corresponding to the rth parameter contained in Ω: $\partial\ell_n(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}_r = 0$. In the application of estimation, the positions of QTL θ are treated as known parameters instead of unknown, although we can also obtain their MLEs through iterative steps. Then a grid search approach can be used to estimate the QTL positions. By assuming there is a putative QTL every 1 or 2 cM at marker intervals, we can get the profile of log-likelihood test statistics throughout the entire genome. The positions with respect to the peak of the profile across a linkage group are the MLEs of the QTL positions.

The computational algorithm is in general described as: for any fixed QTL position, we can use the EM algorithm (Dempster et al., 1977) to find the restricted MLE $\hat{\Omega}$. In the M-step, the Newton–Raphson algorithm is applied to get the maximum likelihood estimation for each putative QTL position every 1 or 2 cM distance. The details of the algorithm derivation are given in Appendix A. From the by-product of the Newton–Raphson algorithm in the M-step of the EM algorithm, we can easily get the approximate standard errors of the parameter estimates. The consistency of the MLEs contained in Ω under the ZIGP mixture model can be shown by using Czado et al.'s (2007) consistency argument and applying the results in Chen and Chen (2005). Consistency of the QTL position estimate θ can be established if there exists a QTL in a testing interval.

2.4. Testing QTL existence

Once the parameters are estimated at each testing position by the grid search algorithm, our main interest would be to test if there exists a QTL at certain interval which is responsible for the variation of the count phenotype with many zeros. The hypothesis for such a test can be formulated as

$$\begin{cases} H_0 : a = d = 0 \\ H_1 : \text{at least one parameter is not zero} \end{cases} \tag{7}$$

The null hypothesis states that there is no genetic effect, i.e., there is no expression difference for the three genotypes in an F_2 population. The likelihood ratio (LR) test has been the standard test in testing the QTL effect. The test statistic is calculated as the log of the LR test statistic of the full model (H_1) over the reduced model (H_0):

$$LR = -2\log[L(\tilde{\Omega}) - L(\hat{\Omega})] \tag{8}$$

where $\tilde{\Omega}$ and $\hat{\Omega}$ denote the MLEs of the unknown parameters under H_0 and H_1, respectively. In the mixture model content, the test statistic LR may not follow an asymptotic χ^2 distribution because of the violation of regularity conditions. The distribution of LR at each test position can be assessed through either parametric bootstrap or permutation test. To assess the genome-wide significance, we use the permutation test proposed by Churchill and Doerge (1994).

2.5. Testing zero inflation

In real application, any observed zeros could be from *sampling* zeros or *structural* zeros, or from both. Structural zeros reflect the nature of true zero status and hence are related to the underlying gene function while sampling zeros are merely due to sampling chance. The two sets of zeros are generally non-distinguishable by visualizing data. Statistical tests have been developed to assess the degree of zero inflation. van den Broek (1995) first proposed a simple score test to test zero inflation with a ZIP model setup. Extending van den Broek's work, Jansakul and Hinde (2002) later developed a score test considering covariates in the ZIP model. However, the authors mentioned that the score test cannot be

applied when the link function for zero probability ω is non-linear (e.g., logit link in the current setting). A composite score test should be applied instead which is essentially equivalent to an AIC type of model selection (Jansakul and Hinde, 2002). Jansakul and Hinde (2002) suggested a score information criterion (SIC) initially proposed by Hart (1999), which is an analogy of AIC (Akaike, 1974) based on the score test statistics. Note that we put a special structure on the link function where the Poisson mean and the zero probability are linked by zero inflation parameter τ. The model parameterization is different from the one proposed by Jansakul and Hinde (2002). Thus, the SIC criterion cannot be applied directly because the score function is not well defined under the null of no zero inflation. For this reason, we apply an AIC type of test to check which model, ZIGP or GPR, fits data better. Since the GPR model assumes no zero inflation, the selection process is equivalent to test if there is zero inflation. A model that favors ZIGP against GPR indicates zero inflation.

2.6. Testing data dispersion

In addition to the advantage of assessing zero inflation, the proposed ZIGP model can also take care of data dispersion. When there are potential under- or over-dispersion, the GPR model outperforms the regular Poisson regression model in QTL mapping as revealed by our previous investigation (Cui et al., 2006). The same conclusion applies when comparing the ZIGP and ZIP model. When the dispersion parameter $\phi = 0$, the ZIGP model reduces to the ZIP model indicating no data dispersion. To assess the adequacy of the ZIGP model over the ZIP model, and to determine whether the data are over- or under-dispersed with respect to the GPR model, a test for the dispersion parameter can be formulated by testing $H_0 : \phi = 0$. When the lower bound for $\hat{\phi}$ is not reached, a Wald type test can be conducted in which $\hat{\phi}/\sigma(\hat{\phi})$ may asymptotically follow a standard normal distribution. Given the mixture distribution, further theoretical study is needed to investigate the validity of the Wald test. Alternatively a LR test can also be applied. The sign of the significant test statistics suggests over- or under-dispersion, where negative estimates indicate under-dispersion and positive estimates indicate over-dispersion. We can also apply an AIC-type model selection procedure to choose which model, ZIP or ZIGP, fits the data better.

3. Simulation

Monte Carlo simulations are conducted to evaluate the performance of the proposed ZIGP mixture model for mapping QTL underlying count trait with many zeros by mimicking practical situations. Consider an F_2 population initiated with two inbred lines with which a 80 cM long linkage group composed of five equidistant markers is constructed. Phenotype count data are simulated assuming there is a putative QTL located at 48 cM from the first marker on the linkage group using the derived ZIGP mixture model. The Haldane map function is used to convert the map distance into the recombination fraction. Data are simulated under different scenarios, namely different sample sizes ($n = 100$, 200, and 400), and different patterns of zero inflation (light, mild, heavy) using the proposed ZIGP mixture model. The root mean squared errors (RMSEs) and the power of detecting QTL are reported.

For each simulation case, 100 replicates are performed. Simulation results are tabulated in Table 1. Since the asymptotic distribution of the LR test statistic for testing the existence of QTL in the current framework is unknown and the permutation test is very time consuming, we can use the empirical LOD score of 3 as the threshold to determine the significance of the LR test in simulation. We can also determine the power by simulation studies. The null distribution for the LR statistic can be simulated assuming no genetic effects, i.e., $a = d = 0$. The powers calculated by using LOD 3 threshold and by simulation studies are indicated as P^1 and P^2 in Table 1, respectively. In real analysis, a permutation test should be applied instead. Without loss of generality, we fix the data dispersion parameter and vary the inflation parameter. The MLE of the parameters and their RMSEs listed in the parenthesis are reported. In general, the ZIGP model can provide accurate parameter estimates with reasonable precision as indicated by the mean and RMSE values under different sample sizes and different zero-inflation conditions. The effect of sample size to the parameter estimation and testing power is remarkable. As we expected, large samples always improve the precision of the parameter estimates and power. For example, the RMSE of QTL position estimation is increased from 14.119 to 4.321 when sample size is increased from 100 to 400 under heavy zero-inflation condition. Similar pattern is also observed for other parameter estimates. Meantime, the power is increased from 71%

Table 1
The mean MLEs with their square root mean square errors (RMSEs) (in parentheses) of the parameters and QTL testing power obtained from 100 simulation replicates with different zero-inflation patterns

Heavy inflation								
n	$p = 48$ cM	$\tau = -0.5$	$\phi = 0.01$	$\mu = 2$	$a = 0.5$	$d = 0.3$	P^1	P^2
100	46.42 (14.119)	−0.504 (0.111)	−0.001 (0.023)	2.01 (0.18)	0.506 (0.151)	0.278 (0.242)	71	90
200	47.48 (8.960)	0.512 (0.084)	0.005 (0.015)	2.01 (0.09)	0.488 (0.092)	0.267 (0.140)	91	100
400	48.08 (4.321)	0.508 (0.053)	0.006 (0.011)	2.00 (0.06)	0.493 (0.069)	0.292 (0.079)	100	100
Mild inflation		$\tau = 0$						
100	47.95 (8.437)	−0.006 (0.083)	0.006 (0.014)	1.99 (0.10)	0.491 (0.103)	0.309 (0.116)	92	99
200	48.20 (4.119)	−0.004 (0.071)	0.007 (0.008)	1.99 (0.06)	0.502 (0.065)	0.312 (0.099)	96	100
400	48.32 (2.395)	−0.004 (0.052)	0.008 (0.006)	2.00 (0.05)	0.498 (0.057)	0.297 (0.065)	100	100
Light inflation		$\tau = 0.5$						
100	48.16 (5.101)	0.501 (0.112)	0.006 (0.011)	1.99 (0.08)	0.507 (0.083)	0.307 (0.103)	95	100
200	48.72 (3.481)	0.514 (0.086)	0.008 (0.007)	2.22 (0.06)	0.494 (0.056)	0.299 (0.076)	99	100
400	48.10 (1.896)	0.501 (0.055)	0.009 (0.005)	2.00 (0.04)	0.496 (0.039)	0.296 (0.051)	100	100

P^1 and P^2 denote power calculated by using the LOD 3 and the simulated thresholds, respectively.

to 100%. We also observe the same trend under other zero-inflation conditions.

The effect of zero inflation on parameter estimation as well as testing power is also significant. We know that the *structural* zero stage is determined by parameter τ as well as the effect size of λ_i from equation $\omega_i = 1/(1+\lambda_i^{\tau})$. Even though the zero probability varies from individual to individual depending on an individual's genotype, parameter τ plays a major role in controlling the zero stage probability for given genetic effects. For example, with the given genetic effects listed in Table 1 and $\tau = 0.5$, the zero probability (ω) is 0.22, 0.24 and 0.32 for individual carrying QTL genotype *QQ*, *Qq* and *qq*, respectively. When τ decreases to -0.5, the zero probability increases to 0.78, 0.76 and 0.68 for the three QTL genotypes, respectively. When $\tau = 0$, the zero probability becomes 0.5 regardless of the underlying QTL genotypes. Therefore, $\tau = -0.5$ indicates heavy zero inflation and large value of τ indicates light zero inflation in the current parameter setup. The effect of zero inflation on the genetic parameter estimation as well as testing power can be clearly seen in Table 1. In general, the increase of zero proportion can reduce the testing power and parameter estimation precision. For example, for fixed sample size (say 200), the RMSE for the additive genetic parameter a increases from 0.056 to 0.092 when τ decreases from 0.5 (light inflation) to -0.5 (heavy inflation), a 39% reduction in precision. For a sample of size 100, the testing power is increased from 71% to 95% when the zero condition is changed from heavy to light (P^1 in Table 1). As sample size increases to 400, this difference is invisible.

When simulated cutoffs are used, the power is increased under the three zero-inflation conditions due to small thresholds used. For example, the power is increased from 71% to 90% when data are heavily zero-inflated under $n = 100$. With the simulated cutoffs, no significant power difference is observed for data showing mild and light inflation. To further show the impact of zero status on testing power, we compared the LR test statistic across the simulated linkage group under the three simulated zero-inflation conditions. Fig. 1 clearly shows the difference of the LR values for different τ values under different sample sizes. For example, when $n = 100$, large LR values are consistently observed across the linkage group with light zero inflation ($\tau = 0.5$). As the proportion of zeros increases, the LR values are significantly decreased. The same trend can be observed under large samples ($n = 400$).

A boxplot of the QTL position estimates comparing the performance of the ZIGP and GPR is given in Fig. 2 which displays the inter-quantile and the range of the estimated QTL position. Outliers are indicated in stars. The notch indicates a robust estimate of the uncertainty about the median. The dotted vertical line represents the true QTL location which is simulated at 48 cM. In all simulation, we fix the dispersion and the genetic parameters and vary the sample size and the inflation parameter. The figure indicates that the ZIGP model gives more precise estimates of the QTL position than the GPR model under different sample sizes and zero-inflation status. Also as we expected, the increase of sample size can dramatically improve the precision of QTL position estimation. The effect of zero inflation on the position estimation can also be clearly seen from the figure in which smaller variation for position estimation is observed under the light inflation condition compared to other status.

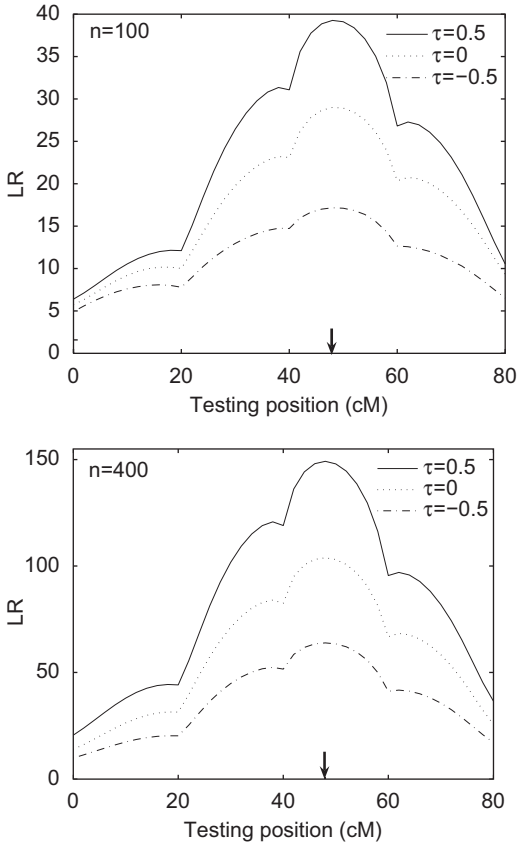

Fig. 1. The LR profile plots averaged over 100 simulation replicates under different sample sizes (100 and 400) assuming different zero-inflation conditions. The arrow sign indicates the simulated QTL position (48 cM).

4. Case study

To show the utility of the developed approach, we apply the *ZIGP*(τ) mixture model to a real data set in mapping QTL underlying cholesterol gallstone formation. Cholesterol gallstones are abnormal masses of a solid mixture of cholesterol crystals. As one of the most common digestive disorders and yet very expensive to treat, gallstone disease has affected people for centuries (Portincasa et al., 2006). Several studies have shown that cholesterol gallstone formation is a complex genetic trait with unique genetic basis (e.g., Lyons et al., 2002, 2003, 2005; Wittenburg et al., 2003). A number of QTLs in determining an individual predisposition to develop cholesterol gallstones have been mapped using inbred mouse strains (Lyons et al., 2002, 2003, 2005; Wittenburg et al., 2003). In these literature, a gallstone scoring approach was applied to calculate an overall phenotypic index measure of gallstone formation in which gallstone number is considered as one of the components (Lyons et al., 2002; Wittenburg et al., 2003). The scores were then used as phenotypes assuming normal distribution for QTL mapping. No study has been reported for QTL mapping focusing on count trait with many zeros.

To apply the data to our ZIGP model, we use the gallstone number as the phenotype for the mapping purpose. The data contain an intercross population generated from two inbred mouse strains PERA/Ei and I/LnJ. Total 279 F_2 mice were collected and genotyped. A genetic linkage map was constructed using 107 genetic markers, with a total length of 1382.3 cM, representing a good coverage of 19 mouse autosomal chromosomes (Wittenburg et al., 2003). The interested count phenotype is the number of formation of cholesterol monohydrate crystals, translucent "sandy" gallstones and opaque "solid" gallstones. Fig. 3 shows the histogram of the cholesterol gallstone counts collected from the 279 F_2 individuals. The large proportion of zeros (~57%) indicates that a ZIGP model considering zero inflation might be more appropriate than a regular GPR model.

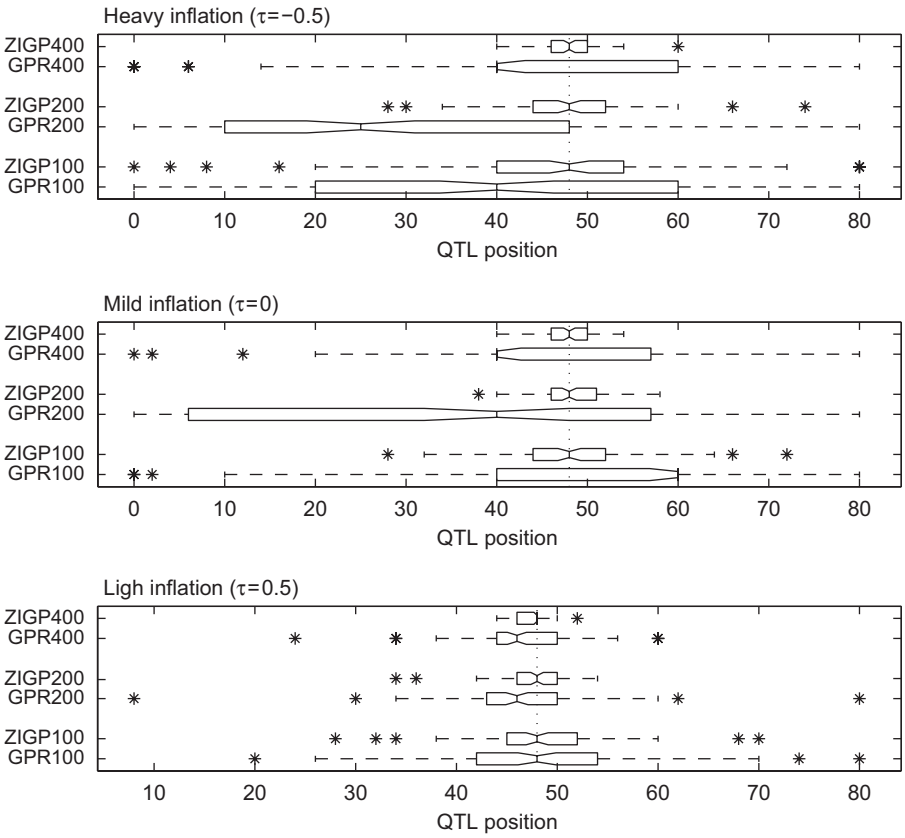

Fig. 2. The boxplot of the estimated QTL position from 100 simulation replicates. Data are simulated using the ZIGP model under different degrees of zero inflation with parameters listed in Table 1, and are analyzed using the GPR and ZIGP models. The true QTL position is simulated at 48 cM away from the first marker indicated by the vertical dotted line.

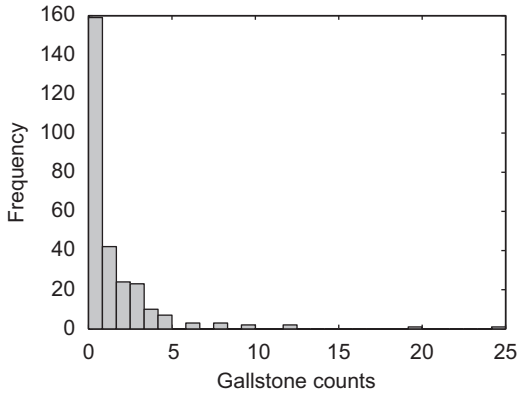

Fig. 3. Histogram of the cholesterol gallstone counts measured in 279 F_2 individuals.

Three types of statistical models, GPR, ZIP and the newly proposed ZIGP are applied to analyze the data. In all three analysis, only genetic factors described in Eq. (5) are considered. A genome-wide linkage scan are conducted and the three models lead to different LR profile throughout the genome. The AIC values are reported at every scan position. Among the three models, smaller AIC values indicate goodness-of-fit of the underlying model to the data. Fig. 4 plots the differences of the AIC values obtained by fitting the model using the ZIP and GPR against the one fitted with the ZIGP model. As indicated by the figure, the ZIP model gives the largest AIC values among the three models across the genome. The GPR and the ZIGP models provide better fits than the ZIP model. Comparing the GPR and the ZIGP models, the later gives consistently smaller AIC values across most testing positions than the ones fitted by the GPR model. Thus, only the results obtained by the ZIGP model are focused in this section. In reality, different genotypes may have different functions, leading to different levels of reaction for an individual when exposed to environmental stimuli, and thus to different levels of gallstone zero status. Models (e.g., ZIGP) that can take care of different zeros status displayed in both structural and sampling forms should therefore be more meaningful and powerful. The real data analysis indicates that the ZIGP model fits the data better than the other models and is more powerful.

By genome-wide scanning for QTLs at every 2 cM within each marker interval across the 19 mouse chromosomes, eight QTLs that trigger effects on mouse gallstone formation are detected. The genome-wide log-LR profile plot is shown in Fig. 5. The dashed horizontal line indicates the 5% genome-wide significance level and the dotted line indicates the 5% chromosome-wide significance level. The figure clearly indicates that there are eight QTLs are detected by the ZIGP model. Only one QTL located on chromosome 10 is significant at the genome-wide significance level. This QTL is located right on the maker position (*D10Mit102*) indicating that this marker is a potential candidate for gallstone formation. All the other QTLs are only significant at the chromosome-wide level and hence are suggestive QTLs. Table 2 tabulates the estimated genetic effects with the asymptotic standard errors given in the parenthesis. The QTLs detected on chromosomes 4, 10, and 15 are consistent with the results obtained by Wittenburg et al. (2003). In addition, we detect three new QTLs located on chromosomes 8, 16 and 19 which are not reported before. Given the close location of the three QTLs

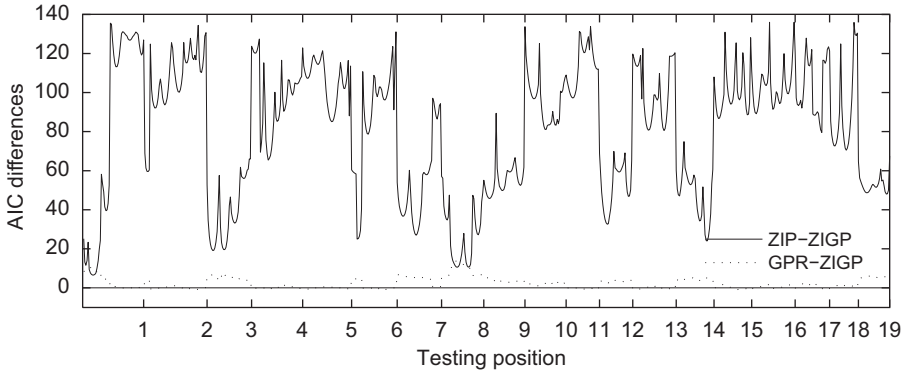

Fig. 4. The differences of the AIC information values calculated from the GPR, ZIP and ZIGP models across 19 chromosomes. The solid and dotted curves represent the AIC differences between the ZIP and ZIGP models, and between the GPR and ZIGP models, respectively.

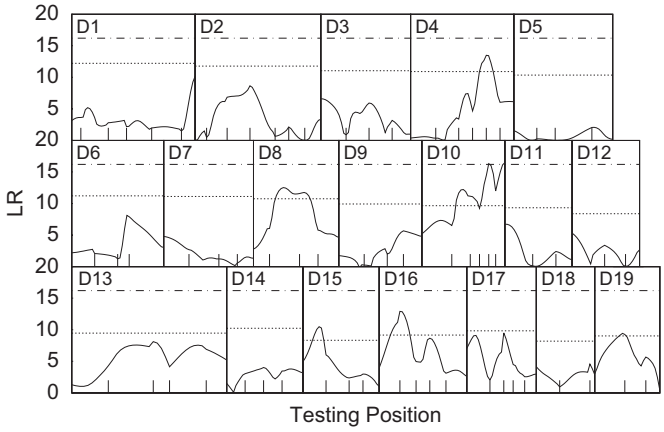

Fig. 5. The profiles of the log-likelihood ratios (LR) between the full and reduced model estimated from the ZIGP mixture model for gallstone numbers across chromosomes 1–19. The genomic positions corresponding to the peaks of the curves are the MLEs of the QTL positions. The genome-wide threshold values for claiming the existence of a QTL is given as the dash-dot horizonal line, and the chromosome-wide threshold value is marked as the horizontal dotted line.

detected on chromosome 10, a multiple QTL model might be needed to test if they are all real.

5. Model extension

The interval mapping approach has certain drawbacks by only considering one QTL at a time (Jansen, 1993; Zeng, 1994). QTLs located not in the testing interval may have interfering effect on the tested QTL and the mapping results might not be conclusive. Moreover, the mapping precision might be reduced as well (Zeng, 1994). A number of approaches have been developed to overcome the problems. Among those, two approaches are most popularly applied, namely the CIM (Zeng, 1994) and the MIM (Kao et al., 1999) methods. Here we extend the developed single-QTL model to multiple QTL analysis based on these two popular approaches.

5.1. Composite interval mapping

The proposed model can be easily extended to fit into the CIM framework. The idea of CIM is to incorporate multiple regression analysis into interval mapping by considering markers outside a testing interval to control background genetic effect in order to improve the precision and power of QTL detection. To extend the proposed model to CIM framework in an F_2 population, we consider the following mean function at a testing interval with flanking markers j and $j+1$:

$$\lambda_i|G_i = \exp\left\{\mu + x_i a + (1-|x_i|)d + \sum_{\ell\neq j,j+1}(x_\ell a_\ell + (1-|x_\ell|)d_\ell)\right\}$$

where μ is the overall mean; x_i represents the QTL genotype taking values 1, 0, and -1 corresponding to QTL genotype QQ, Qq and qq, respectively; x_ℓ is the indicator variable for the background marker genotype taking values 1, 0 and -1 corresponding to marker genotype MM, Mm and mm, respectively. Background marker selection can be done by applying standard methods developed for regular CIM. The EM algorithm derived for interval mapping can be applied to estimate parameters.

5.2. Multiple interval mapping

As shown in literatures, when multiple QTLs are located in the same linkage group, considering one QTL at a time could bias QTL identification and estimation (Jansen, 1993; Zeng, 1994). The problem can be solved by applying a multiple QTL mapping model. One popular approach is the MIM approach (Kao et al., 1999) which uses multiple marker intervals simultaneously to map multiple QTL of epistatic interactions throughout a linkage map. The proposed model can also be extended to fit into the MIM framework. Suppose there are K QTLs, $Q_1,\ldots,Q_K$, located on the genome. The mean function for an individual i who carries genotype G_i can be expressed as

$$\lambda_i|G_i = \exp\left\{\mu + \sum_{k=1}^{K} x_{ik}a_k + \sum_{k=1}^{K} z_{ik}d_k + \sum_{j\neq k}^{K}(x_{ik}x_{ij})i_{a_k a_j} + \sum_{j\neq k}^{K}(x_{ik}z_{ij})i_{a_k d_j} + \sum_{j\neq k}^{K}(z_{ik}x_{ij})i_{d_k a_j} + \sum_{j\neq k}^{K}(z_{ik}z_{ij})i_{d_k d_j}\right\}$$

where x_{ik} and x_{ij} are coded as 1 or -1 corresponding to the QTL genotype QQ and qq, respectively; z_{ik} and z_{ij} are coded as 1 if the QTL genotype is Qq and 0 otherwise; a_k and d_k are the additive and dominant effect for QTL $\mathbf{Q}_k$; $i_{a_k a_j}, i_{a_k d_j}, i_{d_k a_j}$ and $i_{d_k d_j}$ are the pairwise interaction effects between QTL $\mathbf{Q}_k$ and $\mathbf{Q}_j$. Stepwise or chunkwise selection procedure can be implemented to identify and separate linked QTL (Kao et al., 1999).

Table 2
Estimated genetic effects and their asymptotic standard errors (in the parenthesis) of the detected QTLs

Ch	Marker interval	τ	ϕ	μ	a	d	LR/LOD
4	*D4Mit204*	0.8545	0.3954	0.8848	0.3054	−0.4225	13.48/2.93[a]
		(0.013)	(0.051)	(0.044)	(0.031)	(0.031)	
8	*D8Mit147–Mit271*	−0.0665	0.1038	1.5315	0.8429	−0.3190	12.51/2.71[a]
		(0.076)	(0.283)	(0.097)	(0.066)	(0.071)	
10	*D10Mit148–Mit22*	0.3961	0.3058	1.0356	0.3249	−0.55526	12.21/2.65[a]
		(0.018)	(0.060)	(0.041)	(0.029)	(0.029)	
10	*D10Mit66–Mit12*	0.3320	0.2873	1.0657	0.3269	−0.6202	16.3/3.54[a]
		(0.020)	(0.063)	(0.041)	(0.030)	(0.028)	
10	*D10Mit102*	0.9495	0.4053	0.8512	0.3262	−0.4457	16.49/3.58[b]
		(0.012)	(0.051)	(0.046)	(0.034)	(0.031)	
15	*D15Mit174–Mit184*	1.1922	0.4165	0.8629	−0.2146	−0.5256	10.49/2.28[a]
		(0.010)	(0.050)	(0.048)	(0.033)	(0.035)	
16	*D16Mit122*	1.0876	0.4394	0.8338	−0.2359	−0.4409	12.91/2.8[a]
		(0.012)	(0.049)	(0.047)	(0.034)	(0.032)	
19	*D19Mit32–Mit40*	0.31956	0.2302	0.8650	0.3508	0.7613	9.42/2.04[a]
		(0.056)	(0.164)	(0.081)	(0.058)	(0.058)	

Note: The significance is at level 5% through 200 permutation tests.
[a] Refers to chromosome-wide significance.
[b] Refers to genome-wide significance.

6. Discussion

Count traits are often observed in nature. Due to its discrete nature, often many zeros may be occurred. While Poisson regression or other approaches such as the one using the generalized estimating equation can be applied to analyze count data, such approaches often fail when there are many zeros. In this article, we have developed an efficient method in QTL mapping for count data with many zeros. Since zero status may be due to *sampling* zero or *structural* zero, a model that can distinguish these two types of zeros should be more appropriate. The proposed zero-inflated generalized Poisson regression mixture model can take care of both zero inflation and data dispersion and hence should be more appropriate in dissecting the genetic effect of an underlying QTL on a count trait. Computer simulations demonstrate that the model has high power in mapping QTL for zero-inflated count data with reasonable sample size and is quite robust in various situations.

The results in Table 1 show that the mapping power is affected by the degree of zero inflation, especially when sample size is small. High power is obtained when data show light inflation if threshold LOD 3 is used (indicated by P^1). When simulated cutoffs are used, all powers are increased due to small thresholds used (indicated by P^2). The LR cutoffs by simulations under different sample sizes and zero-inflation conditions are ranged from 9.5 to 10.5 which are less than the LOD 3 threshold. The small simulated cutoffs may be due to small linkage group size (80 cM). Real data analysis using permutations indicates that the genome-wide threshold is close to the LOD 3 threshold. The LR profile plots (Fig. 1) clearly show the impact of zero status on LR values and hence on the testing power. Also, the QTL location is more precisely estimated when data show light inflation compared to mild or heavy inflation. The results also indicate that zero inflation does affect QTL parameter estimation. This information suggest us that in real data analysis, one has to be cautious in drawing a conclusion when the proportion of zeros are large, especially with small sample size. The effects of data dispersion on parameter estimate and testing power are also studied. Similar results as reported in Cui et al. (2006) are observed and hence are omitted.

The proposed ZIGP mixture model is a generalization of both the ZIP and GPR mixture models. When the dispersion parameter ϕ is zero, the ZIGP model reduces to the ZIP model. When the zero probability ω approaches zero, the model is reduced to the GPR model. In reality, which model is more appropriate to fit the data can be decided through a model selection procedure. As demonstrated by the real data analysis, the information criterion such as AIC always favors the ZIGP model. More QTL are detected by the ZIGP model than the other models.

Noted that the ZIGP mapping model is derived under the maximum likelihood framework. Hence the likelihood-based inference procedures can be easily applied under the current framework such as the goodness-of-fit test and the residual analysis as described in Cui et al. (2006). The inclusion of potential outliers or influential points may affect QTL effect estimation and inference. They can be easily identified by using these model diagnostic procedures. In this article, we have developed our method in the context of an F_2 population. The model can be easily modified to fit into a more general mapping framework such as CIM or MIM. Extension to other populations such as backcross, RIL or combined crosses are straightforward. A computer program written in R is available upon request.

Acknowledgments

The authors wish to thank B. Paigen for providing the mouse data set and one anonymous referee for valuable comments. This work was supported in part by a NSF grant (DMS 0707031) and by a Michigan State University intramural research grant.

Appendix A

The EM algorithm with the F_2 population is derived as follows. Define $c_i = 2$, 1 or 0 if the QTL genotype is *QQ*, *Qq* or *qq*, respectively, with its distribution function.

$$f(c_i) = \prod_{j=0}^{2} \pi_{ij}^{c_{ij}}$$

where $\pi_{ij} = P(c_{i|j} = j)$. Thus,

$$f(y_i|c_i) = \prod_{j=0}^{2} [p_j(y_i|\lambda_{i|j}, \phi, \tau)]^{c_{ij}}$$

and

$$f(y,c)=\prod_{i=1}^{n}f(y_i,c_i)=\prod_{i=1}^{n}f(y_i|\lambda_{i|j})f(c_i)$$
$$=\prod_{i=1}^{n}\left\{\prod_{j=0}^{2}[p_j(y_i|\lambda_{ij},\phi,\tau)]^{c_{ij}}\pi_{ij}^{c_{ij}}\right\}$$

Then the complete log-likelihood function is given by

$$\ell^c=\sum_{i=1}^{n}\sum_{j=0}^{2}c_{i|j}\log p_j(y_i|\lambda_{i|j},\phi,\tau)+\sum_{i=1}^{n}\sum_{j=0}^{2}c_{i|j}\log\pi_{ij}$$

Since

$$f(c_{i|j}|y_i)=\frac{f(y_i,c_{i|j})}{f(y_i)}=\frac{f(y_i|c_{i|j})f(c_{i|j})}{\sum_{s=0}^{2}\pi_{is}p_s(y_i|\lambda_s,\phi,\tau)}$$
$$=\frac{(\pi_{ij}p_j(y_i|\lambda_{i|j},\phi,\tau))^{c_{ij}}(\pi_{is\neq ij}p_{s\neq j}(y_i|\lambda_{s\neq j},\phi,\tau))^{1-c_{ij}}}{\sum_{s=0}^{2}\pi_{is}p_s(y_i|\lambda_s,\phi,\tau)}$$

Thus, in the E-step, we calculate $\prod_{i|j}$ at the (t)th iteration, which is

$$\prod_{i|j}^{(t)}=E[c_{i|j}|y_i,\pi,\lambda_{i|j},\phi,\tau]=\frac{\pi_{ij}p_j(y_i|\lambda_{i|j},\phi,\tau)}{\sum_{s=0}^{2}\pi_{is}p_s(y_i|\lambda_s,\phi,\tau)}\tag{A.1}$$

And then replace the missing value $c_{i|j}$ by $\prod_{i|j}$ in the log-likelihood function with the complete data. In the M-step, we calculate the MLE of the parameters by using the Newton–Raphson algorithm iteratively by maximizing the complete data likelihood function,

$$Q^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{i|j}^{(t)}\log p_j(y_i|\lambda_{i|j},\phi,\tau)+\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{i|j}^{(t)}\log\pi_{ij}$$

By the Newton–Raphson iteration method, we need to calculate the first and second partial derivatives. In the rest of the derivation, we shall use $\xi_{ij}=\lambda_{ij}(1+\phi\lambda_{ij})^{-1}$, $\eta_{ij}=\lambda_{ij}^{\tau}\exp(-\xi_{ij})$, which are used in the first and the second partial derivatives given below:

$$\frac{\partial}{\partial\boldsymbol{\tau}}\log\ell(\Omega)^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{\frac{\log\lambda_{ij}}{1+\lambda_{ij}^{\tau}}-\frac{\log\lambda_{ij}}{1+\eta_{ij}}I(y_i=0)\right\}$$

$$\frac{\partial}{\partial\boldsymbol{\phi}}\log\ell(\boldsymbol{\Omega})^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{\frac{\xi_{ij}^2\eta_{ij}}{1+\eta_{ij}}I(y_i=0)\right.$$
$$\left.+\left[-\xi_{ij}y_i+\frac{(y_i-1)y_i}{1+\phi y_i}-\frac{\lambda_{ij}(y_i-\lambda_{ij})}{(1+\phi\lambda_{ij})^2}\right]I(y_i>0)\right\}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}_r}\log\ell(\boldsymbol{\Omega})^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{\frac{\tau x_{ij}}{1+\lambda_{ij}^{\tau}}-\frac{\left(\tau+\dfrac{\xi_{ij}^2\eta_{ij}}{\lambda_{ij}}\right)x_{ij}}{1+\eta_{ij}}I(y_i=0)\right.$$
$$\left.+\frac{(y_i-\lambda_{ij})x_{ij}}{(1+\phi\lambda_{ij})^2}I(y_i>0)\right\}$$

$$\frac{\partial^2}{\partial\boldsymbol{\tau}^2}\log\ell(\boldsymbol{\Omega})^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{\eta_{ij}\left(\frac{\log\lambda_{ij}}{1+\eta_{ij}}\right)^2I(y_i=0)\right.$$
$$\left.-\lambda_{ij}^{\tau}\left(\frac{\log\lambda_{ij}}{1+\lambda_{ij}^{\tau}}\right)^2\right\}$$

$$\frac{\partial^2}{\partial\boldsymbol{\phi}^2}\log\ell(\boldsymbol{\Omega})^{(t)}=\sum_{j=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{\frac{\eta_{ij}\xi_{ij}^3(\xi_{ij}-2-2\eta_{ij})}{(1+\eta_{ij})^2}I(y_i=0)\right.$$
$$\left.+\left[y_i\xi_{ij}^2+\frac{2\lambda_{ij}^2(y_i-\lambda_{ij})}{(1+\phi\lambda_{ij})^3}-\frac{y_i^2(y_i-1)}{(1+\phi y_i)^2}\right]I(y_i>0)\right\}$$

$$\frac{\partial^2}{\partial\boldsymbol{\beta}_r^2}\log\ell(\boldsymbol{\Omega})^{(t)}$$
$$=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{-\frac{(\tau x_{ij})^2\lambda_{ij}^{\tau}}{(1+\lambda_{ij}^{\tau})^2}+\left[\frac{x_{ij}^2\xi_{ij}^2\eta_{ij}\lambda_{ij}^{-1}(2\lambda_{ij}^{-1}+\eta_{ij}-\xi_{ij}-1)}{1+\eta_{ij}}\right.\right.$$
$$\left.+\frac{x_{ij}^2\eta_{ij}(3\xi_{ij}^2\lambda_{ij}^{-1}-\tau^2+2\xi_{ij}^2)}{(1+\eta_{ij})^2}\right]I(y_i=0)$$
$$\left.+\frac{x_{ij}^2\lambda_{ij}(2y_i\phi+3\lambda_{ij}\phi+1)}{(1+\phi\lambda_{ij})^3}I(y_i>0)\right\}$$

$$\frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}_r}\log\ell(\boldsymbol{\Omega})^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{\lambda_{ij}^{-1}-\frac{\log(\lambda_{ij})\tau\lambda_{ij}^{\tau}}{\lambda_{ij}(1+\lambda_{ij}^{\tau})^2}\right.$$
$$\left.-\left[\frac{\lambda_{ij}}{1+\eta_{ij}}-\frac{\log(\lambda_{ij})\eta_{ij}(x_{ij}\tau\lambda_{ij}-\xi_{ij}^2)}{(1+\eta_{ij})^2}\right]I(y_i=0)\right\}$$

$$\frac{\partial^2}{\partial\tau\partial\boldsymbol{\phi}}\log\ell(\boldsymbol{\Omega})^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{\frac{\log(\lambda_{ij})\eta_{ij}\xi_{ij}^2}{(1+\eta_{ij})^2}I(y_i=0)\right\}$$

$$\frac{\partial^2}{\partial\boldsymbol{\beta}_r\partial\boldsymbol{\phi}}\log\ell(\boldsymbol{\Omega})^{(t)}=\sum_{i=1}^{n}\sum_{j=0}^{2}\prod_{ij}^{(t)}\left\{-\frac{x_{ij}\tau\eta_{ij}\xi_{ij}^2}{(1+\eta_{ij})^2}\right.$$
$$\left.-\frac{\xi_{ij}\eta_{ij}x_{ij}\lambda_{ij}^{-1}}{1+\eta_{ij}}I(y_i=0)\right\}$$

The Hessian matrix at the (t)th iteration is given by $H^{(t)}=\partial^2Q^{(t)}/\partial\Omega_s\Omega_j$,which leads to the updated parameters Ω at the $(t+1)$th iteration,

$$\Omega^{(t+1)}=\Omega^{(t)}-[H^{(t)}]^{-1}u'\tag{A.2}$$

where u is a vector of the first derivative of $Q^{(t)}$ with respect to Q_r. The EM algorithm is repeated between Eqs. (A.1) and (A.2) until certain convergence criteria are satisfied.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723.

Chen, Z., Chen, H., 2005. On some statistical aspects of the interval mapping for QTL detection. Stat. Sin. 15, 909–925.

Churchill, G.A., Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics 138, 963–971.

Cui, Y.H., Kim, D.-Y., Zhu, J., 2006. On the generalized Poisson regression mixture model for mapping quantitative trait loci with count data. Genetics 174, 2159–2172.

Czado, C., Erhardt, V., Min, A., Wagner, S., 2007. Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. Stat. Modelling 7, 125–153.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via EM algorithm. J. R. Stat. Soc. Ser. B 39, 1–38.

Famoye, F., 1993. Restricted generalized Poisson regression model. Commun. Stat. Theory Methods 22, 1335–1354.

Famoye, F., Singh, K.P., 2006. Zero-inflated generalized Poisson model with an application to domestic violence data. J. Data Sci. 4, 117–130.

Frary, A., Nesbitt, T.C., Frary, A., Grandillo, S., van der Knapp, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K.B., Tanksley, S.D., 2000. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289, 85–88.

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

Hart, J.D., 1999. Testing the fit of functions in fully specified likelihood models. In: Proceedings of the 14th International Workshop on Statistical Modeling, pp. 19–29.

Heilbron, D.C., 1994. Zero-altered and other regression models for count data with added zeros. Biometrical J. 36, 531–547.
Jansakul, N., Hinde, J.P., 2002. Score tests for zero-inflated Poisson models. J. Comput. Stat. Data Anal. 40, 75–96.
Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135, 205–211.
Kao, C.H., Zeng, Z.-B., Teasdale, R.D., 1999. Multiple interval mapping for quantitative trait loci. Genetics 152, 1203–1216.
Kruglyak, L., Lander, E.S., 1995. A nonparametric approach for mapping quantitative trait loci. Genetics 139, 1421–1428.
Lambert, D., 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.
Lander, E.S., Botstein, D., 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121, 185–199.
Lange, C., Whittaker, J.C., 2001. Mapping quantitative trait loci using generalized estimating equations. Genetics 159, 1325–1337.
Li, C.B., Zhou, A.L., Sang, T., 2006. Rice domestication by reducing shattering. Science 311, 1936–1939.
Lynch, M., Walsh, B., 1998. Genetics and Analysis of Quantitative Traits. Sinauer, Sunderland, MA.
Lyons, M.A., Wittenburg, H., Li, R., Walsh, K.A., Leonard, M.R., Churchill, G.A., Carey, M.C., Paigen, B., 2002. New quantitative trait loci that contribute to cholesterol gallstone formation detected in an intercross of CAST/Ei and 129S1/SvImJ inbred mice. Physiol. Genomics 14, 225–239.
Lyons, M.A., Wittenburg, H., Li, R., Walsh, K.A., Leonard, M.R., Korstanje, R., Churchill, G.A., Carey, M.C., Paigen, B., 2003. Lith6: a new QTL for cholesterol gallstones from an intercross of CAST/Ei and DBA/2J inbred mouse strains. J. Lipid Res. 44, 1763–1771.
Lyons, M.A., Korstanje, R., Li, R., Sheehan, S.M., Walsh, K.A., Rollins, J.A., Carey, M.C., Paigen, B., Churchill, G.A., 2005. Single and interacting QTLs for cholesterol gallstones revealed in an intercross between mouse strains NZB and SM. Mamm. Genome 16, 152–163.
Mackay, T.F.C., 2001. Quantitative trait loci in Drosophila. Nat. Rev. Genet. 2, 11–20.
Mullahy, J., 1986. Specification and testing of some modified count data models. J. Econometrics 33, 341–365.
Portincasa, P., Moschetta, A., Palasciano, G., 2006. Cholesterol gallstone disease. Lancet 368, 230–239.
Rebaï, A., 1997. Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. Genetics 69, 69–74.
Sen, S., Churchill, G.A., 2001. A statistical framework for quantitative trait mapping. Genetics 159, 371–387.
Shepel, L.A., Lan, H., Haag, J.D., Brasic, J.M., Gheen, M.E., et al., 1998. Genetic identification of multiple loci that control breast cancer susceptibility in the rat. Genetics 149, 289–299.
Thomson, P., 2003. A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. Genet. Sel. Evol. 35, 257–280.
van den Broek, J., 1995. A score test for zero inflation in a Poisson distribution. Biometrics 51, 738–743.
Wittenburg, H., Lyons, M.A., Li, R., Churchill, G.A., Carey, M.C., Paigen, B., 2003. FXR and ABCG5/ABCG8 as determinants of cholesterol gallstone formation from quantitative trait locus mapping in mice. Gastroenterology 125, 868–881.
Wu, R.L., Ma, C.-X., Lin, M., Casella, G., 2004. A general framework for analyzing the genetic architecture of developmental characteristics. Genetics 166, 1541–1551.
Wu, R.L., Casella, G., Ma, C.-X., 2007. Statistical Genetics of Quantitative Traits: Linkage Maps and Qtl. Springer, New York.
Zeng, Z.-B., 1994. Precision mapping of quantitative trait loci. Genetics 136, 1457–1468.