DOI: 10.1534/genetics.105.047241

# A Logistic Regression Mixture Model for Interval Mapping of Genetic Trait Loci Affecting Binary Phenotypes

**Weiping Deng,* Hanfeng Chen† and Zhaohai Li*,‡,[1]**

**Department of Statistics, George Washington University, Washington, District of Columbia 20052, †Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio 43403 and ‡Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, Maryland 20892*

Manuscript received June 24, 2005
Accepted for publication October 24, 2005

ABSTRACT

Often in genetic research, presence or absence of a disease is affected by not only the trait locus genotypes but also some covariates. The finite logistic regression mixture models and the methods under the models are developed for detection of a binary trait locus (BTL) through an interval-mapping procedure. The maximum-likelihood estimates (MLEs) of the logistic regression parameters are asymptotically unbiased. The null asymptotic distributions of the likelihood-ratio test (LRT) statistics for detection of a BTL are found to be given by the supremum of a $\chi^2$-process. The limiting null distributions are free of the null model parameters and are determined explicitly through only four (backcross case) or nine (intercross case) independent standard normal random variables. Therefore a threshold for detecting a BTL in a flanking marker interval can be approximated easily by using a Monte Carlo method. It is pointed out that use of a threshold incorrectly determined by reading off a $\chi^2$-probability table can result in an excessive false BTL detection rate much more severely than many researchers might anticipate. Simulation results show that the BTL detection procedures based on the thresholds determined by the limiting distributions perform quite well when the sample sizes are moderately large.

A variety of quantitative and/or qualitative traits in plants and animals are attributed mainly to genotypes at certain chromosome locations (so-called trait loci). Extensive studies on quantitative trait locus (QTL) mapping have been conducted since Sax (1923). For experimental organisms, there are several statistical methods developed for mapping QTL systematically. Soller and Brody (1976) proposed the traditional approach of detecting a QTL by using a nearby marker as a surrogate for a QTL. Lander and Botstein (1989) developed the interval-mapping (IM) method to test for a QTL in a flanking marker interval. As the genotypes of the QTL are unobservable, the distribution of the quantitative trait follows a finite mixture model, say a mixture of two normal distributions as many authors proposed. For some recent developments on the QTL mapping method, see, *e.g.*, Zeng (1994), Hackett and Weller (1995), Kao and Zeng (1997), Lange and Whittaker (2001), Broman (2003), Thomson (2003), Chen and Chen (2005), Sillanpaa and Bhattacharjee (2005), and Xu *et al.* (2005).

The QTL mapping methods and corresponding statistical procedures discussed above are for quantitative traits with continuous phenotypic values. In real life, however, some disease traits are qualitative, and in many cases they are binary (binary trait locus, BTL). For example, the fusiform rust disease in loblolly pine is described as presence or absence of the formation of galls on a polygenic basis. Current approaches to the BTL mapping usually assume that the binary trait of interest is to be controlled by an underlying continuous liability so that the QTL mapping methods can be modified for the BTL mapping. On the other hand, as several authors remarked (see, *e.g.*, Xu and Atchley 1996), mapping genes for binary traits is more complicated than that for continuous traits. Although many descriptive procedures are proposed, the development of inferential procedures has been left behind due to the great challenge met in determining the threshold for detecting a BTL with a controlled false BTL detection rate. A reasonable account of large sample studies for the BTL interval-mapping methods has been lacking. As a result, the thresholds for significance tests and critical values of interval estimates of the parameters have to be established using an empirical approach (Yi and Xu 1999b) or a repeated sampling technique, *e.g.*, a permutation test (Churchill and Doerge 1994) or bootstrapping test (Visscher *et al.* 1996). It is also common practice due to the lack of large sample theory that the thresholds for detecting a BTL on the basis of the log-likelihood ratio or the LOD score are incorrectly determined by reading off a $\chi^2$-probability table as suggested by the classic asymptotic $\chi^2$-theory, with the perception that the actual thresholds may be only *slightly*

[1]*Corresponding author:* Department of Statistics, George Washington University, 2140 Pennsylvania Ave., NW, Washington, DC 20052. E-mail: zli@gwu.edu

larger than those from a $\chi^2$-table (see, *e.g.*, Haley and Knott 1992; Zeng 1994; Xu and Atchley 1996).

In this article, the finite logistic regression mixture models for the BTL interval mapping and the methods under these models are developed and investigated with a quite general scheme: (a) both the backcross and intercross populations are considered; (b) readily available covariates are accommodated as one of the main features of the models; and (c) other markers in addition to the two flanking markers can be naturally incorporated into the models as part of the covariates to control the genetic background of other chromosomal regions. The research is intended to establish the direct asymptotic procedures for determining the thresholds for the BTL interval-mapping methods on the basis of the log-likelihood ratio or equivalently the LOD score.

Note that some covariates such as environmental and nutritional conditions, age, sex, and location of field are often readily available in genetic research. While including several explanatory covariates can substantially increase the efficiency of the statistical analysis, the identifiability of the logistic regression mixture model in the parameters of interest can also be augmented via the covariates (Follmann and Lambert 1991). It is widely recognized that identifiability is a central issue in the finite mixture models (see, *e.g.*, Chen and Chen 2001).

## MODELS AND METHODS

Suppose that a BTL of an experimental organism is flanked by two markers, say marker I and marker II. Assume that the putative BTL has two alleles $B$ and $b$, marker I has two alleles $M_1$ and $m_1$, and marker II has $M_2$ and $m_2$. To detect a BTL in the flanked marker interval, a series of experiments can be arranged so that eventually two inbred lines $P_1$ and $P_2$ can be obtained, where any individual from $P_1$ has homozygous genotype $M_1BM_2/M_1BM_2$ at the three locations on a chromosome: marker I, the BTL, and marker II, while any individual from $P_2$ has homozygous genotype $m_1bm_2/m_1bm_2$. Thus, all individuals in the $F_1$ population, the progeny of $P_1$ and $P_2$, have the same heterozygous genotype $M_1BM_2/m_1bm_2$. Let $\gamma$ be the recombination fraction between marker I and marker II. Assume that $\gamma$ is specific or can be preestimated. Also assume that the two flanking markers, marker I and marker II, are linked; *i.e.*, $\gamma < 0.5$. In fact, in interval mapping $\gamma$ is usually chosen to be substantially below 0.5. An experimental population can then be produced either by $F_1$ individuals backcrossing with one of the parents $P_1$ and $P_2$, resulting in a backcross population, or by $F_1$ intermating with $F_1$ itself, resulting in an intercross population also called an $F_2$ population.

Let $Y$ be the observable binary trait of concern: $Y = 1$ or 0 according to presence or absence of the trait on a randomly selected individual. Suppose that in addition to the putative BTL, the binary trait $Y$ may also be affected by some explanatory covariates, say $x = (x_1, \ldots, x_p)^\tau$, where $a^\tau$ is the transpose of a vector $a$. The covariate vector $x$ can be fixed effects or random effects, or part of it is fixed and part of it random. In the case of random effects, a conditional model of $Y$ given $x$ is adopted to form the likelihood function. As such, part of the explanatory covariates can be the index variables of genotypes of some other markers in addition to marker I and marker II. For example, if an individual is homozygous at the $j$th marker, define $x_j = 1$; otherwise, $x_j = 0$. As commented by Kao and Zeng (1997), combining interval mapping with multiple markers may greatly improve the identification of a BTL. Other explanatory covariates can be environmental and nutritional conditions, age, sex, location of field, and so on.

We now present the models of $Y$ and the methods of BTL detection combining the IM method in the backcross and intercross populations, respectively.

**Model and method with backcross population:** For specification, consider a backcross population to be the progeny of $P_1$ and $F_1$ so that the individuals in the backcross population have four different genotypes at marker I and marker II: $M_1M_2/M_1M_2$, $M_1M_2/M_1m_2$, $M_1M_2/m_1M_2$, and $M_1M_2/m_1m_2$, coded as 1, 2, 3, and 4. Let $J$ denote the code of the genotype for a randomly selected individual at marker I and marker II, $J = 1, 2, 3$, and 4. The genotypes of markers, *i.e.*, the values of $J$, are observable with the probabilities

$$q_{\mathrm{bc}}(1) = q_{\mathrm{bc}}(4) = (1-\gamma)/2 \quad \text{and} \quad q_{\mathrm{bc}}(2) = q_{\mathrm{bc}}(3) = \gamma/2. \tag{1}$$

At a putative BTL, the genotype can be homozygous $BB$ or heterozygous $Bb$. Let $r$ be the recombination fraction between marker I and the BTL and $s$ the recombination fraction between the BTL and marker II. Without interference, $s$ can be computed from $\gamma$ and $r$ as $s = (\gamma - r)/(1 - 2r)$. The genetic association among the two flanking markers and the BTL can be described by the conditional probabilities $p_{\mathrm{r}}(j) = P(BB \mid J = j)$ that a randomly selected individual has homozygous genotype $BB$ given $J = j$ as follows:

$$\begin{aligned} p_{\mathrm{r}}(1) &= 1 - p_{\mathrm{r}}(4) = (1-r)(1-r-\gamma)/\{(1-\gamma)(1-2r)\} \\ p_{\mathrm{r}}(2) &= 1 - p_{\mathrm{r}}(3) = (1-r)(\gamma-r)/\{\gamma(1-2r)\}. \end{aligned} \tag{2}$$

Suppose that given covariate value $x$ and genotype value $u$ at the BTL ($u = 1$ if $BB$ and $u = 0$ if $Bb$), the probability distribution of $Y = y$ follows a logistic model,

$$\begin{aligned} &\pi(y \mid x, \beta + \alpha^{1-u}, \lambda) \\ &\quad = \exp\{(\beta + \alpha^{1-u} + \lambda^\tau x)y\}/[1 + \exp\{\beta + \alpha^{1-u} + \lambda^\tau x\}], \end{aligned} \tag{3}$$

(Bonney 1986; Hosmer and Lemeshow 1989), where the parameters $\beta$ and $\alpha$ and the $1 \times p$ parameter vector $\lambda$ are unknown. Since the genotype of the putative BTL is unobservable, the probability distribution for $Y$ given $J = j$ and $x$ can be constructed hieratically according to the homozygous or heterozygous genotype of the BTL, so that a logistic regression mixture model for detection of a BTL by the interval-mapping method is induced as follows: for $y = 0, 1$,

$$\begin{aligned} f(y \mid j,\ x,\ \theta) &= P\{Y = y \mid J = j, x\} \\ &= p_{\mathrm{r}}(j)\pi(y \mid x,\ \beta,\ \lambda) + \{1 - p_{\mathrm{r}}(j)\}\pi(y \mid x,\ \beta + \alpha,\ \lambda), \end{aligned} \tag{4}$$

where $\theta = (r, \beta, \alpha, \lambda^{\tau})$ is the parameter vector of interest with $p + 3$ components, and $p_{\mathrm{r}}(j)$ and $\pi$ are given by (2) and (3). The parameter $r$ is the genetic distance between marker I and the BTL, $\beta$ is the base-line gene effect on the binary trait and $\alpha$ is the effect of the BTL, and $\lambda$ is the effects of the explanatory covariate vector $x$. When a random sample $y_i, j_i, x_i, i = 1, \ldots, n$, of size $n$ is observed on the binary trait $Y$, the flanking marker genotype code $J$, and the explanatory covariate $x$, the log-likelihood function for $\theta$ is

$$l_n(\theta) = \sum_{i=1}^{n} \log\{p_{\mathrm{r}}(j_i)\pi_i(\beta, \lambda) + \{1 - p_{\mathrm{r}}(j_i)\}\pi_i(\beta + \alpha, \lambda)\}, \tag{5}$$

where

$$\pi_i(\beta, \lambda) = \pi(y_i \mid x_i, \beta, \lambda). \tag{6}$$

The maximum-likelihood estimate (MLE) $\hat{\theta}$ for $\theta$ is such that it solves the partial derivative equation: $\partial l_n(\theta)/\partial\theta = 0$. A computational algorithm can be as follows. For any fixed $r$, the EM algorithm (Dempster *et al.* 1977) is used to find the restricted MLE $\hat{\theta}(r)$, with Newton-Raphson employed in the M-step (detailed instructions are given in the appendix). Then to obtain $\hat{r}$ and hence $\hat{\theta} = \hat{\theta}(\hat{r})$, vary $r$ over the interval $[0, \gamma]$ with a small increment of 1 or 2 cM at a time.

Clearly the identifiability question of the model (4) in $\theta$ must be settled before one can meaningfully discuss any inferential procedure on the basis of the likelihood function $l_n(\theta)$. First, it is clear that if $\alpha = 0$, $f(y|j, x, \theta) = \pi(y|x, \beta, \lambda)$ and the mixture model reduces to the ordinary logistic regression model so that the model is unidentifiable in $r$ since the reduced model is free of $r$. To settle the identifiability problem in other cases, define a cumulative probability function

$$G_j(\mu) = p_{\mathrm{r}}(j)I(\mu \geq \beta) + \{1 - p_{\mathrm{r}}(j)\}I(\mu \geq \beta + \alpha),$$

where $I\{\cdot\}$ is the indicator function. The mixture model $f(y|j, x, \theta)$ can be reparameterized by $\lambda$ and the mixing distribution $G_j$ as follows:

$$f(y \mid j,\ x,\ \theta) = f(y \mid x,\ \lambda,\ G_j) = \int \pi(y \mid x,\ \mu,\ \lambda)\, dG_j(\mu).$$

In light of Theorem 2 of Follmann and Lambert (1991), $f(y \mid x, \lambda, G_j) = f(y \mid x, \lambda', G_j')$ holds for all $y$, $x$, and $j = 1, 2, 3, 4$, if and only if $\lambda = \lambda'$ and $G_j = G_j'$ for $j = 1, 2, 3, 4$. On the other hand, for $r$ fixed, $(\beta, \alpha)$ identifies $G_j$; if $\alpha \neq 0$, $(r, \beta, \alpha)$ identifies $G_j$. These results imply the following:

i. The parameter $(\beta, \alpha, \lambda)$ identifies the model (4) with $r$ fixed.
ii. The parameter $(r, \beta, \alpha, \lambda)$ identifies the model (4) if $\alpha \neq 0$.
iii. When $\alpha = 0$, the model (4) reduces to the ordinary logistic regression model and is unidentifiable in $r$.

Applying Wald's (1949) consistency argument and using the techniques developed in Chen and Chen (2005), the MLEs of $\beta$, $\alpha$, and $\lambda$ under the logistic regression mixture model (4) are consistent, but the MLE of $r$ is inconsistent. However, if indeed there is a BTL in the flanking interval, *i.e.*, $\alpha \neq 0$, then the MLE of $r$ is consistent.

Detecting a BTL in the flanking marker interval is accomplished by testing $H_0$: $\alpha = 0$ *vs.* $H_1$: $\alpha \neq 0$. Denote the MLE of $\theta$ under the full model (4) by $\hat{\theta} = (\hat{r}, \hat{\beta}, \hat{\alpha}, \hat{\lambda})$, while the MLE under the null model is denoted by $\hat{\theta}_0 = (\gamma, \hat{\beta}_0, 0, \hat{\lambda}_0)$. The null hypothesis is rejected if the likelihood-ratio test (LRT) statistic

$$T_{\mathrm{n}} = 2\{l_n(\hat{\theta}) - l_n(\hat{\theta}_0)\}$$

exceeds a threshold value, where $l_n(\theta)$ is given by (5). When the sample size $n$ is large, the null distribution of $T_{\mathrm{n}}$ can be approximated by that of

$$T = \sup_{0 \leq r \leq \gamma} \left[ \sum_{j=1}^{4} \sqrt{q_{\mathrm{bc}}(j)}\{p_{\mathrm{r}}(j) - \tfrac{1}{2}\} Z_j / \tau_{\mathrm{r}} \right]^2, \tag{7}$$

where $Z_1, \ldots, Z_4$ are four independent standard normal variables, $\tau_{\mathrm{r}}^2 = \sum_{j=1}^{4} q_{\mathrm{bc}}(j)\{p_{\mathrm{r}}(j) - \frac{1}{2}\}^2$, and $q_{\mathrm{bc}}(j)$ and $p_{\mathrm{r}}(j)$ are defined by (1) and (2), respectively. Thus a threshold or a $P$-value based on the LRT statistic $T_n$ can be determined asymptotically by the distribution of $T$. Justification of this asymptotic testing procedure is given in the appendix.

It should be noted that if $\log_{10}$ is chosen to define the logarithm of likelihood function, the log-likelihood ratio becomes the so-called LOD. Although $\log_{10}$ might be a more popular choice than $\log_e$ historically in the genetics community, $\log_e$ is adopted in this article to enjoy convenience with the calculus and asymptotic analysis of $T_{\mathrm{n}}$. Needless to say, the LOD score and $T_{\mathrm{n}}$ are only a scale transform of each other ($T_{\mathrm{n}} = 2\mathrm{LOD}/\log_{10} e$) and hence equivalent for summarizing the evidence for a BTL. For those who would feel more

**TABLE 1**

**Thresholds determined by the limiting distribution ($T$) of the LRT under the backcross population**

| | Marker interval length (cM) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Level (%) | 2 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | $\chi^2_1$ | $\chi^2_2$ |
| 10 | 3.05 | 3.23 | 3.41 | 3.54 | 3.63 | 3.71 | 3.77 | 3.82 | 3.87 | 2.71 | 4.61 |
| 5 | 4.25 | 4.45 | 4.65 | 3.79 | 4.89 | 4.97 | 5.04 | 5.10 | 5.15 | 3.84 | 5.99 |
| 1 | 7.11 | 7.36 | 7.60 | 7.76 | 7.88 | 7.98 | 8.06 | 8.12 | 8.18 | 6.63 | 9.21 |

The last two columns display the thresholds by $\chi^2$-distributions for comparisons.

comfortable with the LOD score than with the LRT statistic $T_n$, an adjusting formula and procedure is immediate: the threshold of the LOD score can be asymptotically determined by the distribution of

$$\frac{\log_{10} e}{2} T,$$

or $0.217T$ approximately, where $T$ is given by (7).

Note that the distribution of $T$ is convenient to implement for determining a threshold at a prespecified significance level via a Monte Carlo method. The null limiting distribution is determined solely by four independent standard normal random variables, free of the nuisance parameters $\beta$ and $\lambda$ and free of the distribution of the covariate $x$. Table 1 displays the thresholds determined by the distribution of $T$ at the levels 1, 5, and 10% with nine different marker interval sizes (between 2 and 40 cM). In each case, 1,000,000 Monte Carlo repetitions of four independent standard normal variables were performed. The supremum of the $\chi^2$-process, as defined by (7), was obtained by exhaustive search over $r \in [0, \gamma]$ in a grid increment of 1 cM (the centimorgan unit was converted to the recombination fraction $r$ by Haldane's mapping function). From Table 1, the difference between an actual threshold and the $\chi^2$-threshold is clearly notable. The consequences of the difference on the false BTL detection rate are shown in the end of this section.

**Model and method with intercross population:** In addition to the four genotypes at marker I and marker II in the backcross population, there are another five observable genotypes in the intercross population: $M_1m_1/M_2m_2$, $M_1m_1/m_2m_2$, $m_1m_1/M_2M_2$, $m_1m_1/M_2m_2$, and $m_1m_1/m_1m_2$, coded as 5, 6, 7, 8, and 9, respectively. Continue to denote by $J$ the code of the genotype of a randomly selected individual from the intercross population. Let $P(J = j) = q_{ic}(j)$, $j = 1, \ldots, 9$. Without interference,

$$\begin{aligned}
&q_{ic}(1) = q_{ic}(9) = (1-\gamma)^2/4\\
&q_{ic}(2) = q_{ic}(4) = q_{ic}6) = q_{ic}(8) = \gamma(1-\gamma)/2\\
&q_{ic}(3) = q_{ic}(7) = \gamma^2/4\\
&q_{ic}(5) = \{(1-\gamma)^2 + \gamma^2\}/2.
\end{aligned}$$

At the putative BTL, there are now three different genotypes: *BB*, *Bb*, and *bb*. Let $p_{r,1}(j) = P(BB \mid J = j)$, $p_{r,2}(j) = P(bb \mid J = j)$, and $p_{r,3}(j) = P(Bb \mid J = j)$. Recall that $s = (\gamma - r)/(1 - 2r)$ is the recombination fraction between the BTL and marker II. Then

$$\begin{aligned}
&p_{r,1}(1) = (1-r)^2(1-s)^2/(1-\gamma)^2\\
&p_{r,1}(2) = (1-r)^2 s(1-s)/\{\gamma(1-\gamma)\}\\
&p_{r,1}(3) = (1-r)^2 s^2/\gamma^2\\
&p_{r,1}(4) = r(1-r)(1-s)^2/\{\gamma(1-\gamma)\}\\
&p_{r,1}(5) = 2rs(1-r)(1-s)/\{\gamma^2 + (1-\gamma)^2\}\\
&p_{r,1}(6) = r(1-r)s^2/\{\gamma(1-\gamma)\}\\
&p_{r,1}(7) = r^2(1-s)^2/\gamma^2\\
&p_{r,1}(8) = r^2 s(1-s)/\{\gamma(1-\gamma)\}\\
&p_{r,1}(9) = r^2 s^2/(1-\gamma)^2
\end{aligned}$$

and for $j = 1, 2, \ldots, 9$,

$$\begin{aligned}
p_{r,2}(j) &= p_{r,1}(9 - j + 1),\\
p_{r,3}(j) &= 1 - p_{r,1}(j) - p_{r,2}(j).
\end{aligned}$$

For the $i$th randomly selected individual with $Y = y_i$, $x = x_i$, and $J = j_i$, because there are three different but unobservable genotypes at the BTL, the probability model is a mixture of three logistic regression distributions,

$$\begin{aligned}
&p_{r,1}(j)\pi(y_i \mid x_i, \beta, \lambda) + p_{r,2}(j)\pi(y_i \mid x_i, \beta + \alpha, \lambda)\\
&\quad + p_{r,3}(j)\pi(y_i \mid x_i, \beta + \nu, \lambda),
\end{aligned} \tag{8}$$

where $\eta = (r, \beta, \alpha, \nu, \lambda^\tau)$ is the parameter vector with $4 + p$ components. With a random sample of size $n$, the log-likelihood function for $\eta$ is

$$\begin{aligned}
l_n(\eta) = \sum_{i=1}^{n} \log\{&p_{r,1}(j)\pi_i(\beta, \lambda) + p_{r,2}(j)\pi_i(\beta + \alpha, \lambda)\\
&+ p_{r,3}(j)\pi_i(\beta + \nu, \lambda)\},
\end{aligned} \tag{9}$$

where $\pi_i(\beta, \lambda) = \pi(y_i \mid x_i, \beta, \lambda)$ as defined in (6).

Similar to the backcross case, it can be seen that the MLEs of $\beta$, $\alpha$, $\nu$, and $\lambda$ are always consistent under the model (8), although the MLE of $r$ is inconsistent when $\alpha = \nu = 0$. However, when either $\alpha$ or $\nu$ or both are not 0, the MLE of $r$ is consistent.

**TABLE 2**

**Thresholds determined by the limiting distribution (*S*) of the LRT under the intercross population**

| | Marker interval length (cM) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Level (%) | 2 | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | $\chi^2_2$ | $\chi^2_3$ |
| 10 | 5.14 | 5.41 | 5.65 | 5.82 | 5.95 | 6.04 | 6.12 | 6.18 | 6.23 | 4.61 | 6.25 |
| 5 | 6.55 | 6.84 | 7.14 | 7.31 | 7.46 | 7.57 | 7.66 | 7.72 | 7.78 | 5.99 | 7.81 |
| 1 | 9.86 | 10.21 | 10.57 | 10.73 | 10.87 | 10.97 | 11.05 | 11.14 | 11.19 | 9.21 | 11.34 |

The last two columns display the $\chi^2$-thresholds for comparisons.

Detecting a BTL in the flanking interval is now accomplished by testing $H_0$: $\alpha = \nu = 0$ *vs.* $H_1$: $\alpha \neq 0$ or $\nu \neq 0$. The LRT statistic is

$$S_n = 2\{l_n(\hat{\eta}) - l_n(\hat{\eta}_0)\},$$

where $l_n(\eta)$ is defined by (9), $\hat{\eta}$ is the unrestricted MLE of $\eta$, and $\hat{\eta}_0 = (\gamma, \hat{\beta}_0, 0, 0, \hat{\lambda}_0^\tau)$ is the MLE of $\eta$ under the null model. The null hypothesis is rejected and hence a BTL is detected if $S_n$ exceeds a threshold that can be determined asymptotically by the distribution of

$$S = \sup_{0\le r\le\gamma}\left\{\left[\sum_{j=1}^{9}\sqrt{q_{ic}(j)}a_r(j)Z_j/\tau_{r,1}\right]^2 + \left[\sum_{j=1}^{9}\sqrt{q_{ic}(j)}b_r(j)Z_j/\tau_{r,2.1}\right]^2\right\}, \quad (10)$$

where $Z_1, \ldots, Z_9$ are nine independent standard normal variables, and

$$\begin{aligned}
a_r(j) &= p_{r,1}(j) - \tfrac{1}{4}\\
b_r(j) &= \{p_{r,2}(j) - \tfrac{1}{4}\} - (\tau_{r,12}/\tau^2_{r,1})\{p_{r,1}(j) - \tfrac{1}{4}\}\\
\tau^2_{r,i} &= \sum_{j=1}^{9} q_{ic}(j)\{p_{r,i}(j) - \tfrac{1}{4}\}^2, \quad i = 1, 2\\
\tau_{r,12} &= \sum_{j=1}^{9} q_{ic}(j)\{p_{r,1}(j) - \tfrac{1}{4}\}\{p_{r,2}(j) - \tfrac{1}{4}\}\\
\tau^2_{r,2.1} &= \tau^2_{r,2} - \tau^2_{r,12}/\tau^2_{r,1}.
\end{aligned}$$

A proof of the limiting distribution (10) of $S_n$ is outlined in the APPENDIX.

If the LOD score is used to measure the evidence of a BTL, it is immediate that a threshold of the LOD can be determined asymptotically by the distribution of $(\log_{10}e/2)S$ or $0.217S$.

As in the backcross population case, the limiting distribution (10) is free of nuisance parameters $\beta$ and $\lambda$ and free of the distribution of $x$ and is entirely determined by nine independent standard normal variables. Therefore, (10) is easy to use to approximate a threshold or a *P*-value of the LRT $S_n$ via a Monte Carlo method. Note that $\sum_{j=1}^{9} a_r(j)b_r(j) = 0$, implying $S > \chi^2_2$ stochastically. Table 2 displays the thresholds determined by the distribution of $S$ at levels of 1, 5, and 10% with nine different marker interval sizes (between 2 and 40 cM). As expected, the thresholds determined by the distribution of $S$ are between $\chi^2_2$ and $\chi^2_3$ and can be very close to $\chi^2_3$, depending on the size of the flanking marker interval.

**Consequences of using a $\chi^2$-threshold:** The differences between the asymptotic thresholds determined correctly by the limiting distribution and incorrectly by a $\chi^2$-distribution have been noted in Tables 1 and 2. The consequences of the differences can further be described in terms of the false BTL detection rate. A false detection of a BTL over the flanking marker interval occurs when there is actually not any BTL presented in the marker interval, but the LRT statistic (or equivalently the LOD score) exceeds a threshold preset to satisfy a desired level of false BTL detection, say 5%. It is thus clear that if the threshold used is lower than it is supposed to be, the resulting false BTL detection rate will be higher than expected.

Table 3 displays the actual asymptotic false BTL detection rates of the LRT statistic when a $\chi^2$-threshold at a level of 5% is incorrectly used (LANDER and BOTSTEIN 1989; XU and ATCHLEY 1996; LYNCH and WALSH 1998, p. 447). Five different marker interval lengths (2, 10, 20, 30, and 40 cM) are considered. As seen from Table 3, the actual false BTL detection rate can be substantially higher than expected and is more than doubled in both the backcross and intercross cases, when the interval length is as short as 40 cM. Even in the event of an unusually short interval length of 2 cM, the actual false BTL detection is >1% higher than the desired 5%.

## SIMULATION STUDIES

Simulation studies were intended primarily to assess the consequences of using the limiting distributions ($n = \infty$) of $T$ and $S$ in real life ($n < \infty$). Therefore, for simplicity, only two flanking markers and one continuous covariate were included in the simulation models.

The simulation studies were planned with the following consideration: sample sizes $n = 200$ and 500, marker interval size = 20 cM (*i.e.*, $\gamma = 0.091$), absence or

**TABLE 3**

**Asymptotic false BTL detection rates by the LRT statistic when a $\chi^2$-threshold at a level of 5% is incorrectly used**

| Population | $\chi^2$-Threshold | Marker interval length (cM) | | | | |
|---|---|---|---|---|---|---|
| | | 2 | 10 | 20 | 30 | 40 |
| Backcross | 3.84 | 0.063 | 0.078 | 0.089 | 0.096 | 0.101 |
| Intercross | 5.99 | 0.066 | 0.085 | 0.098 | 0.106 | 0.112 |

$P(T > 3.84)$ in the backcross case and $P(S > 5.99)$ in the intercross case.

presence of a BTL at the location of 4 cM ($r = 0.038$) from marker I, and a standard normal covariate $x$ with coefficient $\lambda = 1$. Both the backcross and intercross populations with different combinations of effects were considered. The nominal significance level was 5%. The simulated thresholds, *i.e.*, 4.89 in the backcross case (Table 1) and 7.14 in the intercross case (Table 2), were adopted.

In each scenario, 5000 Monte Carlo repetitions were performed. With each Monte Carlo sample, the EM algorithm was used to obtain the MLEs and the supremum of a $\chi^2$-process was obtained by exhaustive search from 0 to 20 cM in a grid increment of 1 cM. The simulation results are summarized in Tables 4 and 5.

The simulation results demonstrate that the asymptotic distributions approximate the sampling distributions well. Specifically, (i) the probabilities of type I errors are very close to the nominal one, (ii) the detection rates for a BTL correctly reflect the sample sizes and the sizes of discrepancy between an alternative and the null, and (iii) the estimates of the parameters confirm the expected consistency behaviors.

The simulation results also show the weak performance of the estimates for the BTL locations. As a referee remarked, it is often the case that the QTL estimates are drawn to the center of the flanking marker interval. This phenomenon is also evident in Tables 4 and 5. It appears that performance of the estimates of $r$ depends largely on the BTL's genetic distance to the flanking markers as well as on the discrepancy between the null model and the alternative. The performance gets weaker when the BTL is closer to a flanking marker and is best when the BTL is in the middle of the marker interval. It should be noted that this interesting dynamic consistency property of the MLE of $r$ does not show up in Xu and Atchley's (1996) simulation report because they considered only the BTLs located in the middles of the marker intervals.

## CONCLUSIONS AND DISCUSSION

The finite logistic regression mixture models for the BTL interval mapping and the methods under these models have been developed and investigated.

**TABLE 4**

**Simulated powers of the LRT and mean MLEs with the standard deviations (in parentheses) under the backcross population**

| | Loc | $\beta$ | $\alpha$ | Mean MLE (SD) | | | | Power |
|---|---|---|---|---|---|---|---|---|
| | | | | Lôc | $\hat{\beta}$ | $\hat{\alpha}$ | $\hat{\lambda}$ | |
| $n = 200$ | — | 0 | 0 | 9.95 (9.08) | 0.00 (0.26) | 0.00 (0.40) | 1.03 (.19) | 0.050 |
| | 4 | 0 | 0.5 | 7.97 (8.09) | −0.02 (0.24) | 0.55 (0.37) | 1.03 (0.20) | 0.308 |
| | 4 | 0 | 0.8 | 6.31 (6.77) | −0.02 (0.23) | 0.85 (0.36) | 1.03 (0.20) | 0.617 |
| | 4 | 0 | 1 | 5.61 (5.99) | −0.01 (0.23) | 1.05 (0.36) | 1.03 (0.21) | 0.802 |
| | 10 | 0 | 0.5 | 10.26 (8.22) | −0.01 (0.24) | 0.54 (0.37) | 1.03 (0.20) | 0.293 |
| | 10 | 0 | 0.8 | 10.15 (7.36) | −0.01 (0.24) | 0.84 (0.36) | 1.03 (0.20) | 0.594 |
| | 10 | 0 | 1 | 9.96 (6.82) | −0.01 (0.24) | 1.03 (0.37) | 1.03 (0.21) | 0.776 |
| $n = 500$ | — | 0 | 0 | 9.77 (9.04) | 0.00 (0.16) | 0.00 (0.25) | 1.01 (0.12) | 0.056 |
| | 4 | 0 | 0.5 | 6.37 (6.88) | −0.01 (0.15) | 0.52 (0.22) | 1.01 (0.12) | 0.623 |
| | 4 | 0 | 0.8 | 5.10 (5.10) | −0.01 (0.15) | 0.82 (0.22) | 1.01 (0.12) | 0.944 |
| | 4 | 0 | 1 | 4.67 (4.32) | −0.01 (0.15) | 1.02 (0.23) | 1.01 (0.13) | 0.995 |
| | 10 | 0 | 0.5 | 10.14 (7.37) | −0.01 (0.15) | 0.51 (0.22) | 1.01 (0.12) | 0.598 |
| | 10 | 0 | 0.8 | 10.05 (5.91) | 0.00 (0.15) | 0.81 (0.23) | 1.01 (0.12) | 0.931 |
| | 10 | 0 | 1 | 10.05 (5.12) | 0.00 (0.15) | 1.01 (0.24) | 1.01 (0.13) | 0.990 |

The asymptotic thresholds at level 5% given in Table 1 are used; two sample sizes, $n = 200$ and 500 are considered. The BTL location is in centimorgans and denoted by Loc. The covariate $x \sim N(0, 1)$ and $\lambda = 1$. Monte Carlo repetition size is 5000 in each case.

**TABLE 5**

**Simulated powers of the LRT and mean MLEs with the standard deviations (in parentheses) under the intercross population**

| | | | | | Mean MLE (SD) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Loc | β | α | $\nu$ | Lôc | $\hat{\beta}$ | $\hat{\alpha}$ | $\hat{\nu}$ | $\hat{\lambda}$ | Power |
| $n = 200$ | — | 0 | 0 | 0 | 10.02 (8.77) | −0.01 (0.39) | 0.01 (0.55) | 0.01 (0.52) | 1.04 (0.20) | 0.056 |
| | 4 | 0 | 0 | 0.5 | 7.89 (7.89) | −0.01 (0.38) | 0.01 (0.54) | 0.54 (0.51) | 1.04 (0.20) | 0.199 |
| | 4 | 0 | 0 | 0.8 | 6.01 (6.49) | −0.02 (0.37) | 0.01 (0.52) | 0.86 (0.50) | 1.04 (0.20) | 0.450 |
| | 4 | 0 | 0.5 | 1 | 5.93 (6.43) | −0.02 (0.36) | 0.54 (0.53) | 1.06 (0.50) | 1.04 (0.21) | 0.492 |
| | 4 | 0 | 0 | 1 | 5.16 (5.49) | −0.02 (0.36) | 0.01 (0.51) | 1.06 (0.50) | 1.04 (0.21) | 0.643 |
| | 10 | 0 | 0 | 0.5 | 9.98 (7.99) | 0.00 (0.38) | 0.01 (0.55) | 0.52 (0.51) | 1.04 (0.20) | 0.180 |
| | 10 | 0 | 0 | 0.8 | 9.88 (7.03) | 0.00 (0.37) | 0.01 (0.53) | 0.82 (0.50) | 1.04 (0.20) | 0.401 |
| | 10 | 0 | 0.5 | 1 | 9.85 (7.09) | 0.00 (0.37) | 0.53 (0.54) | 1.02 (0.50) | 1.04 (0.21) | 0.438 |
| | 10 | 0 | 0 | 1 | 9.84 (6.37) | 0.00 (0.37) | 0.01 (0.53) | 1.02 (0.51) | 1.04 (0.21) | 0.568 |
| $n = 500$ | — | 0 | 0 | 0 | 9.93 (8.73) | −0.01 (0.24) | 0.01 (0.34) | 0.00 (0.32) | 1.02 (0.12) | 0.052 |
| | 4 | 0 | 0 | 0.5 | 5.90 (6.39) | −0.01 (0.23) | 0.01 (0.32) | 0.52 (0.30) | 1.02 (0.12) | 0.440 |
| | 4 | 0 | 0 | 0.8 | 4.51 (4.32) | −0.01 (0.22) | 0.01 (0.31) | 0.83 (0.30) | 1.02 (0.12) | 0.856 |
| | 4 | 0 | 0.5 | 1 | 4.45 (4.35) | −0.02 (0.22) | 0.52 (0.32) | 1.03 (0.30) | 1.01 (0.12) | 0.891 |
| | 4 | 0 | 0 | 1 | 4.17 (3.61) | −0.01 (0.22) | 0.01 (0.31) | 1.03 (0.30) | 1.02 (0.13) | 0.970 |
| | 10 | 0 | 0 | 0.5 | 9.94 (6.93) | 0.00 (0.23) | 0.00 (0.33) | 0.50 (0.31) | 1.01 (0.12) | 0.388 |
| | 10 | 0 | 0 | 0.8 | 9.93 (5.36) | 0.00 (0.23) | 0.01 (0.32) | 0.80 (0.31) | 1.01 (0.12) | 0.785 |
| | 10 | 0 | 0.5 | 1 | 9.90 (5.41) | 0.00 (0.23) | 0.51 (0.33) | 1.00 (0.31) | 1.01 (0.12) | 0.838 |
| | 10 | 0 | 0 | 1 | 9.96 (4.51) | 0.00 (0.23) | 0.01 (0.32) | 1.00 (0.32) | 1.01 (0.13) | 0.937 |

The asymptotic thresholds at level 5% given in Table 2 are used; two sample sizes, $n = 200$ and 500 are considered. The BTL location is in centimorgans and denoted by Loc. The covariate $x \sim N(0, 1)$ and $\lambda = 1$. Monte Carlo repetition size is 5000 in each case.

The binary trait of interest is assumed to be affected by a BTL and some explanatory covariates. Since the genotype of the BTL is unobservable, the finite logistic regression mixture models are used to develop the procedures for detecting a BTL in a flanking marker interval. Compared to the existing methods, the new approach accommodates trait-associated covariates and allows multiple markers that are expected to help improve the efficiency of BTL mapping. Both the backcross and intercross populations are considered. It should be pointed out that the logistic regression mixture models in this research assume no interaction among the covariates and the genotype of the putative BTL.

The MLEs of the logistic regression parameters are asymptotically unbiased. The null asymptotic distributions of the likelihood-ratio test statistics for detecting a BTL are presented and found to be given by the supremum of a $\chi^2$-process. Nevertheless, the limiting null distributions are free of the nuisance parameters and free of the null model parameters and the distribution of the covariates and are explicitly determined through four (backcross case) or nine (intercross case) independent standard normal variables. Thus the thresholds for detecting a BTL in a flanking marker interval can be easily determined by the limiting distributions via a Monte Carlo method. The asymptotic thresholds turn out to be substantially larger than those given by the ordinary $\chi^2$-distribution and the actual false BTL detection rates can be much higher than many researchers might have anticipated when a $\chi^2$-threshold is incorrectly used.

The asymptotic procedures developed are applicable to large samples in general. For small or moderate samples, Churchill and Doerge (1994) described how to perform a permutation test in QTL mapping. The permutation procedure can be extended to BTL mapping by permuting the binary data along with the associated covariates.

Finally, note that the present research is for detecting only one putative BTL on a chromosome using two flanking markers. One disadvantage of this method, similar to that in Lander and Botstein's (1989) IM method, is that if there is more than one BTL on a chromosome, the estimate of a BTL effect and the location in a tested interval is likely to be affected by the BTLs at other locations. The main purpose of this article is to provide a basic method in genetic locus mapping for binary traits and it is hoped that future research can extend it for testing multiple BTL simultaneously. For some recent research on BTL mapping, see, *e.g.*, Yi and Xu (1999a,b, 2000), Kadarmideen *et al.* (2000, 2001), McIntyre *et al.* (2001), and Arechavaleta-Velasco and Hunt (2004).

The authors thank the associate editor J. Bruce Walsh and two reviewers for their constructive suggestions and comments leading to substantial improvements of the manuscript. They also acknowledge helpful discussions with Hong Zhang. This research is supported in part by National Institutes of Health grant EY014478 (Z. L.).

## LITERATURE CITED


Arechavaleta-Velasco, M. E., and G. J. Hunt, 2004 Binary trait loci that influence honey bee (Hymenoptera: Apidae) guarding behavior. Ann. Entomol. Soc. Am. **97:** 177–183.

Bonney, G. E., 1986 Regressive logistic models for familial disease and other binary traits. Biometrics **42:** 611–625.

Broman, K. W., 2003 Mapping quantitative trait loci in the case of a spike in the phenotype distribution. Genetics **163:** 1169–1175.

Chen, H., and J. Chen, 2001 The likelihood ratio test for homogeneity in finite mixture models. Can. J. Stat. **29:** 201–215.

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

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 967–971.

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

Follmann, D. A., and D. Lambert, 1991 Identifiability of finite mixtures of logistic regression models. J. Stat. Plan. Inference **27:** 375–381.

Hackett, C. A., and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics **51:** 1252–1263.

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

Hosmer, D. W., and S. Lemeshow, 1989 *Applied Logistic Regression.* John Wiley & Sons, New York.

Kadarmideen, H. N., L. L. G. Janss and J. C. M. Dekkers, 2000 Power of quantitative trait locus mapping for polygenic binary traits using generalized and regression interval mapping in multi-family half-sib designs. Genet. Res. **76:** 305–317.

Kadarmideen, H. N., L. L. G. Janss and J. C. M. Dekkers, 2001 Generalized marker regression and interval QTL mapping methods for binary traits in half-sib family designs. J. Anim. Breed. Genet. **118:** 297–309.

Kao, C. H., and Z-B. Zeng, 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, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lange, C., and J. C. Whittaker, 2001 Mapping quantitative trait loci using generalized estimating equations. Genetics **159:** 1325–1337.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

McIntyre, L. M., C. J. Coffman and R. W. Doerge, 2001 Detection and localization of a single binary trait locus in experimental populations. Genet. Res. **78:** 79–92.

Sax, K., 1923 The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris.* Genetics **8:** 552–560.

Sillanpaa, M. J., and M. Bhattacharjee, 2005 Bayesian association-based fine mapping in small chromosomal segments. Genetics **169:** 427–439.

Soller, M., and T. Brody, 1976 On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47:** 35–39.

Thomson, P. C., 2003 A generalized estimating equations approach to quantitative trait locus detection of non-normal traits. Genet. Sel. Evol. **35:** 257–280.

Visscher, P. M., C. S. Haley and S. A. Knott, 1996 Mapping QTLs for binary traits in backcross and F-2 populations. Genet. Res. **68:** 55–63.

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

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

Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary disease using line crosses. Genetics **143:** 1417–1424.

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

Yi, N. J., and S. Z. Xu, 1999b A random model approach to mapping quantitative trait loci for complex binary traits in outbred populations. Genetics **2:** 1029–1040.

Yi, N. J., and S. Z. Xu, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155:** 1391–1403.

Zeng, Z-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.


Communicating editor: J. B. Walsh

## APPENDIX

**Assumptions:** All the asymptotic results assume that (i) the parametric space of the logistic regression model parameters is a compact super-rectangular in $R^{2+p}$ for the backcross population and in $R^{3+p}$ for the intercross and $r \in [0, \gamma]$ with $\gamma < \frac{1}{2}$ being specific. On the covariate $x$, assume that $E\{xx^\tau\} > 0$ is positive definite and the distribution $g$ of $x$ satisfies $E\{|\log g(x)|\} < \infty$.

**Null limiting distribution of $T_n$:** Suppose that under the null hypothesis, the true parameter is $\theta_0 = (\gamma, \beta_0, 0, \lambda_0)$; then, the probability mass function of the trait value $Y$ is

$$\pi_0(y|x) = \exp\{(\beta_0 + \lambda_0^\tau x)y\}/\{1 + \exp\{\beta_0 + \lambda_0^\tau x\}\}.$$

Rewrite

$$T_n = 2[l_n(\hat{\theta}) - l_n(\theta_0)] - 2[l_n(\hat{\theta}_0) - l_n(\theta_0)] \equiv T_{n1} - T_{n2}.$$

Let

$$R_n(\theta) = 2\{l_n(\theta) - l_n(\theta_0)\} = 2\sum_{i=1}^{n} \log\left\{p_r(j_i)\frac{\pi(y_i \mid x_i, \beta, \lambda)}{\pi_0(y_i \mid x_i)} + [1 - p_r(j_i)]\frac{\pi(y_i \mid x_i, \beta + \alpha, \lambda)}{\pi_0(y_i \mid x_i)}\right\} = 2\sum_{i=1}^{n} \log\{1 + \delta_i(\theta)\},$$

where

$$\delta_i(\theta) = p_r(j_i)\left\{\frac{\pi(y_i \mid x_i, \beta, \lambda)}{\pi_0(y_i \mid x_i)} - 1\right\} + [1 - p_r(j_i)]\left\{\frac{\pi(y_i \mid x_i, \beta + \alpha, \lambda)}{\pi_0(y_i \mid x_i)} - 1\right\}.$$

Recall that the MLEs of $\beta$, $\alpha$, and $\lambda$ are consistent. So for $r$ fixed, consider the Taylor expansion of $\delta_i(\theta)$ at $\theta = \theta_0$. Rearranging the leading terms, we have

$$\delta_i(\theta) = (\lambda - \lambda_0)^\tau x_i W_i + (\beta - \beta_0 + \alpha/2) W_i - \alpha[p_r(j_i) - \tfrac{1}{2}] W_i + \epsilon_i(\theta), \tag{A1}$$

where $W_i = Y_i - 1 + (1/\exp\{\beta_0 + \lambda_0^\tau x_i\})$ and $\epsilon_i(\theta)$ is the remainder.

By the zero-mean conditions

$$E((\lambda - \lambda_0)^\tau x_i W_i) = E(W_i) = E\{[p_r(j_i) - \tfrac{1}{2}] W_i\} = 0$$

and

$$E\{(\lambda - \lambda_0)^\tau x_i W_i [p_r(j_i) - \tfrac{1}{2}] W_i\} = E\{W_i [p_r(j_i) - \tfrac{1}{2}] W_i\} = 0,$$

it can be shown (see Chen and Chen 2005 for technical details in a similar situation) that

$$T_{n1} = R_n(\hat{\theta}) = Q_n + \sup_{0 \le r \le \gamma} \frac{[\sum_{i=1}^n (p_r(j_i) - \frac{1}{2}) W_i]^2}{\sum_{i=1}^n (p_r(j_i) - \frac{1}{2})^2 W_i^2} + o_p(1),$$

where

$$Q_n = \left( \sum_{i=1}^n W_i x_i^\tau, \sum_{i=1}^n W_i \right) \begin{pmatrix} \sum_{i=1}^n W_i^2 x_i x_i^\tau & \sum_{i=1}^n x_i W_i^2 \\ \sum_{i=1}^n W_i^2 x_i^\tau & \sum_{i=1}^n W_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n x_i W_i \\ \sum_{i=1}^n W_i \end{pmatrix} \tag{A2}$$

and $o_p(1)$ means convergence to zero in probability. A standard analysis for $T_{n2}$ yields

$$T_{n2} = R_n(\hat{\theta}_0) = Q_n + o_p(1).$$

Hence,

$$T_n = T_{n1} - T_{n2} = \sup_{0 \le r \le \gamma} \frac{[\sum_{i=1}^n (p_r(j_i) - \frac{1}{2}) W_i]^2}{\sum_{i=1}^n (p_r(j_i) - \frac{1}{2})^2 W_i^2} + o_p(1).$$

Uniformly in $r$,

$$n^{-1} \sum_{i=1}^n (p_r(j_i) - \tfrac{1}{2})^2 W_i^2 \to E(W_1^2) \tau_r^2$$

almost surely, and as $n \to \infty$,

$$n^{-1/2} \sum_{i=1}^n (p_r(j_i) - \tfrac{1}{2}) W_i \to \sqrt{E(W_1^2)} \sum_{j=1}^4 \sqrt{q_{bc}(j)} \{p_r(j) - \tfrac{1}{2}\} Z_j$$

in distribution. Therefore, by Slutsky's convergence theorem, the asymptotic distribution of $T_n$ is given by (7).

**Null limiting distribution of $S_n$:** The proof is similar to the case of $T_n$, but just adjust the orthogonal decomposition of $\delta_i$ in (A1) as follows:

$$\begin{aligned} \delta_i(\eta) = {} & (\lambda - \lambda_0)^\tau x_i W_i + \{2(\beta - \beta_0) + (\alpha/4) + 3(\nu/2)\} W_i \\ & + \{-\nu + (\tau_{r,12}/\tau_{r,1}^2)(\alpha - \nu)\} a_r(j_i) W_i \\ & + (\alpha - \nu) b_r(j_i) W_i. \end{aligned}$$

Note that the first two terms contribute to $Q_n$ defined in (A2) that is canceled out with the leading term of $T_{n2}$ and the last terms are orthogonal to the first two and orthogonal to each other. And finally, since

$$n^{-1/2} \sum_{i=1}^n a_r(j_i) W_i \to \sqrt{E(W_1^2)} \sum_{j=1}^9 \sqrt{q_{ic}(j)}\, a_r(j) Z_j$$

and

$$n^{-1/2}\sum_{i=1}^{n} b_{\mathrm{r}}(j_i)W_i \to \sqrt{E(W_1^2)}\sum_{j=1}^{9}\sqrt{q_{\mathrm{ic}}(j)}\,b_{\mathrm{r}}(j)Z_j$$

uniformly in $r$, the limiting distribution of $S_{\mathrm{n}}$ is implied and given by (10).

**EM algorithm for obtaining MLEs:** The EM algorithm with the backcross population is given here, while the EM algorithm with the intercross is similar and so omitted. Fix $r$. Let $u_i$ be the genotype of the BTL ($u_i = 1$ if $BB$ and $= 0$ if $Bb$). Since $u_i$ is unobservable, it is treated as a missing value. Denote

$$\pi_{i1} = \pi(1|x_i, \beta, \lambda) \quad \text{and} \quad \pi_{i2} = \pi(1|x_i, \beta + \alpha, \lambda).$$

Then the log-likelihood function with the complete data is

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^{n} \log\{[\pi_{i1}^{y_i}(1-\pi_{i1})^{1-y_i}]^{u_i}[\pi_{i2}^{y_i}(1-\pi_{i2})^{1-y_i}]^{1-u_i}\} \\ &= \sum_{i=1}^{n}\{u_i \log\{\pi_{i1}^{y_i}(1-\pi_{i1})^{1-y_i}\} + (1-u_i)\log\{\pi_{i2}^{y_i}(1-\pi_{i2})^{1-y_i}\}\}. \end{aligned}$$

Thus in the E-step of the $(t+1)$th EM iteration, we need to calculate only

$$\begin{aligned} w_i^{(t)} &= E[u_i \,|\, y_i;\, x_i,\, j_i,\, \theta^{(t)}(r)] = P[u_i = 1 \mid y_i;\, x_i,\, j_i,\, \theta^{(t)}(r)] \\ &= \frac{p_{\mathrm{r}}(j_i)[\pi_{i1}^{(t)}]^{y_i}[1-\pi_{i1}^{(t)}]^{1-y_i}}{p_{\mathrm{r}}(j_i)[\pi_{i1}^{(t)}]^{y_i}[1-\pi_{i1}^{(t)}]^{1-y_i} + [1-p_{\mathrm{r}}(j_i)][\pi_{i2}^{(t)}]^{y_i}[1-\pi_{i2}^{(t)}]^{1-y_i}}. \end{aligned}$$

Replace the missing value $u_i$ by $w_i^{(t)}$ in the log-likelihood function with the complete data; then in the M-step, we maximize

$$Q^{(t)} = \sum_{i=1}^{n}\{w_i^{(t)}\log[\pi_{i1}^{y_i}(1-\pi_{i1})^{1-y_i}] + (1-w_i^{(t)})\log[\pi_{i2}^{y_i}(1-\pi_{i2})^{1-y_i}]\}$$

with respect to $\beta$, $\alpha$, and $\lambda$. To do so, we can use the Newton-Raphson iteration method that needs the first and second partial derivatives given below:

$$\begin{aligned} \frac{\partial Q^{(t)}}{\partial \beta} &= \sum_{i=1}^{n}\{w_i^{(t)}(y_i - \pi_{i1}) + (1-w_i^{(t)})(y_i - \pi_{i2})\} \\ \frac{\partial Q^{(t)}}{\partial \alpha} &= \sum_{i=1}^{n}(1-w_i^{(t)})(y_i - \pi_{i2}) \\ \frac{\partial Q^{(t)}}{\partial \lambda} &= \sum_{i=1}^{n} x_i\{w_i^{(t)}(y_i - \pi_{i1}) + (1-w_i^{(t)})(y_i - \pi_{i2})\} \\ \frac{\partial^2 Q^{(t)}}{\partial \beta^2} &= -\sum_{i=1}^{n}\{w_i^{(t)}\pi_{i1}(1-\pi_{i1}) + (1-w_i^{(t)})\pi_{i2}(1-\pi_{i2})\} \\ \frac{\partial^2 Q^{(t)}}{\partial \alpha^2} &= -\sum_{i=1}^{n}(1-w_i^{(t)})\pi_{i2}(1-\pi_{i2}) \\ \frac{\partial^2 Q^{(t)}}{\partial \lambda^2} &= -\sum_{i=1}^{n} x_i^2\{w_i^{(t)}\pi_{i1}(1-\pi_{i1}) + (1-w_i^{(t)})\pi_{i2}(1-\pi_{i2})\} \\ \frac{\partial^2 Q^{(t)}}{\partial \beta \partial \alpha} &= \frac{\partial^2 Q^{(t)}}{\partial \alpha^2} \\ \frac{\partial^2 Q^{(t)}}{\partial \beta \partial \lambda} &= -\sum_{i=1}^{n} w_i^{(t)} x_i \pi_{i1}(1-\pi_{i1}) \\ \frac{\partial^2 Q^{(t)}}{\partial \alpha \partial \lambda} &= -\sum_{i=1}^{n}(1-w_i^{(t)}) x_i \pi_{i2}(1-\pi_{i2}). \end{aligned}$$