DOI: 10.1534/genetics.103.023903

Mapping Quantitative Trait Loci With Censored Observations

Guoqing Diao, D. Y. Lin[1] and Fei Zou

Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599-7420

Manuscript received October 31, 2003
Accepted for publication June 2, 2004

ABSTRACT

The existing statistical methods for mapping quantitative trait loci (QTL) assume that the phenotype follows a normal distribution and is fully observed. These assumptions may not be satisfied when the phenotype pertains to the survival time or failure time, which has a skewed distribution and is usually subject to censoring due to random loss of follow-up or limited duration of the experiment. In this article, we propose an interval-mapping approach for censored failure time phenotypes. We formulate the effects of QTL on the failure time through parametric proportional hazards models and develop efficient likelihood-based inference procedures. In addition, we show how to assess genome-wide statistical significance. The performance of the proposed methods is evaluated through extensive simulation studies. An application to a mouse cross is provided.

QUANTITATIVE trait analysis plays an important role in the understanding of genetic variations in plants and animals. Mapping quantitative trait loci (QTL) can lead to improvements in economic traits, such as yield and quality in crop plants and milk production in cows. QTL mapping in animals can also provide valuable insights into the genetic etiologies of complex human diseases (Hilbert *et al.* 1991; Jacob *et al.* 1991; Shepel *et al.* 1998).

Much of the modern statistical methodology for QTL mapping in experimental crosses originates from the seminal work of Lander and Botstein (1989). The Lander-Botstein interval-mapping method postulates that QTL occur at a series of positions within a set of adjacent marker intervals and that the trait value depends on the QTL genotype through a linear regression model. The distance between each pair of genetic markers is assumed known. The method steps through the genome in specified increments, say every 1 or 2 cM, and calculates the likelihood-ratio statistic for testing no QTL present at each position. The position with the largest value of the likelihood-ratio statistic is declared to be the candidate QTL location provided that the value exceeds a certain threshold level. It is widely recognized that the interval-mapping method has higher power and requires fewer progenies than the single-marker analysis (Lander and Botstein 1989; Haley and Knott 1992; Zeng 1994). Doerge *et al.* (1997) described in greater detail this method and various extensions.

Most of the existing QTL-mapping methods require that the phenotype be normally distributed and fully observed. These assumptions are likely to be false when the phenotype pertains to the survival time or failure time. The Weibull distribution and other skewed distributions with long right tails are more appropriate than the normal distribution. Furthermore, the failure time is often subject to censoring so that the trait value is known only to be beyond the censoring time. An example of failure time in plant experiments is the flowering time, which may be censored due to limited duration of the experiment; see Ferreira *et al.* (1995). In animal studies, the failure times of interest include time to tumor and time to death (*i.e.*, survival time), which may be subject to censoring because of limited study duration or death due to unrelated causes. One particular example is a mice cross presented by Broman (2003), in which the trait of interest is time to death after a bacterial infection and in which 30% of the mice are still alive at the end of the study period. Symons *et al.* (2002) presented another interesting study, in which the phenotype is the time until terminal illness due to tumor for Eμ-v-abl transgenic mice.

The incompleteness of the trait values presents major challenges in the application of the interval-mapping approach. Broman (2003) considered a cure model in which the mice that are alive at the end of the study are regarded as cured and in which the survival times among the deaths follow a log-normal distribution. This is a specialized model, which can deal only with the situations in which the potential censoring times are equal among all study subjects. Symons *et al.* (2002) utilized a variant of the expectation-maximization (EM) algorithm (Lipsitz and Ibrahim 1998) to map QTL with censored observations. This method is computationally intensive and its properties have not been investigated.

In this article, we provide a simple and rigorous exten-

[1] *Corresponding author:* Department of Biostatistics, University of North Carolina, 3101E McGavran-Greenberg Hall, CB 7420, Chapel Hill, NC 27599-7420. E-mail: lin@bios.unc.edu

sion of the interval-mapping approach of Lander and Botstein (1989) to censored quantitative traits. Specifically, we formulate the effects of QTL on the failure time through proportional hazards models (Kalbfleish and Prentice 2002, Sect. 2.3). We then develop efficient likelihood-based methods for locating QTL and estimating the effects of QTL. In addition, we show how to assess genome-wide statistical significance by extending the analytical results of Lander and Botstein (1989) and Dupuis and Siegmund (1999) and by developing an accurate and efficient Monte Carlo procedure. We conduct extensive simulation studies to evaluate the performance of the proposed methods. Finally, we provide an application to the mice data of Broman (2003).

METHODS

Interval mapping: In this section, we develop an interval-mapping method for potentially censored failure-time traits in an F_2 intercross population by modeling a single QTL. Expanding the results to other crosses is straightforward. Consider n progenies from an intercross between two inbred strains. Let T_i denote the quantitative trait for the ith subject, which pertains to a failure time that can potentially be censored and thus incompletely observed. Let C_i be the censoring time for the ith subject. The observation on the trait value of the ith subject consists of two components: $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, where $I(\mathscr{A})$ is the indicator function for event $\mathscr{A}$. The failure time T_i is fully observed only when it is uncensored, *i.e.*, $\Delta_i = 1$.

Suppose that we have data on a set of genetic markers with a known genetic map. Let $\mathbf{M}_i$ denote the multipoint marker genotype data for the ith subject. We consider a putative QTL locus d in the genome with two possible alleles q and Q from the two inbred parents and define $G_i = -1$, 0, or 1 according to whether the ith subject has genotype qq, Qq, or QQ, respectively, at the QTL. We specify a proportional hazards model for the effects of the QTL genotype on the failure time such that, conditional on the QTL genotype G_i, the hazard function of T_i takes the form

$$\lambda(t|G_i) = \lambda_0(t)e^{\beta_1 G_i + \beta_2(1-|G_i|)}, \quad i = 1, \ldots, n, \qquad (1)$$

where β_1 and β_2 pertain to the additive and dominant effects of the QTL, and $\lambda_0(\cdot)$ is an unknown baseline hazard function (Kalbfleish and Prentice 2002, Sect. 2.3). In this article, we assume a parametric model for λ_0. In particular, we consider a Weibull hazard function $\lambda_0(t) = \gamma_1\gamma_2 t^{\gamma_2 - 1}$, $\gamma_1 > 0$, $\gamma_2 > 0$ (Kalbfleish and Prentice 2002, p. 33).

Write $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$. At each locus, we may calculate $\pi_{i,g} = \Pr(G_i = g|\mathbf{M}_i)$ ($g = -1, 0, 1$; $i = 1, \ldots, n$), which are the conditional probabilities of the QTL genotypes given the observed marker data. Under the assumptions of no crossover interference and no genotyping errors, these probabilities are determined by the genotypes of the two flanking markers and the location of the QTL; see Equation 15.2 of Lynch and Walsh (1998).

We assume that censoring is noninformative (Kalbfleish and Prentice 2002, p. 195) and that the censoring time is independent of the failure time and QTL genotype. The likelihood for the vector of parameters $\boldsymbol{\theta}$ based on the complete data $(Y_i, \Delta_i, G_i, \mathbf{M}_i)$ $(i = 1, \ldots, n)$ is proportional to

$$\prod_{i=1}^{n} \prod_{g=-1,0,1} \left\{\lambda^{\Delta_i}(Y_i|g)\, e^{-\int_0^{Y_i}\lambda(t|g)\,dt}\pi_{i,g}\right\}^{I(G_i=g)}, \qquad (2)$$

while the likelihood based on the observed data $(Y_i, \Delta_i, \mathbf{M}_i)$ $(i = 1, \ldots, n)$ is proportional to

$$\prod_{i=1}^{n} \sum_{g=-1,0,1} \lambda^{\Delta_i}(Y_i|g)\, e^{-\int_0^{Y_i}\lambda(t|g)\,dt}\pi_{i,g}. \qquad (3)$$

To obtain the maximum-likelihood estimator (MLE) of $\boldsymbol{\theta}$, we may maximize the observed-data likelihood (3) directly. An alternative approach is to apply the EM algorithm (Dempster *et al.* 1977) to (2). The expected value of the complete-data log-likelihood given the observed data can be shown to be, up to a constant,

$$\sum_{i=1}^{n} \sum_{g=-1,0,1} p_{i,g}(\theta)\left\{\Delta_i \ln \lambda(Y_i|g) - \int_0^{Y_i}\lambda(t|g)\,dt\right\}, \qquad (4)$$

where

$$p_{i,g}(\boldsymbol{\theta}) = \frac{p^{\dagger}_{i,g}(\boldsymbol{\theta})}{\sum_{v=-1,0,1} p^{\dagger}_{i,v}(\boldsymbol{\theta})}, \quad g = -1, 0, 1;\ i = 1, \ldots, n,$$

and

$$p^{\dagger}_{i,g}(\boldsymbol{\theta}) = \pi_{i,g}\exp\left\{\Delta_i(\beta_1 g + \beta_2(1-|g|)) - \int_0^{Y_i}\lambda(t|g)\,dt\right\}.$$

In the E-step, we evaluate $p_{i,g}(\boldsymbol{\theta})$ at the current estimate of $\boldsymbol{\theta}$. The M-step can proceed in a similar manner to the case of complete data since expression (4), with $\boldsymbol{\theta}$ in $p_{i,g}(\boldsymbol{\theta})$ fixed, takes the same form as the complete-data log-likelihood. We begin the EM algorithm by assigning an initial value to $\boldsymbol{\theta}$ and iterate until convergence. The initial value for $\boldsymbol{\beta}$ is set to $\mathbf{0}$ and that of $\boldsymbol{\gamma}$ to some value in the parameter space of $\boldsymbol{\gamma}$. The resulting MLE is denoted by $\hat{\boldsymbol{\theta}}$. See appendix a for further detail.

We test the null hypothesis of no QTL effects, *i.e.*, H_0: $\boldsymbol{\beta} = \mathbf{0}$, by the likelihood-ratio statistic

$$\mathrm{LR} = 2\ln\frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})},$$

where $L(\cdot)$ is the observed-data likelihood, and $\tilde{\boldsymbol{\theta}} = (\mathbf{0}, \tilde{\boldsymbol{\gamma}})$ with $\tilde{\boldsymbol{\gamma}}$ being the restricted MLE of $\boldsymbol{\gamma}$ under H_0. The LOD score is LR/(2 ln 10). Under H_0, LR is asymptotically χ^2-distributed with 2 d.f. (appendix b). Note that $p_{i,g}(\boldsymbol{\theta})$, $\hat{\boldsymbol{\theta}}$, $L(\hat{\boldsymbol{\theta}})$, LR, and LOD all depend on the locus d through the dependence of $\pi_{i,g}$ on d. In the sequel, we include d in the expressions to emphasize their dependence on d if ambiguity arises. Note also that $\tilde{\boldsymbol{\theta}}$ and $L(\tilde{\boldsymbol{\theta}})$ do not depend on d. Thus, as in the case of standard interval mapping, the likelihood under H_0 is calculated once while the likelihood under the alternative is evaluated at each location in the genome to produce a LOD curve for each chromosome. The position with the largest value of the LOD score is declared to be the QTL location provided that the value exceeds a certain threshold level. We show how to determine the threshold level in the following section.

Thresholds: When searching the entire chromosome or whole genome for QTL, one should select a threshold level for the LOD score such that the probability (under the null hypothesis) that LOD or some other test statistic exceeds this level anywhere in the genome equals the desired false-positive rate. The pointwise significance level based on the χ^2-approximation is inadequate because of the multiple tests while the Bonferroni correction is too conservative because of the dependence of the test statistics at different points in the genome. In APPENDIX C, we show that the likelihood-ratio statistic $\mathrm{LR}(d)$ can be partitioned into the sum of the squares of two asymptotically independent Ornstein-Uhlenbeck processes. This result is analogous to those of LANDER and BOTSTEIN (1989) and DUPUIS and SIEGMUND (1999) and implies that the analytical approximations of thresholds for normal traits can be applied to the case of censored failure time observations. These analytical results assume that the markers are dense or equally spaced with no missing data and thus may not work well in practice. Using results of DAVIES (1977, 1987), REBAI *et al.* (1994) provided approximate thresholds in backcross (BC) and F_2, which are applicable in the intermediate map density case. The calculations are formidable, even for F_2, and do not accommodate missing marker data.

To overcome the limitations of the analytical approximations, we propose a novel resampling approach to determining the thresholds for genome-wide statistical significance. This approach allows arbitrary distributions of the markers as well as arbitrary test positions. It also accommodates missing marker data and dominant markers.

For evaluating the correlations of the test statistics among different locations, it is more convenient to work with the score statistic than the likelihood-ratio statistic. Let $\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}}; d)$ be the score function (based on the observed-data likelihood) for $\boldsymbol{\beta}$ at location d, which can be approximated by the sum of n independent zero-mean random vectors $\Sigma_{i=1}^{n}\tilde{\mathbf{U}}_i(\boldsymbol{\theta}_0; d)$, where $\boldsymbol{\theta}_0 = (\mathbf{0}, \boldsymbol{\gamma}_0)$ is the true parameter value under H_0; see APPENDIX B. Thus, $n^{-1/2}\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}}; d)$ is asymptotically zero-mean normal with covariance matrix $\mathbf{V}(d)$ that is the limit of n^{-1} $\Sigma_{i=1}^{n}\tilde{\mathbf{U}}_i(\boldsymbol{\theta}_0; d)\tilde{\mathbf{U}}_i^T(\boldsymbol{\theta}_0; d)$. We replace the unknown quantities in $\tilde{\mathbf{U}}_i(\boldsymbol{\theta}_0; d)$ by their sample estimators to yield $\hat{\mathbf{U}}_i(d)$ shown in APPENDIX B. The score statistic for testing H_0: $\boldsymbol{\beta} = \mathbf{0}$ against H_1: $\boldsymbol{\beta} \neq \mathbf{0}$ takes the form

$$W(d) = n^{-1}\mathbf{U}_\beta^T(\tilde{\boldsymbol{\theta}}; d)\hat{\mathbf{V}}^{-1}(d)\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}}; d),$$

where $\hat{\mathbf{V}}(d) = n^{-1}\Sigma_{i=1}^{n}\hat{\mathbf{U}}_i(d)\hat{\mathbf{U}}_i^T(d)$ is a consistent estimator of the limiting covariance matrix $\mathbf{V}(d)$. It can be shown that $W(d)$ is asymptotically equivalent to $\mathrm{LR}(d)$ (COX and HINKLEY 1974, Sect. 9.3).

In general, the limiting distribution of $\sup_d W(d)$ is not analytically tractable. We use a resampling approach similar to that of LIN *et al.* (1993) to approximate the null distributions of $n^{-1/2}\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}}; d)$ and $\sup_d W(d)$. From APPENDIX B, it is easy to see that $n^{-1/2}\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}}; d)$ converges to a zero-mean Gaussian process with covariance function $\boldsymbol{\psi}(d_1, d_2)$ that is the limit of $n^{-1}\Sigma_{i=1}^{n}\tilde{\mathbf{U}}_i(\boldsymbol{\theta}; d_1)\tilde{\mathbf{U}}_i^T(\boldsymbol{\theta}; d_2)$ at (d_1, d_2) and that $\boldsymbol{\psi}(d_1, d_2)$ can be consistently estimated by $n^{-1}\Sigma_{i=1}^{n}\hat{\mathbf{U}}_i(d_1)\hat{\mathbf{U}}_i^T(d_2)$. Define $\hat{\mathbf{U}}(d) = \Sigma_{i=1}^{n}$ $\hat{\mathbf{U}}_i(d)Z_i$, where Z_i ($i = 1, \ldots, n$) are independent standard normal random variables that are independent of the observed data. Conditional on the observed data, $n^{-1/2}\hat{\mathbf{U}}(d)$ is a Gaussian process with mean $\mathbf{0}$ and covariance function $n^{-1}\Sigma_{i=1}^{n}\hat{\mathbf{U}}_i(d_1)\hat{\mathbf{U}}_i^T(d_2)$ at (d_1, d_2), which converges to $\boldsymbol{\psi}(d_1, d_2)$. Thus, the conditional distribution of $n^{-1/2}\hat{\mathbf{U}}(d)$ given the observed data converges to the limiting distribution of $n^{-1/2}\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}}; d)$. As a result, the distribution of $W(d)$ can be approximated by that of

$$\hat{W}(d) = n^{-1}\hat{\mathbf{U}}^T(d)\hat{\mathbf{V}}^{-1}(d)\hat{\mathbf{U}}(d).$$

To approximate the distribution of $\sup_d W(d)$, we generate the normal random sample $(Z_1, \ldots, Z_n)$ a large number of times while holding the observed data fixed; for each sample, we calculate $\hat{W}(d)$ and $\sup_d \hat{W}(d)$. The $100(1 - \alpha)$th percentile of the simulated $\sup_d \hat{W}(d)$ is the threshold value for the genome-wide significance level of α. This resampling approach is computationally much more efficient than the use of permutation (CHURCHILL and DOERGE 1994) and other simulation methods because it involves only simulation of normal random variables and does not entail repeated analysis of simulated data sets.

SIMULATIONS

To investigate the operating characteristics of the proposed methods in practical situations, we performed extensive simulation studies. We generated the failure times from the Weibull distributions with baseline hazard function $\lambda_0(t) = \gamma_1\gamma_2 t^{\gamma_2 - 1}$, where $\gamma_1 = 0.01$ and $\gamma_2 = 2$. We reparameterized γ_1 and γ_2 according to $\gamma_k = e^{\alpha_k}$ ($k = 1, 2$) so as to ensure that the estimates of γ_1 and γ_2 are positive. The censoring times were generated from the uniform $(0, \tau)$ distribution, where τ was chosen to yield ~30% censored observations. Assuming no crossover interference, we generated the marker data from the Markov chain. The interval-mapping step size was set at 1 cM.

We considered a chromosome with a total length of 100 cM. Genetic maps with different numbers of equally spaced markers were simulated. Under H_1, one QTL

TABLE 1

Summary statistics for the estimator of the additive QTL effect at the true QTL location

No. of	H_0: $\beta_1 = 0$, $\beta_2 = 0$				H_1: $\beta_1 = 0.35$, $\beta_2 = 0.30$			
markers	Mean[a]	SE[b]	SEE[c]	CP[d]	Mean[a]	SE[b]	SEE[c]	CP[d]
6	0.002	0.108	0.107	94.8	0.355	0.110	0.108	94.4
11	0.002	0.105	0.104	94.6	0.355	0.107	0.105	94.8
51	−0.001	0.101	0.099	94.4	0.353	0.103	0.101	94.6
101	0.000	0.099	0.098	94.9	0.354	0.100	0.100	94.8

[a] Mean is the mean of the parameter estimator.
[b] SE is the standard error of the parameter estimator.
[c] SEE is the mean of the standard error estimator.
[d] CP is the coverage probability of the 95% confidence interval.

located at 35 cM was simulated with $\beta_1 = 0.35$ and $\beta_2 = 0.30$. We generated 10,000 replicates of 300 observations from an F_2 population. We evaluated the finite-sample properties of the MLEs of the QTL effects at the true QTL location. The results for the estimator of the additive QTL effect are summarized in Table 1. The proposed estimator appears to be virtually unbiased. The standard error estimator reflects accurately the true variation. The confidence intervals have proper coverage probabilities. We obtained similar results for the estimator of the dominant QTL effect (data not shown). We also examined the performance of the proposed interval-mapping methods for locating the QTL and estimating the QTL effects at the location with the maximum LOD. The results are summarized in Table 2. There is little bias for the estimator of the QTL location or the estimators of the QTL effects. The mapping is more precise for denser marker maps.

We conducted additional simulation studies to evaluate the performance of the analytical and resampling methods for determining genome-wide statistical significance. We generated both equally and unequally spaced markers. We also simulated data with missing marker genotypes and dominant markers, which are more comparable with real data. We considered one chromosome with a total length of 100 cM. For the cases of unevenly spaced markers, we placed m markers at the following locations,

$$\mathrm{LOC}_j = \begin{cases} 50(j-1)/(m-1), & j = 1, \ldots, [m/2], \\ 100(j-1)/(m-1), & \text{otherwise}, \end{cases}$$

where LOC_j is the jth marker location and $[m/2]$ is the largest integer that is less than or equal to $m/2$. In these settings, the first half of the markers is denser than the second half of the markers. We generated 10,000 replicates of 300 observations from an F_2 population. The dense-map and sparse-map approximations were obtained from Equation C1 in APPENDIX C . The thresholds for the resampling method were based on 10,000 normal samples. The results are summarized in Tables 3 and 4.

The thresholds based on the resampling method are close to the empirical values, whether the data are generated under H_0 or H_1; consequently, the LR tests based on these thresholds have proper type I error and power. This is true of any genetic map, with or without missing marker genotypes and dominant markers. The dense-map approximations are too conservative and thus result in power loss, while the sparse-map approximations tend to be too liberal. We also assessed the approximations by Rebai *et al.* (1994), which turn out to be conservative when the genetic map is dense. For example, in

TABLE 2

Sampling means of the estimators for the QTL location and for the QTL effects at the estimated QTL location in the simulation studies

	H_0: $\beta_1 = 0$, $\beta_2 = 0$			H_1: $\beta_1 = 0.35$, $\beta_2 = 0.30$		
		QTL effects			QTL effects	
No. of markers	QTL location (cM)	β_1	β_2	QTL location (cM)	β_1	β_2
6	49.6 (34.2)	0.001 (0.141)	−0.012 (0.235)	36.6 (11.3)	0.359 (0.107)	0.304 (0.173)
11	49.9 (32.8)	0.000 (0.143)	−0.007 (0.233)	35.8 (10.9)	0.359 (0.101)	0.300 (0.162)
51	50.2 (31.8)	−0.001 (0.148)	−0.008 (0.251)	35.7 (9.3)	0.362 (0.098)	0.314 (0.154)
101	50.3 (31.4)	0.001 (0.150)	−0.009 (0.255)	35.5 (8.9)	0.365 (0.100)	0.319 (0.152)

Standard errors are shown in parentheses.

TABLE 3

Analytical and resampling-based thresholds at the targeted genome-wide significance level of α

				Resampling[c]				Analytical			
		Empirical[b]		H_0		H_1[d]		Dense-map		Sparse-map	
No. of markers	Marker pattern[a]	α = 5%	1%	α = 5%	1%	α = 5%	1%	α = 5%	1%	α = 5%	1%
6	1	10.05	13.57	9.80 (0.16)	13.34 (0.25)	9.81 (0.16)	13.36 (0.25)	13.37	17.12	9.15	12.39
11	1	10.33	13.90	10.41 (0.17)	13.99 (0.25)	10.42 (0.16)	14.00 (0.25)	13.37	17.12	10.15	13.53
51	1	11.79	15.60	11.47 (0.19)	15.12 (0.27)	11.47 (0.18)	15.12 (0.27)	13.37	17.12	11.80	15.37
6	2	9.94	13.65	9.63 (0.16)	13.18 (0.25)	9.64 (0.17)	13.18 (0.25)	13.37	17.12	9.15	12.39
11	2	10.65	14.27	10.23 (0.17)	13.80 (0.25)	10.24 (0.17)	13.81 (0.26)	13.37	17.12	10.15	13.53
51	2	11.54	15.08	11.17 (0.19)	14.80 (0.27)	11.17 (0.18)	14.81 (0.26)	13.37	17.12	11.80	15.37

[a] Under pattern 1, markers are evenly spaced with no missing marker genotypes or dominant markers; under pattern 2, markers are unevenly spaced with 20% missing marker genotypes and 5% dominant markers.

[b] Percentiles of the test statistic based on 10,000 simulated data sets.

[c] Average thresholds from 10,000 simulated data sets. The values in parentheses are the standard errors of the thresholds.

[d] QTL is located at 35 cM with $\beta_1 = 0.35$ and $\beta_2 = 0.3$.

the case of 51 markers with $\alpha = 0.05$, the sizes are 2.66 and 2.26% for marker patterns 1 and 2, respectively.

APPLICATION

To illustrate our methods, we consider the mice data previously analyzed by Broman (2003). A total of 116 female mice from an intercross between the BALB/cByJ and C57BL/6ByJ strains were genotyped at 133 markers, including 2 on the X chromosome. The phenotype of interest is the time to death following infection with *Listeria monocytogenes*. Approximately 30% of the survival times are censored.

Broman (2003) proposed a nonparametric (NP) approach and a two-part model. The NP approach is an extension of the Kruskal-Wallis statistic (Lehmann 1975, Sect. 5.2) by assigning a prior weight ($\pi_{i,g}$) to the rank of the *i*th observation for each QTL genotype group *g*. In this approach, the censored observations are treated as the true failure times and an average rank is assigned to those observations. In the two-part approach, Broman considered a cure model in which the mice that are alive at the end of the study are regarded as cured while the survival times among the deaths follow a log-normal distribution.

We applied the proposed methods to these data, assuming a Weibull baseline hazard, and the results are shown in Table 5 and Figures 1 and 2. The threshold for the LOD score at the 5% genome-wide significance level based on the resampling approach is 3.43, which is close to 3.27, the threshold obtained by permutation for the NP approach of Broman (2003). Our results are fairly consistent with those of Broman (2003). We detect almost the same QTL on chromosomes 5, 13, and 15 except that we detect an additional QTL on chromosome 6 rather than on chromosome 1. The QTL

TABLE 4

Sizes/powers (%) according to the analytical and resampling-based thresholds

						Analytical							
		Resampling				Dense-map				Sparse-map			
		H_0		H_1[b]		H_0		H_1[b]		H_0		H_1[b]	
No. of markers	Marker pattern[a]	α = 5%	1%	α = 5%	1%	α = 5%	1%	α = 5%	1%	α = 5%	1%	α = 5%	1%
6	1	5.6	1.1	91	80	1.1	0.2	79	63	7.5	1.8	93	83
11	1	4.8	1.0	93	82	1.2	0.2	84	69	5.4	1.2	93	83
51	1	5.5	1.1	94	85	2.6	0.6	90	78	5.0	1.1	94	84
6	2	5.6	1.2	77	57	1.1	0.2	56	36	7.1	1.8	80	62
11	2	5.9	1.2	82	65	1.5	0.3	67	47	6.3	1.4	83	66
51	2	5.8	1.2	85	68	2.1	0.5	75	56	4.4	0.9	82	65

[a] Under pattern 1, markers are evenly spaced with no missing marker genotypes or dominant markers; under pattern 2, markers are unevenly spaced with 20% missing marker genotypes and 5% dominant markers.

[b] QTL is located at 35 cM with $\beta_1 = 0.35$ and $\beta_2 = 0.3$.

TABLE 5

Estimates of the QTL positions and QTL effects along with the maximum LOD scores for the data on survival time following infection with *Listeria monocytogenes* in 116 intercross mice

	Proposed method				Nonparametric		Two-part model	
Chromosome	Pos (cM)	LOD	β_1	β_2	Pos (cM)	LOD	Pos (cM)	LOD
1	75	1.94	−0.456	−0.542	76	3.38	81	5.45
5	28	9.01	1.149	0.100	27	5.41	28	6.79
6	59	3.66	0.559	0.563	59	2.45	10	4.09
13	26	6.64	−0.614	−0.740	26	6.71	26	7.38
15	23	4.49	0.370	−0.935	23	3.49	16	4.61

Pos, position.

on chromosome 5 appears to have a strong additive effect and the hazard ratio of the survival time with genotype *QQ vs. qq* is ∼9.95. Genotypes *qq* and *Qq* at the QTL on chromosome 6 seem to have similar effects. The QTL on chromosome 13 appears to have both additive and dominant effects. The QTL on chromosome 15 appears to have a strong dominant effect. At most detected QTL locations, our LOD scores are larger than those of Broman's NP approach. This suggests that our approach may be more efficient in detecting QTL.

Figure 1 shows the LOD curves from the three methods: proposed method, nonparametric method, and two-part model. The LOD scores from the two-part model are larger than those of the other two methods at some QTL locations because there are two more free parameters in the two-part model than in the other two methods. This will decrease the power to detect QTL since a larger threshold (*i.e.*, 4.91) is required. To evaluate different methods on a common scale, we converted the LOD curves to the estimated pointwise *P*-values. Figure 2 displays the values of $-\log_{10}P$ for chromosomes 1, 5, 6, 13, and 15. Comparisons with the nonparametric method and two-part model reveal that the proposed method yields more significant results on the above chromosomes except chromosome 1. Incidentally, the resampling method is ∼100 times faster than the permutation method in this application.

To get some ideas about the adequacy of the Weibull

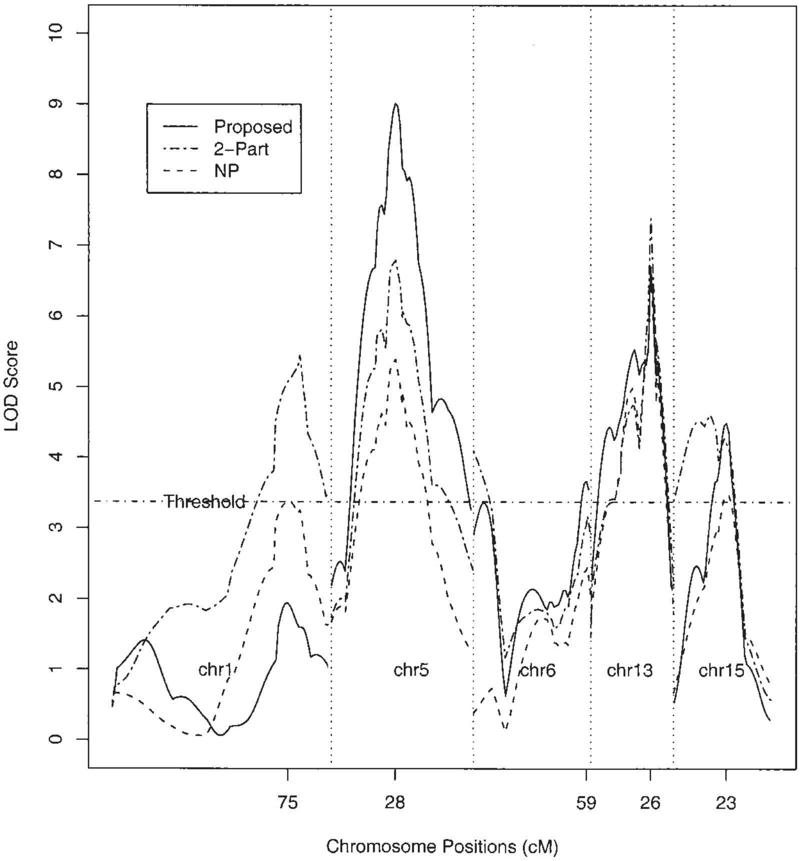

Figure 1.—The LOD scores from three QTL mapping methods for the data on survival time following infection with *Listeria monocytogenes* in 116 intercross mice. The threshold pertains to the 5% genome-wide significance level under the resampling method.

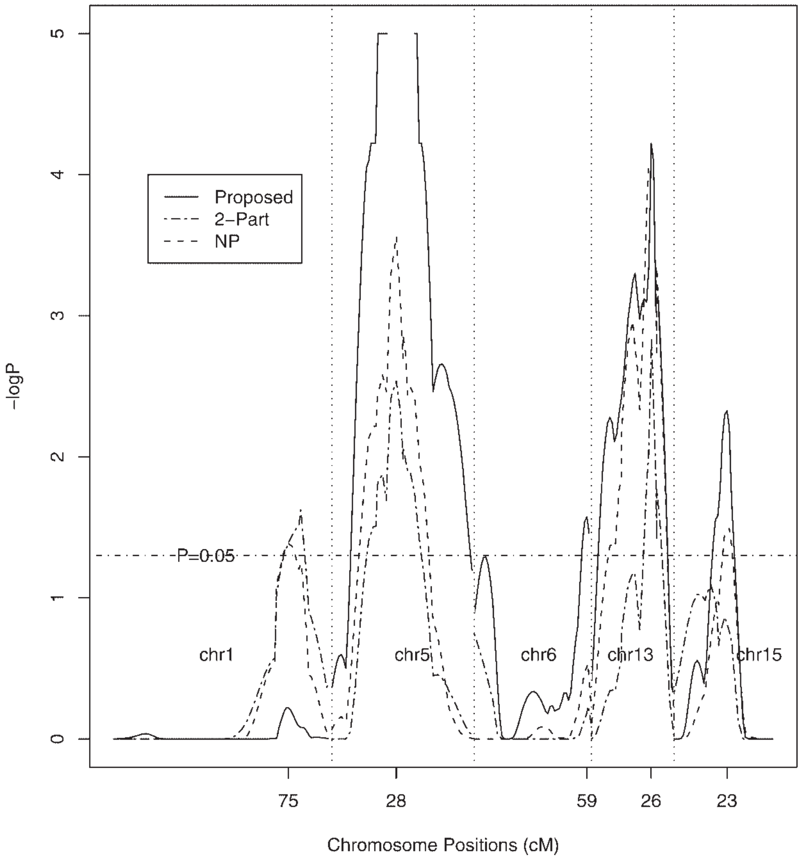

Figure 2.—Plot of the $-\log_{10}P$-values for three QTL mapping methods for the data on survival time following infection with *Listeria monocytogenes* in 116 intercross mice. The P-values for the proposed method are based on 100,000 normal samples. In the region between 26 and 30 cM on chromosome 5, the P-values are $<10^{-5}$ and thus are not displayed. The P-values for the NP and two-part models are based on 11,000 permutation replicates.

distribution in describing the Listeria data, we fitted both the semiparametric model and the Weibull model at marker D5M357, which is close to the peak of the LOD score on chromosome 5. The estimated QTL effects from the two models are very similar. It would be worthwhile to develop formal goodness-of-fit methods for assessing the adequacy of the parametric survival model at the true QTL location.

EXTENSIONS

In this section, we extend the single-QTL model to multiple QTL. The approach of interval mapping (IM) considers one putative QTL at a time. The QTL located elsewhere on the genome can have interfering effects, so that the estimators for the locations and effects of QTL may be biased and the power of detecting QTL may be compromised (Lander and Botstein 1989; Haley and Knott 1992; Zeng 1994). Boer *et al.* (2002) showed that the IM method fails to detect three interacting QTL with no main effects through simulation studies. A variety of approaches have been proposed for mapping multiple QTL. These methods can increase the power to detect QTL and reduce biases in the estimators of the QTL effects and locations. In this section, we consider mainly composite-interval mapping (CIM; Jansen 1993; Zeng 1993, 1994) and multiple-interval mapping (MIM; Kao *et al.* 1999) for censored traits.

Composite-interval mapping: The idea of CIM is to combine IM with multiple regression analysis in mapping QTL by conditioning on markers outside a region of interest to account for the effects of other QTL. To extend the original CIM model to censored traits, we consider the following proportional hazards model,

$$\lambda(t|G_i) = \lambda_0(t)\exp\left\{\beta_1 G_i + \beta_2(1 - |G_i|) + \sum_{k\neq j,j+1} (\beta_{1k}M_{ik} + \beta_{2k}(1 - |M_{ik}|))\right\}, \tag{5}$$

where j and $j + 1$ conform to two flanking markers bordering the putative QTL, and M_{ik} is the indicator variable for the marker genotype, which takes values -1, 0, and 1 for genotypes *aa*, *Aa*, and *AA*, respectively. We may further enhance the model by considering the interaction effects between the putative QTL and controlling markers. Replacing $\lambda(t|G_i)$ in (2) and (3) with (5), we obtain the complete-data and observed-data likelihood functions, respectively. As in the case of standard interval mapping, we can maximize the observed-data likelihood directly or apply the EM algorithm to obtain the MLEs. We can test H_0: $\boldsymbol{\beta} = \mathbf{0}$ at any position in the genome.

The CIM approach requires that the sample size be large relative to the number of markers included in the model. In practice, the sample size is generally not very large. Thus, Zeng (1994) suggested including in the model only those markers that are more or less evenly spaced in the genome or those preidentified markers that explain most of the genetic variation in the genome. This suggestion also applies to our setting.

Multiple-interval mapping: The MIM approach proposed by Kao *et al.* (1999) uses multiple marker intervals simultaneously to fit multiple putative QTL directly in the model. Consider K QTL, $Q_1, \ldots, Q_K$, located at $d_1, \ldots, d_K$ in the genome. There are 3^K possible QTL genotypes. Some of the K QTL may exhibit epistasis. We formulate the effects of the K QTL on the failure time through a proportional hazards model, such that, conditional on the joint genotype $\mathbf{G}_i = (G_{1i}, \ldots, G_{Ki})$, the hazard function of T_i takes the form

$$\lambda(t|\mathbf{G}_i) = \lambda_0(t)\exp\left\{\sum_{j=1}^{K}\boldsymbol{\beta}_j^T\mathbf{x}_{ij} + \sum_{j\neq k}^{K}\delta_{jk}(\mathbf{x}_{ij}^T\mathbf{B}_{jk}\mathbf{x}_{ik})\right\}, \quad (6)$$

where $\mathbf{x}_{ij} = (G_{ij}, 1 - |G_{ij}|)^T$, δ_{jk} is an indicator variable for epistasis between Q_j and Q_k, and $\boldsymbol{\beta}_j$ and $\mathbf{B}_{jk}$ pertain to the main effects and epistatic effects, respectively. The variable δ_{jk} indicates, by the values 1 *vs.* 0, whether or not Q_j and Q_k interact. Given the marker data $\mathbf{M}_i$ for the ith subject and assuming no crossover interference, we may calculate $\pi_{i,g}(g = 1, \ldots, 3^K)$, the conditional probabilities of the 3^K possible genotypes of the K QTL. The complete-data likelihood takes the same form as Equation 2 except that the summation of g is now over $(1, \ldots, 3^K)$. To obtain the MLEs and LOD scores, we can again apply the EM algorithm.

Since the true number and locations of the QTL are unknown, model selection is a critical issue in the MIM approach. Kao *et al.* (1999) suggested stepwise and chunkwise selection with the likelihood-ratio test statistic as a selection criterion to identify QTL, to separate linked QTL, and to analyze epistasis between QTL. Broman and Speed (2002) developed a modified Bayesian information criterion (Schwarz 1978) for model selection. When many QTL are included in the MIM model, the computation can be very intensive. Sen and Churchill (2001) reduced the computation burden of MIM by replacing the EM algorithm with a Monte Carlo algorithm. All these strategies can be applied to our setting.

DISCUSSION

We have described our methods in the context of an F_2 population. All the proposed methods can be easily generalized to other crosses such as BC, F_3, or even more complicated crosses such as combined crosses (Zou *et al.* 2001). We may also extend our methods to accommodate covariates, such as block factors in an agriculture field trial or cage number in a mouse cross.

Symons *et al.* (2002) formulated the effects of the QTL on the failure time through the semiparametric proportional hazards model. They utilized a variant of the EM algorithm developed by Lipsitz and Ibrahim (1998), in which Monte Carlo simulation is used to approximate the conditional expectation of the complete-data partial-likelihood score function given the observed data. The solution to this conditional expectation is not a maximum partial-likelihood estimator, so that the method is not statistically efficient. The Monte Carlo simulation is time-consuming. There is no formal justification for the method of Lipsitz and Ibrahim (1998) or that of Symons *et al.* (2002). Our method is computationally much simpler than Monte Carlo simulation. It is based on the maximum-likelihood estimator and is thus statistically efficient. Furthermore, we have established the theoretical properties of our method and assessed its empirical performance through simulation studies. The method of Symons *et al.* (2002) has the advantage that the baseline hazard function is unspecified.

The proposed resampling approach to determining genome-wide significance is applicable to arbitrary genetic maps and accommodates missing marker data and dominant markers. For the CIM and MIM models, no analytical thresholds are available and the permutation method is extremely time-consuming. The proposed resampling approach can be applied to the CIM and MIM models since all the relevant formulas have been presented for an arbitrary likelihood. For the MIM method, the resampling approach produces appropriate thresholds for testing each putative QTL given the others, with or without adjustment for the fact that multiple QTL are tested simultaneously.

We have written a computer program in C to implement the proposed method. This program is available from the authors upon request.

LITERATURE CITED

Boer, M. P., C. J. Braak and R. C. Jansen, 2002 A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. Genetics **162:** 951–960.

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

Broman, K. W., and T. P. Speed, 2002 A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J. R. Stat. Soc. B **64:** 731–775.

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

Cox, D. R., and D. V. Hinkley, 1974 *Theoretical Statistics.* Chapman & Hall, London.

Davies, R. B., 1977 Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **64:** 247–254.

Davies, R. B., 1987 Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika **74:** 33–43.

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

Doerge, R. W., Z-B. Zeng and B. S. Weir, 1997 Statistical issues in the search for genes affecting quantitative traits in experimental populations. Stat. Sci. **12:** 195–219.

Dupuis, J., and D. Siegmund, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics **151:** 373–386.

Ferreira, M. E., J. Satagopan, B. S. Yandell, P. H. Williams and T. C. Osborn, 1995 Mapping loci controlling vernalization requirement and flower time in *Brassica napus.* Theor. Appl. Genet. **90:** 727–732.

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.

Hilbert, P., K. Lindpaintner, J. S. Beckmann, T. Serikawa, F. Soubrier *et al.*, 1991 Chromosomal mapping of two genetic loci associated with blood-pressure regulation in hereditary hypertensive rats. Nature **353:** 521–529.

Jacob, H. J., K. Lindpaintner, S. E. Lincoln, K. Kusumi, R. K. Bunker *et al.*, 1991 Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell **67:** 213–224.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

Kalbfleish, J. D., and R. L. Prentice, 2002 *The Statistical Analysis of Failure Time Data*, Ed. 2. Wiley, Hoboken, NJ.

Kao, C. H., Z-B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lehmann, E. L., 1975 *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

Lin, D. Y., L. J. Wei and Z. Ying, 1993 Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika **80:** 557–572.

Lipsitz, S. R., and J. G. Ibrahim, 1998 Estimating equations with incomplete categorical covariates in the Cox model. Biometrics **54:** 1002–1013.

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

Rebai, A., B. Gofinet and B. Mangin, 1994 Approximate thresholds of interval mapping test for QTL detection. Genetics **138:** 235–240.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. **6:** 461–464.

Sen, S., and G. A. Churchill, 2001 A statistical framework of quantitative trait mapping. Genetics **159:** 371–387.

Shepel, L. A., H. Lan, J. D. Haag, G. M. Brasic, M. E. Gheen *et al.*, 1998 Genetic identification of multiple loci that control breast cancer susceptibility. Genetics **149:** 289–299.

Siegmund, D., 1985 *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.

Symons, R. C., M. J. Daly, J. Fridlyand, T. P. Speed, W. D. Cook *et al.*, 2002 Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in Eμ-v-abl transgenic mice. Proc. Natl. Acad. Sci. USA **99:** 11299–11304.

Zeng, Z-B., 1993 Theoretical basis of precision mapping of quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

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

Zou, F., B. S. Yandell and J. P. Fine, 2001 Statistical issues in the analysis of quantitative traits in combined crosses. Genetics **158:** 1339–1346.

Communicating editor: G. Churchill

APPENDIX A: COMPUTATIONS OF MLE AND COVARIANCE MATRIX

In this section, we present the formulas for the M-step of the EM algorithm under the Weibull model. In addition, we provide the observed information matrix as well as a consistent estimator of the covariance matrix of MLE $\hat{\boldsymbol{\theta}}$. In the M-step of the $(k+1)$th iteration, the first and second derivatives of the expected value of the log-likelihood given the observed data and current estimate $\hat{\boldsymbol{\theta}}^{(k)}$ are given by

$$\mathbf{U}^{(k+1)}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{g=-1,0,1} p_{i,g}(\hat{\boldsymbol{\theta}}^{(k)})\mathbf{U}_{i,g}(\boldsymbol{\theta}), \tag{A1}$$

$$\mathbf{I}^{(k+1)}(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{g=-1,0,1} p_{i,g}(\hat{\boldsymbol{\theta}}^{(k)})\mathbf{I}_{i,g}(\boldsymbol{\theta}), \tag{A2}$$

where

$$\mathbf{U}_{i,g}(\boldsymbol{\theta}) = \begin{pmatrix} \Delta_i\mathbf{g} - \Lambda_{i,g}\mathbf{g} \\ \Delta_i - \Lambda_{i,g} \\ \Delta_i(1 + e^{\alpha_2}\log Y_i) - \Lambda_{i,g}e^{\alpha_2}\log Y_i \end{pmatrix},$$

$$\mathbf{I}_{i,g}(\boldsymbol{\theta}) = \begin{pmatrix} -\Lambda_{i,g}\mathbf{g}\mathbf{g}^T & -\Lambda_{i,g}\mathbf{g} & -\Lambda_{i,g}e^{\alpha_2}\log Y_i\mathbf{g} \\ -\Lambda_{i,g}\mathbf{g}^T & -\Lambda_{i,g} & -\Lambda_{i,g}e^{\alpha_2}\log Y_i \\ -\Lambda_{i,g}e^{\alpha_2}\log Y_i\mathbf{g}^T & -\Lambda_{i,g}e^{\alpha_2}\log Y_i & e^{\alpha_2}\log Y_i\{\Delta_i - \Lambda_{i,g}(1 + e^{\alpha_2}\log Y_i)\} \end{pmatrix},$$

$\mathbf{g} = (g, 1 - |g|)^T$, and $\Lambda_{i,g} = \int_0^{Y_i}\lambda(t|g)\,dt$, which is the cumulative hazard function conditional on the QTL genotype g. Then, we can apply the Newton-Raphson algorithm to update the current estimate with the new maximizer $\hat{\boldsymbol{\theta}}^{(k+1)}$.

The observed information matrix is given by

$$\mathbf{I}(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{g=-1,0,1} p_{i,g}(\boldsymbol{\theta})\mathbf{I}_{i,g}(\boldsymbol{\theta}) - \sum_{i=1}^{n}\sum_{g=-1,0,1} p_{i,g}(\boldsymbol{\theta})\{\mathbf{U}_{i,g}(\boldsymbol{\theta})\}^{\otimes 2} + \sum_{i=1}^{n}\left\{\sum_{g=-1,0,1} p_{i,g}(\boldsymbol{\theta})\mathbf{U}_{i,g}(\boldsymbol{\theta})\right\}^{\otimes 2},$$

where for a column vector $\mathbf{a}$, $\mathbf{a}^{\otimes 2}$ denotes the matrix $\mathbf{a}\mathbf{a}^T$. Thus, a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by the inverse of the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$, *i.e.*, $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})$.

APPENDIX B: ASYMPTOTIC PROPERTIES OF SCORE AND LIKELIHOOD-RATIO STATISTICS

In this section, we show that LR(d) is asymptotically χ^2_2-distributed under H$_0$ and provide the necessary ingredients for deriving the thresholds. Let $\mathbf{U}(\boldsymbol{\theta}; d)$ and $\mathbf{I}(\boldsymbol{\theta}; d)$ be the observed-data score function and information matrix at location d with the following partitions to conform with the partition $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ of $\boldsymbol{\theta}$,

$$\mathbf{U}(\boldsymbol{\theta};\, d) = \begin{pmatrix} \mathbf{U}_\beta(\boldsymbol{\theta};\, d) \\ \mathbf{U}_\gamma(\boldsymbol{\theta};\, d) \end{pmatrix},$$

and

$$\mathbf{I}(\boldsymbol{\theta};\, d) = \begin{pmatrix} \mathbf{I}_{\beta\beta}(\boldsymbol{\theta};\, d) & \mathbf{I}_{\beta\gamma}(\boldsymbol{\theta};\, d) \\ \mathbf{I}_{\gamma\beta}(\boldsymbol{\theta};\, d) & \mathbf{I}_{\gamma\gamma}(\boldsymbol{\theta};\, d) \end{pmatrix}.$$

Under H_0, $n^{-1}\mathbf{I}(\tilde{\boldsymbol{\theta}})$ converges to $\Sigma(d)$. Denote

$$\Sigma^{-1}(d) = \begin{pmatrix} \Sigma^{\beta\beta}(d) & \Sigma^{\beta\gamma}(d) \\ \Sigma^{\gamma\beta}(d) & \Sigma^{\gamma\gamma}(d) \end{pmatrix}.$$

It can be shown that $\mathrm{LR}(d) = n^{-1}\mathbf{U}_\beta^T(\tilde{\boldsymbol{\theta}};\, d)\Sigma^{\beta\beta}(d)\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}};\, d)$ asymptotically (Cox and Hinkley 1974, Sect. 9.3). Through Taylor series expansions, $n^{-1/2}\mathbf{U}_\beta(\tilde{\boldsymbol{\theta}};\, d) = n^{-1/2}\Sigma_{i=1}^{n}\tilde{\mathbf{U}}_i(\boldsymbol{\theta}_0;\, d)$ asymptotically, where $\tilde{\mathbf{U}}_i(\boldsymbol{\theta}_0;\, d) = \mathbf{U}_{\beta,i}(\boldsymbol{\theta}_0;\, d) - \Sigma_{\beta\gamma}(d)\Sigma_{\gamma\gamma}^{-1}(d)\mathbf{U}_{\gamma,i}(\boldsymbol{\theta}_0;\, d)$ and $\boldsymbol{\theta}_0 = (\mathbf{0}, \boldsymbol{\gamma}_0)$. The replacements of the unknown quantities in $\tilde{\mathbf{U}}_i(\boldsymbol{\theta}_0;\, d)$ with their sample estimators yield $\hat{\mathbf{U}}_i(d) = \mathbf{U}_{\beta,i}(\tilde{\boldsymbol{\theta}};\, d) - \mathbf{I}_{\beta\gamma}(\tilde{\boldsymbol{\theta}};\, d)\mathbf{I}_{\gamma\gamma}^{-1}(\tilde{\boldsymbol{\theta}};\, d)\mathbf{U}_{\gamma,i}(\tilde{\boldsymbol{\theta}};\, d)$.

Let $\mathbf{z}(d) = (\Sigma^{\beta\beta}(d))^{1/2}(n^{-1/2}\mathbf{U}_\beta(\boldsymbol{\theta}_0;\, d) - \Sigma_{\beta\gamma}(d)\Sigma_{\gamma\gamma}^{-1}(d)n^{-1/2}\mathbf{U}_\gamma(\boldsymbol{\theta}_0;\, d))$. Then $\mathbf{z}(d)$ converges to a normal distribution with mean $\mathbf{0}$ and an identity 2×2 covariance matrix. Thus, $\mathrm{LR}(d)$ is asymptotically distributed as χ_2^2 under H_0.

APPENDIX C: ANALYTICAL APPROXIMATIONS OF THRESHOLDS

We show in this appendix that, for infinitely dense markers, the null distribution of $\mathrm{LR}(d)$ can be approximated by an Ornstein-Uhlenbeck process. Under H_0, $\Sigma(d)$ does not depend on d. Let d_1 and d_2 denote two points on the chromosome, and r be the recombination fraction corresponding to the genetic distance $|d_1 - d_2|$. Under the assumption of no crossover interference, it is easy to show that the correlation between $\mathbf{g}(d_1)$ and $\mathbf{g}(d_2)$ is given by

$$\mathrm{Corr}(\mathbf{g}(d_1),\, \mathbf{g}(d_2)) \approx \begin{pmatrix} 1 - 2r & 0 \\ 0 & 1 - 4r \end{pmatrix}$$

provided that r is small. Since $\mathbf{U}_{\beta,i}(\boldsymbol{\theta};\, d) = (\Delta_i - \Lambda_0(Y_i))\mathbf{g}_i(d)$ under H_0, we have

$$\mathrm{Corr}(\mathbf{z}(d_1),\, \mathbf{z}(d_2)) = \mathrm{Corr}(\mathbf{g}(d_1),\, \mathbf{g}(d_2)) \approx \begin{pmatrix} 1 - 2r & 0 \\ 0 & 1 - 4r \end{pmatrix}.$$

The above result implies that $z_1(d)$ and $z_2(d)$, the first and second components of $\mathbf{z}(d)$, are approximately independent Ornstein-Uhlenbeck processes with means zero and variances $1 - 2r$ and $1 - 4r$, respectively. By the arguments of Dupuis and Siegmund (1999), the tail distribution of $\sup_d \mathrm{LR}(d)$ under H_0 satisfies

$$P(\sup_d \mathrm{LR}(d) \geq a^2) \approx 1 - \exp\{-(C + 3va^2L)\exp(-a^2/2)\}, \tag{C1}$$

where Δ is the average marker distance (in morgans), C is the number of chromosomes, L is the total length of the genome (in morgans), and $v = v(a(6\Delta)^{1/2})$, the definition of which can be found in Siegmund (1985). When $\Delta = 0$, the above formula reduces to that of Lander and Botstein (1989). These results imply that all the analytical thresholds for the normal trait can be applied to our case.