Biometrics 61, 789–798
September 2005

DOI: 10.1111/j.1541-0420.2005.00346.x

Semiparametric Methods for Mapping Quantitative Trait Loci with Censored Data

Guoqing Diao and D. Y. Lin*

Department of Biostatistics, CB No. 7420, University of North Carolina, Chapel Hill,
North Carolina 27599-7420, U.S.A.
**email:* lin@bios.unc.edu

Summary. Statistical methods for the detection of genes influencing quantitative traits with the aid of genetic markers are well developed for normally distributed, fully observed phenotypes. Many experiments are concerned with failure-time phenotypes, which have skewed distributions and which are usually subject to censoring because of random loss to follow-up, failures from competing causes, or limited duration of the experiment. In this article, we develop semiparametric statistical methods for mapping quantitative trait loci (QTLs) based on censored failure-time phenotypes. We formulate the effects of the QTL genotype on the failure time through the Cox (1972, *Journal of the Royal Statistical Society, Series B* **34,** 187–220) proportional hazards model and derive efficient likelihood-based inference procedures. In addition, we show how to assess statistical significance when searching several regions or the entire genome for QTLs. Extensive simulation studies demonstrate that the proposed methods perform well in practical situations. Applications to two animal studies are provided.

Key words: Experimental population; Interval mapping; Nonparametric likelihood; Proportional hazards; QTL; Survival data.

1. Introduction

Modern techniques in molecular biology have stimulated the development of many statistical methods for detecting quantitative trait loci (QTLs) along the genome with the aid of genetic markers. Most of these methods stem from the landmark work of Lander and Botstein (1989). The Lander–Botstein interval mapping method postulates that the QTLs exist at some unknown positions bracketed by known genetic markers and that the trait value depends on the QTL genotype through a linear regression model. The likelihood-ratio statistic for testing the null hypothesis of no QTL present is calculated at each position along the genome. The position with the largest value of the test statistic is declared to be the putative QTL location provided that the value exceeds a certain threshold level. Doerge, Zeng, and Weir (1997) provided an excellent review of this method and various extensions.

Most of the existing methods for QTL mapping require that the phenotype be normally distributed, possibly after a simple transformation, and be fully observed. These assumptions are likely to be violated when the phenotype pertains to the survival time or failure time, which has a skewed distribution and which is often subject to censoring. Ferreira et al. (1995) described a plant experiment on flower times, some of which were censored due to limited duration of the experiment. Broman (2003) presented data from an experiment using mice, in which the trait of interest was time to death after a bacterial infection and in which 30% of the mice were still alive at the end of the study period. Symons et al. (2002) described another mouse study, in which the phenotype was the time until terminal illness due to tumor for Eμ-v-abl transgenic mice and in which censoring was induced by deaths due to unrelated causes or the end of the study.

Censoring presents a major challenge in the application of the interval mapping approach. Broman (2003) considered a cure model by regarding the mice alive at the end of the study as cured and by postulating a log-normal distribution for the survival times among the deaths. This method was not really developed for censored phenotypes, and can only deal with the situations in which the potential censoring times are equal among all study subjects. Symons et al. (2002) specified a Cox proportional hazards model for the effects of the QTL genotype on the failure time and estimated the model parameters by a variant of the expectation–maximization (EM) algorithm (Lipsitz and Ibrahim, 1998). The properties of this method have not been carefully investigated. Recently, Diao, Lin, and Zou (2004) developed simple likelihood-based methods under the same model but with the baseline hazard function parameterized.

In this article, we consider the same model as Symons et al. (2002), and develop efficient and rigorous methods based on the nonparametric (NP) likelihood. We also show how to assess the statistical significance when searching multiple regions or the entire genome for QTLs. These methods are presented in the next section. Section 3 reports the results of our simulation studies, while Section 4 illustrates the proposed methods with the data of Broman (2003) and Symons et al. (2002). The results are discussed in Section 5.

2. Methods

2.1 *Interval Mapping*

In this section, we develop a semiparametric interval mapping method for potentially censored failure-time traits in an F_2 intercross population. It is straightforward to extend the results to other crosses. 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 $\Delta_i = I(T_i \le C_i)$, where $I(\cdot)$ is the indicator function.

Suppose that we have data on a set of genetic markers with a known linkage map. Let $\mathbf{M}_i$ denote the multipoint marker genotype data for the ith subject. We consider a putative QTL 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

$$\lambda(t \mid G_i) = \alpha(t) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{G}_i}, \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{G}_i = (G_i, 1 - |G_i|)^{\mathrm{T}}, \boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$, and $\alpha(\cdot)$ is an unspecified baseline hazard function. Note that β_1 and β_2 pertain to the additive and dominant effects of the QTL, respectively.

At each locus, we may calculate $\pi_{i,g} = \Pr(G_i = g \mid \mathbf{M}_i)$ ($g = -1, 0, 1$; $i = 1, \ldots, n$), 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 (Kalbfleisch and Prentice, 2002, p. 195) and that the censoring time is independent of the failure time and unobserved QTL genotype. Then, the likelihood for the complete data (Y_i, Δ_i, G_i, $\mathbf{M}_i$) ($i = 1, \ldots, n$) is proportional to

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

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

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

Throughout this article, the summation or product over g is over $\{-1, 0, 1\}$, corresponding to three possible QTL genotypes. Write $\boldsymbol{\theta} = (\boldsymbol{\beta}, A)$, where $A(t) \equiv \int_0^t \alpha(u)\, du$ is the cumulative baseline hazard function. It is straightforward to see that the conditional distribution of QTL genotype G_i given the observed data on the ith subject $X_i \equiv (Y_i, \Delta_i, \mathbf{M}_i)$ is multinomial with probabilities

$$\begin{aligned} p_{i,g}(\boldsymbol{\theta}) &= P(G_i = g \mid X_i, \boldsymbol{\theta}) \\ &= \frac{\pi_{i,g} \exp\left\{\Delta_i \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g} - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}}\right\}}{\sum_{\tilde{g}} \pi_{i,\tilde{g}} \exp\left\{\Delta_i \boldsymbol{\beta}^{\mathrm{T}} \tilde{\mathbf{g}} - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \tilde{\mathbf{g}}}\right\}}, \end{aligned} \tag{4}$$

where $\mathbf{g} = (g, 1 - |g|)^{\mathrm{T}}$ and $\tilde{\mathbf{g}} = (\tilde{g}, 1 - |\tilde{g}|)^{\mathrm{T}}$.

Our goal is to estimate $\boldsymbol{\beta}$ and $A(t)$ by using the observed-data likelihood given in (3). As in the standard Cox regression, the maximum of the likelihood over the parameter space $R^2 \times$ {absolutely continuous cumulative hazards} does not exist. This can be shown by contradiction. Suppose that there exists an estimator $(\hat{A}(t), \hat{\boldsymbol{\beta}})$ that maximizes the likelihood. Because $\hat{A}(t)$ is absolutely continuous, we can always find another absolutely cumulative hazard function $\tilde{A}(t)$ with a larger derivative at $Y_{(1)}$ and $\tilde{A}(t) = \hat{A}(t)$ for $t \ge Y_{(1)}$, where $Y_{(1)}$ is the first observed failure time. The new estimator $(\tilde{A}(t), \hat{\boldsymbol{\beta}})$ will yield a larger likelihood than $(\hat{A}(t), \hat{\boldsymbol{\beta}})$, so that there is a contradiction. To resolve this problem, we extend the parameter space to $R^2 \times$ {right continuous cumulative hazards} to allow for a discrete estimator, and replace $\alpha(t)$ in (2) and (3) by $A\{t\}$, the jump size of A at the time point t. The resulting NP likelihood functions are given by

$$L_n^f(\boldsymbol{\beta}, A) = \prod_{i=1}^{n} \prod_{g} \left[A^{\Delta_i}\{Y_i\} \exp\left(\Delta_i \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g} - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}}\right)\right]^{I(G_i = g)} \tag{5}$$

and

$$L_n(\boldsymbol{\beta}, A) = \prod_{i=1}^{n} \sum_{g} A^{\Delta_i}\{Y_i\} \exp\left(\Delta_i \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g} - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}}\right) \pi_{i,g}. \tag{6}$$

We assume that there exists a finite time point τ such that $P(C \ge \tau) = P(C = \tau) > 0$. We also assume that $P(T > \tau) > 0$ and $P(T \le C \mid G) > 0$ for any QTL genotype G, which ensure that we can observe failures on the entire interval and therefore can estimate A on the entire interval. For $\boldsymbol{\beta}$ to be identifiable, we assume that if $P(\mathbf{c}_1^{\mathrm{T}} \boldsymbol{\pi} / \mathbf{c}_2^{\mathrm{T}} \boldsymbol{\pi} = c_0) = 1$, then $c_{1j}/c_{2j} = c_0 (j = 1, 2, 3)$, where $\boldsymbol{\pi}$ is the vector of conditional probabilities of the QTL genotypes given the marker data. This assumption is equivalent to that the QTL genotype probability matrix $\boldsymbol{\Pi} \equiv (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n)$ is positive definite. Finally, we assume that $\boldsymbol{\beta}_0$ lie in a known compact set, say $\mathcal{B}, \alpha_0(\cdot) > 0$, and $A_0(\tau) = \int_0^\tau \alpha_0(u)\, du < \infty$, where $\boldsymbol{\beta}_0$ and A_0 are the true values of $\boldsymbol{\beta}$ and A.

From the fact that $\boldsymbol{\beta}_0$ lies in a compact set and that the NP likelihood function L_n is continuous with respect to the unknown parameters, it is simple to show that the maximizer of L_n exists. Denote the maximizer of $L_n(\boldsymbol{\beta}, A)$ by $\hat{\boldsymbol{\theta}} \equiv (\hat{\boldsymbol{\beta}}_n, \hat{A}_n)$, which is referred to as the NP maximum likelihood estimator (NPMLE). It is easy to see that $\hat{A}_n$ must be discrete with positive jump sizes at the observed failure times only.

Instead of maximizing (6) directly, we apply the EM algorithm to (5), as is commonly done for the proportional hazards frailty model (Klein, 1992; Nielsen et al., 1992; Andersen et al., 1993, Chapter IX). Specifically, in the $(k + 1)$th iteration of the E-step, we calculate the expected value of the complete-data log likelihood $\log L_n^f(\boldsymbol{\beta}, A)$ given the observed data and current estimate $\hat{\boldsymbol{\theta}}^{(k)}$:

$$\sum_{i=1}^{n} \sum_{g} p_{i,g}\big(\hat{\boldsymbol{\theta}}^{(k)}\big)\big\{\Delta_i\big(\log A\{Y_i\} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{g}\big) - A(Y_i)e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{g}}\big\}. \quad (7)$$

In the $(k + 1)$th iteration of the M-step, we maximize (7) to update the estimates of $\boldsymbol{\beta}$ and A. The estimate $\hat{\boldsymbol{\beta}}^{(k+1)}$ is the solution to the following equation

$$\sum_{i=1}^{n} \Delta_i \left\{ \sum_{g} p_{i,g}\big(\hat{\boldsymbol{\theta}}^{(k)}\big)\mathbf{g} - \frac{\displaystyle\sum_{j=1}^{n} I(Y_j \geq Y_i) \sum_{g} p_{j,g}\big(\hat{\boldsymbol{\theta}}^{(k)}\big)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{g})\mathbf{g}}{\displaystyle\sum_{j=1}^{n} I(Y_j \geq Y_i) \sum_{g} p_{j,g}\big(\hat{\boldsymbol{\theta}}^{(k)}\big)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{g})} \right\} = \mathbf{0},$$

and the estimate for $A(t)$ takes the form

$$\hat{A}^{(k+1)}(t) = \sum_{i=1}^{n} \frac{\Delta_i I(Y_i \leq t)}{\displaystyle\sum_{j=1}^{n} I(Y_j \geq Y_i) \sum_{g} p_{j,g}\big(\hat{\boldsymbol{\theta}}^{(k)}\big)\exp\big(\hat{\boldsymbol{\beta}}^{(k+1)^{\mathrm{T}}}\mathbf{g}\big)}.$$

We start the EM algorithm by assigning an initial value to $\boldsymbol{\theta}$ and iterate until convergence. The computation of the covariance matrix for the parameter estimators is given in Appendix A. We show in Appendix B that $\hat{\boldsymbol{\beta}}_n$ is consistent, asymptotically normal, and asymptotically efficient.

Symons et al. (2002) also considered model (1). They utilized a variant of the EM algorithm described 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 computation is demanding. The estimator does not maximize the observed-data likelihood, and is thus expected to be asymptotically inefficient.

We test the null hypothesis of no QTL effects, i.e., $\mathrm{H}_0 : \boldsymbol{\beta} = \mathbf{0}$, by the likelihood-ratio statistic

$$LR = 2\log\frac{L_n(\hat{\boldsymbol{\theta}}_n)}{L_n(\tilde{\boldsymbol{\theta}}_n)},$$

where $\tilde{\boldsymbol{\theta}}_n = (\mathbf{0}, \tilde{A}_n)$, with $\tilde{A}_n$ being the restricted NPMLE of A under H_0. In Appendix B, we show that under H_0, LR is asymptotically χ^2 distributed with 2 degrees of freedom. Note that $p_{i,g}(\boldsymbol{\theta}), \hat{\boldsymbol{\theta}}_n, L_n(\hat{\boldsymbol{\theta}}_n)$, and LR all depend on the locus d through the dependence of $\pi_{i,g}$ on d. In the sequel, we will include d in the expressions to emphasize their dependence on d if necessary. Note also that $\tilde{\boldsymbol{\theta}}_n$ and $L_n(\tilde{\boldsymbol{\theta}}_n)$ do not depend on d. In fact, $\tilde{A}_n$ is the conventional Nelson–Aalen estimator based on (Y_i, Δ_i) $(i = 1, \ldots, n)$. 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 curve of LR for each chromosome. The position associated with the largest value of LR 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. In the genetic literature, LOD score is often used as the test statistic, which has a simple relationship with LR: $LOD = LR/(2\log 10)$.

2.2 *Thresholds*

When searching the entire chromosome or whole genome for QTLs, one should select a threshold level such that the probability (under the null hypothesis) that LR 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 among different locations in the genome. Lander and Botstein (1989) and Dupuis and Siegmund (1999) showed that the likelihood-ratio test statistic can be approximated by an Ornstein–Uhlenbeck process for normal traits. Diao et al. (2004) obtained analogous results for failure-time traits under parametric survival models. 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 the results of Davies (1977, 1987) and Rebai, Gofinet, and Mangin (1994) provided approximate thresholds for backcross 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.

Diao et al. (2004) proposed a resampling approach to determining the thresholds for genome-wide statistical significance under parametric survival models, which allows arbitrary distributions of the markers and arbitrary test positions, and which also accommodates missing marker data and dominant markers. The resampling approach is computationally much more efficient than the permutation method proposed by Churchill and Doerge (1994). We extend this approach to the semiparametric model.

It follows from equation (B.2) in Appendix B that, under H_0, $LR(d)$ is asymptotically equivalent to

$$n^{-1}\left\{\sum_{i=1}^{n} \tilde{\mathbf{l}}_0^{\mathrm{T}}(X_i; d)\right\} \tilde{\mathbf{I}}_0^{-1}(d) \left\{\sum_{i=1}^{n} \tilde{\mathbf{l}}_0(X_i; d)\right\},$$

where $\tilde{\mathbf{l}}_0$ is the efficient score function for $\boldsymbol{\beta}$, and $\tilde{\mathbf{I}}_0$ is its covariance matrix or the efficient information matrix. The expressions for $\tilde{\mathbf{l}}_0$ and $\tilde{\mathbf{I}}_0$ are given in Appendix B. We replace the unknown quantities in $\tilde{\mathbf{l}}_0$ by their sample estimators to yield $\hat{\mathbf{l}}$, and estimate $\tilde{\mathbf{I}}_0(d)$ consistently by $\hat{\mathbf{I}}(d) \equiv n^{-1}\sum_{i=1}^{n} \hat{\mathbf{l}}(X_i; d)\hat{\mathbf{l}}^{\mathrm{T}}(X_i; d)$. Indeed, $\hat{\mathbf{l}}$ can be obtained from the results in Appendix A in a similar fashion as the parametric models. We regard the jump sizes of the A at the observed failure times rather than the Euclidean parameters which parameterize the baseline hazard as the nuisance parameters.

Following Diao et al. (2004), we define $\hat{\mathbf{U}}(d) = \sum_{i=1}^{n} \hat{\mathbf{l}}(X_i; d)Z_i$, 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 $n^{-1}\sum_{i=1}^{n}\hat{\mathbf{I}}(X_i, d_1)\hat{\mathbf{I}}^{\mathrm{T}}(X_i, d_2)$ at $(d_1,\ d_2)$, which converges to the covariance function of the limiting Gaussian process of $n^{-1/2}\sum_{i=1}^{n}\tilde{\mathbf{I}}_0(X_i; d)$. Thus, the conditional distribution of $n^{-1/2}\hat{\mathbf{U}}(d)$ given the observed data converges to the limiting distribution of $n^{-1/2}\sum_{i=1}^{n}\tilde{\mathbf{I}}_0(X_i; d)$. As a result, the distribution of $LR(d)$ can be approximated by that of

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

To approximate the distribution of $\sup_d LR(d)$, 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.

3. Simulation Studies

Extensive simulation studies were performed to investigate the operating characteristics of the proposed methods in practical situations. We generated the failure times from the proportional hazards model $\lambda(t \mid \mathbf{G}) = \lambda_0(t)e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{G}}$ with Weibull baseline hazard function $\lambda_0(t) = \gamma_1\gamma_2 t^{\gamma_2 - 1}$, where $\gamma_1 = 0.01$ and $\gamma_2 = 2$. The censoring times were generated from the uniform $(0, \tau)$ distribution, where τ was chosen to yield approximately 30% censored observations. Assuming no crossover interference, we generated the marker data using a Markov chain. Specifically, we simulated the genotype of the first marker on a chromosome from a multinomial distribution with probabilities 0.25, 0.50, and 0.25 corresponding to three possible genotypes. We then generated the subsequent marker genotypes from the conditional distribution given the previous marker genotype. The likelihood-ratio statistic was calculated at the 1-cM increment.

In all the simulation studies, we considered a chromosome with a total length of 100 cM. We generated both equally and unequally spaced markers. We also simulated data with missing marker genotypes and dominant markers. For the cases of unevenly spaced markers, we placed m markers at the following locations:

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

where LOC_j is the jth marker location and $[\frac{m}{2}]$ is the largest integer that is less than or equal to $\frac{m}{2}$. In these settings, the first half of the markers are denser than the second half of the markers. Under H_0, the QTL effects β_1 and β_2 are 0 at all locations on the chromosome. We assume only one QTL located at 35 cM with $\beta_1 = 0.5$ and $\beta_2 = 0.4$ under H_1. We generated 10,000 replicates of 200 observations from an F_2 population.

We first evaluated the finite-sample properties of the NPMLE at the true QTL location of H_1. For comparisons, we also evaluated Symons et al.'s (2002) estimator and Diao et al.'s (2004) parametric estimator. The results for the estimators of the additive QTL effect are summarized in Table 1. The NPMLE at the true QTL location appears to be virtually unbiased. The standard error estimator reflects accurately the standard error. The confidence intervals have proper coverage probabilities. The NPMLE is nearly as efficient as the parametric estimator and tends to be more efficient than Symons et al.'s estimator. We obtained similar results for the estimation of the dominant QTL effect (results not shown).

We also examined the properties of the estimated QTL location, i.e., the chromosome position where the likelihood-ratio test statistic $LR(d)$ is maximized, and the corresponding estimators of the genetic effects. The results are summarized in Table 2. There is little bias for the estimator of the QTL location or the estimators of the genetic effects.

We conducted additional simulation studies to evaluate the ability of the semiparametric method in identifying the

Table 1

Summary statistics for the estimators of the additive QTL effects

No. of markers	Marker pattern	Semiparametric				Symons et al.		Parametric	
		Mean	SE	SEE	CP	Mean	SE	Mean	SE
				$\mathrm{H}_0 : \boldsymbol{\beta} = (0.0, 0.0)$					
6	1	0.000	0.135	0.132	95.0	0.000	0.135	0.000	0.137
11	1	0.002	0.131	0.129	94.8	0.002	0.130	0.002	0.132
51	1	0.000	0.124	0.123	95.0	0.000	0.124	0.000	0.123
6	2	−0.001	0.162	0.162	95.3	−0.001	0.170	−0.001	0.163
11	2	0.000	0.151	0.151	95.4	0.000	0.156	0.000	0.151
51	2	0.000	0.145	0.144	95.1	0.001	0.149	0.000	0.145
				$\mathrm{H}_1 : \boldsymbol{\beta} = (0.5, 0.4)$					
6	1	0.506	0.142	0.141	95.0	0.509	0.144	0.508	0.137
11	1	0.505	0.138	0.136	95.0	0.507	0.140	0.505	0.134
51	1	0.508	0.132	0.131	95.2	0.508	0.132	0.507	0.127
6	2	0.502	0.175	0.173	95.5	0.518	0.186	0.508	0.165
11	2	0.507	0.166	0.161	95.1	0.519	0.173	0.510	0.158
51	2	0.507	0.155	0.153	95.1	0.515	0.160	0.509	0.148

Notes: Under marker 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. Mean and SE are the mean and standard error of the parameter estimator; SEE is the mean of the standard error estimator; CP is the coverage probability of the 95% confidence interval. Each entry is based on 10,000 simulated data sets.

Table 2

Sampling means of the QTL location estimator $\hat{d}$ and the corresponding estimators of the genetic effects in the simulation studies

		$H_0 : \beta_1 = 0, \beta_2 = 0$			$H_1 : \beta_1 = 0.5, \beta_2 = 0.4$		
			Genetic effects			Genetic effects	
No. of markers	Marker pattern	$\hat{d}$ (cM)	β_1	β_2	$\hat{d}$ (cM)	β_1	β_2
6	1	49.6	0.001	−0.013	36.6	0.519	0.412
11	1	50.1	0.000	−0.013	35.4	0.516	0.404
51	1	50.4	0.002	−0.010	35.3	0.527	0.418
6	2	48.5	0.000	−0.006	34.4	0.496	0.365
11	2	49.1	−0.001	−0.004	35.7	0.494	0.367
51	2	50.6	0.000	−0.011	35.2	0.495	0.361

Notes: Under marker 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. Each entry is based on 10,000 simulated data sets.

Table 3

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

				Resampling				Analytical			
		Empirical		H_0		H_1		Dense map		Sparse map	
No. of markers	Marker pattern	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%
6	1	10.07	13.57	9.77	13.30	9.79	13.33	13.37	17.12	9.15	12.39
11	1	10.62	14.22	10.38	13.94	10.41	13.97	13.37	17.12	10.15	13.53
51	1	11.84	15.39	11.40	15.04	11.46	15.09	13.37	17.12	11.80	15.37
6	2	9.80	13.49	9.61	13.14	9.62	13.16	13.37	17.12	9.15	12.39
11	2	10.48	14.15	10.20	13.77	10.22	13.79	13.37	17.12	10.15	13.53
51	2	11.54	15.28	11.12	14.74	11.15	14.78	13.37	17.12	11.80	15.37

Notes: Under marker 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. Empirical thresholds pertain to the percentiles of the test statistic based on 10,000 simulated data sets under H_0. The resampling-based thresholds shown in the table are the mean thresholds from 10,000 simulated data sets. Under H_1, QTL is located at 35 cM with $\beta_1 = 0.50$ and $\beta_2 = 0.40$.

QTL when the analytical or resampling approach is used to control genome-wide statistical significance. The dense-map and sparse-map approximations were obtained from formula (C1) in Diao et al. (2004). 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

Table 4

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

						Analytical							
		Resampling				Dense map				Sparse map			
		H_0		H_1		H_0		H_1		H_0		H_1	
No. of markers	Marker pattern	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%	$\alpha = 5\%$	1%
6	1	5.62	1.14	92.64	81.29	1.10	0.21	81.08	64.57	7.28	1.79	94.08	84.77
11	1	5.66	1.10	93.76	83.30	1.48	0.24	85.32	70.63	6.12	1.39	94.29	84.73
51	1	5.98	1.24	95.37	87.14	2.62	0.37	91.64	80.36	5.07	1.04	94.67	86.17
6	2	5.37	1.16	77.46	56.64	1.09	0.21	55.25	34.94	6.60	1.59	80.18	61.04
11	2	5.64	1.13	82.28	62.97	1.39	0.24	65.38	45.16	5.77	1.28	82.63	64.48
51	2	6.04	1.26	84.58	66.70	2.23	0.45	74.05	53.93	4.49	0.97	81.80	63.51

Notes: Sizes (powers) are the probabilities of rejecting the null hypothesis under H_0 (H_1). Under marker 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. Under H_1, QTL is located at 35 cM with $\beta_1 = 0.50$ and $\beta_2 = 0.40$.

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				Weibull				Nonparametric		Two-part model	
Chromosome	Pos. (cM)	*LOD*	β_1	β_2	Pos. (cM)	*LOD*	β_1	β_2	Pos. (cM)	*LOD*	Pos. (cM)	*LOD*
1	75	2.61	−0.527	−0.561	75	1.94	−0.456	−0.542	76	3.38	81	5.45
5	28	6.50	0.952	0.113	28	9.01	1.149	0.100	27	5.41	28	6.79
6	59	2.71	−0.499	0.467	59	3.66	−0.559	0.563	59	2.45	10	4.09
13	26	6.15	−0.573	−0.713	26	6.64	−0.614	−0.740	26	6.71	26	7.38
15	23	3.64	0.384	−0.778	23	4.49	0.370	−0.935	23	3.49	16	4.61

Notes: The thresholds for the *LOD* score at the 5% genome-wide significance level based on the resampling approach are 3.36 for the proposed method and 3.43 for the parametric Weibull method. The threshold for the *LOD* score at the 5% genome-wide significance level based on the permutation approach are 3.27 for the nonparametric method and 4.91 for the two-part model.

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 the case of 51 markers with $\alpha = 0.05$, the sizes are 2.78% and 2.31% for marker patterns 1 and 2, respectively.

4. Applications

To illustrate our methods, we now consider two studies. The first study was previously considered by Broman (2003) and Diao et al. (2004), while the second was considered by Symons et al. (2002). For each study, we fitted a semiparametric Cox regression model. In the first study, we also fitted a parametric model with a Weibull baseline hazard as in Diao et al. (2004).

4.1 Listeria Monocytogenes *Mice Study*

The data used in this example, originally published by Boyartchuk et al. (2001), contain observations on the time to death following infection with *Listeria monocytogenes* in 116 female F_2 mice from an intercross between the BALB/cByJ and C57BL/6ByJ strains. The mice were genotyped at 133 markers, including 2 on the X chromosome. Approximately 30% of the mice were alive at the end of the study.

Broman (2003) proposed an NP approach and a two-part model. The NP approach is an extension of the Kruskal–Wallis statistic (Lehmann, 1975, Section 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. Diao et al. (2004) fitted a proportional hazards model with a Weibull baseline hazard.

We applied the proposed methods to these data, and the results are shown in Table 5 and Figures 1 and 2. For simplicity, we only present the results on those chromosomes that contain QTLs identified by at least one of four methods: the proposed method, Diao et al.'s (2004) parametric Weibull method, and Broman's (2003) NP method and two-part model. The threshold for the *LOD* score at the 5% genome-wide significance level based on the resampling approach is 3.36, which is close to the threshold (i.e., 3.27) obtained by permutation for the NP approach of Broman (2003). The threshold based on the resampling approach for the parametric method is 3.43 (Diao et al., 2004). Our results are fairly consistent with those of Broman (2003) and Diao et al. (2004). We detect almost the same QTLs on chromosomes 5, 13, and 15, although no significant QTL effects are detected on chromosomes 1 and 6 at the 5% significance level. The QTL on chromosome 5 appears to have a strong additive effect and the hazard ratio with genotype *QQ* versus *qq* is about 6.71. The QTL on chromosome 13 appears to have both additive and dominant effects. The QTL on chromosome 15 appears to have a strong dominant

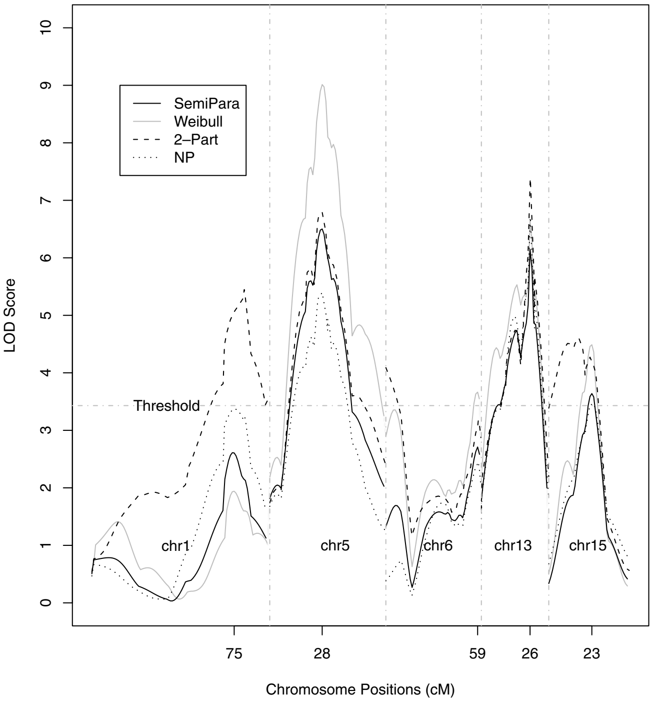

Figure 1. The *LOD* scores from four QTL mapping methods for the data on survival time following infection with *Listeria monocytogenes* in 116 intercross mice. The threshold pertains to the 5% genome-wide significance level under the resampling method.

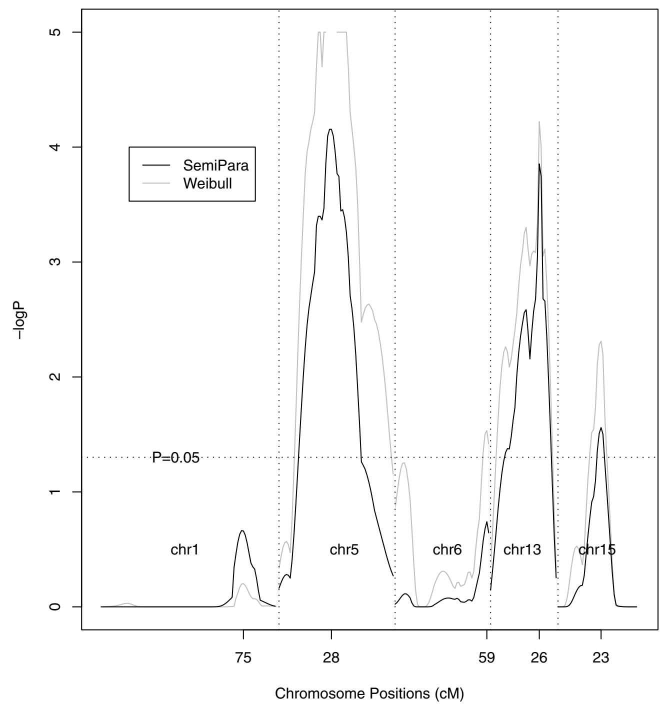

Figure 2. Plots of the $-\log_{10}P$ values for the proposed method and Weibull model for the data on survival time following infection with *Listeria monocytogenes* in 116 intercross mice. The P-values are based on 100,000 normal samples. For the parametric model, in the region between 26 cM and 30 cM on chromosome 5, the P-values are less than 10^{-5} and thus are not displayed.

effect. The QTLs on both chromosomes 13 and 15 appear to be overdominant, i.e., $|\beta_2| > |\beta_1|$.

Figure 1 shows the *LOD* curves from the aforementioned four methods. The *LOD* scores from the two-part model are larger than those of the other three methods at some QTL locations. There are, however, two more free parameters in the two-part model than in the other three methods. This will decrease the power to detect QTLs since a larger threshold (i.e., 4.91) is required. In order 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 Figure 2 of Broman (2003) reveal that the proposed method yields more significant results than the two-part method on the aforementioned chromosomes except chromosome 1.

4.2 *Eμ-v-abl Transgenic Mice Study*

This example is based on the data reported and analyzed by Symons et al. (2002) on 835 F_2 Eμ-v-abl transgenic mice. The phenotype of interest is the survival time until the onset of terminal illness due to the presence of tumors. Genetic map was constructed using the software MAPMAKER (Lander et al., 1987) and the resulting 145 autosomal markers are spaced on 19 chromosomes with a sparse density (>20 cM). Approximately 41% of the survival times were censored at the time points ranging from 49 to 476 days.

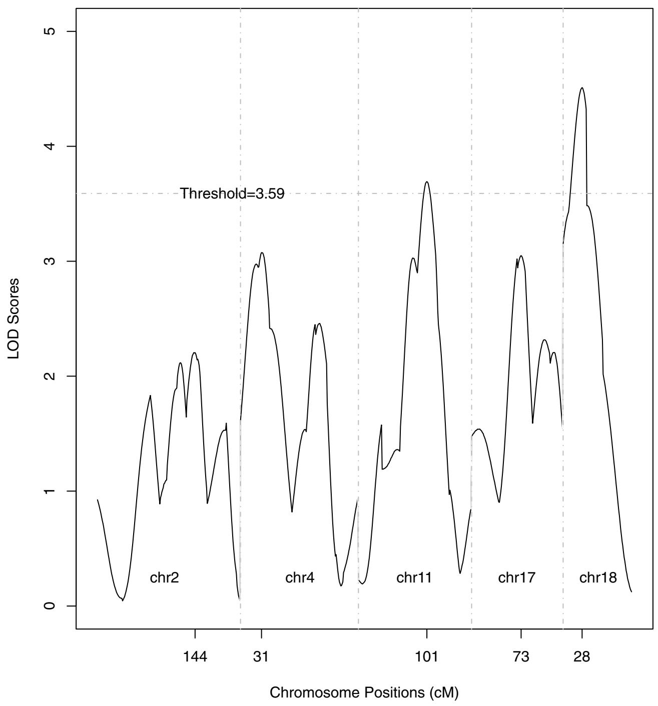

Figure 3. The *LOD* scores for the data on the time until terminal illness due to tumor in Eμ-v-abl transgenic mice. The threshold pertains to the 5% genome-wide significance level under the resampling method.

We applied the proposed methods to these data. Figure 3 presents the *LOD* score curves on chromosomes 2, 4, 11, 17, and 18. Compared with Figure 4 of Symons et al. (2002), the shapes of the *LOD* curves are similar, although there are differences in the magnitude of the *LOD* scores. Based on the threshold for the *LOD* score at the 5% genome-wide significance level from the resampling approach, i.e., 3.59, two QTLs on chromosome 11 and 18 are detected. Symons et al. (2002) adopted the genome-wide threshold based on the dense-map approximation (i.e., 4.3), which might not be appropriate for such a sparse map.

5. Discussion

We have extended the interval mapping approach of Lander and Botstein (1989) to potentially censored quantitative traits by formulating the QTL effects on the failure time through the Cox proportional hazards model. Compared to the parametric method of Diao et al. (2004), the proposed semiparametric method leaves the baseline hazard function completely unspecified and is thus more robust. The semiparametric nature of the model entails an infinite-dimensional nuisance parameter, which cannot be handled by standard statistical arguments. We have tackled the technical challenges by appealing to the modern empirical process theory. The proposed semiparametric method can also be applied to noncensored quantitative traits. Because no parametric distribution is assumed and the inference is based on the ranks of the trait values, this method is more robust than the traditional parametric QTL mapping methods.

Although we have described our method in the context of an F_2 population, extensions to other types of experimental populations, such as backcrosses, F_3 populations, and combined crosses, are straightforward. It is also simple to expand the vector **g** to include environmental variables, such as block factors in an agriculture field trial or cage number in a mouse cross. Finally, we can generalize the proposed method to multiple QTL models along the lines of Zeng (1993, 1994), Jansen (1993), and Kao, Zeng, and Teasdale (1999).

Acknowledgements

The authors are grateful to a referee and an associate editor for their helpful comments. This research was supported by the National Cancer Institute.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

Boyartchuk, V. L., Broman, K. W., Mosher, R. E., and Dietrich, W. (2001). Multigenic control of *Listeria monocytogenes* susceptibility in mice. *Nature Genetics* **27,** 259–260.

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

Chen, H. Y. and Little, R. J. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* **94,** 896–908.

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

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

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.

Diao, G. Q., Lin, D. Y., and Zou, F. (2004). Mapping quantitative trait loci with censored observations. *Genetics* **168,** 1689–1698.

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

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

Ferreira, M. E., Satagopan, J., Yandell, B. S., Williams, P. H., and Osborn, T. C. (1995). Mapping loci controlling vernalization requirement and flower time in *Brassica napus. Theoretical and Applied Genetics* **90,** 727–732.

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

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken, New Jersey: Wiley.

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

Klein, J. P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **48,** 785–806.

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

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., and Newburg, L. (1987). MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1,** 174–181.

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

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

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sunderland, Massachusetts: Sinauer.

Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95,** 449–485.

Murphy, S. A., Rossini, A. J., and Van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association* **92,** 968–976.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics* **19,** 25–43.

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

Rudin, W. (1973). *Functional Analysis.* New York: McGraw-Hill.

Symons, R. C. A., Daly, M. J., Fridlyand, J., Speed, T. P., Cook, W. D., Gerondakis, S., Harris, A. W., and Foote, S. J. (2002). Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in Eμ-v-abl transgenic mice. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 11299–11304.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* New York: Springer Verlag.

Zeng, Z. B. (1993). Theoretical basis of precision mapping of quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **90,** 10972–10976.

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

Received March 2004. *Revised November* 2004.
Accepted December 2004.

Appendix A

Derivatives and Covariance Estimates

Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^{\mathrm{T}}$ be the vector of the discrete baseline hazards at the distinct failure time points $T_{(1)}, \ldots, T_{(L)}$, where $T_{(l)}$ is the lth smallest distinct failure time and L is the total number of distinct failure time points. Note that $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \boldsymbol{\gamma})$ is an $(L+2) \times 1$ vector. From (6), the observed-data log likelihood for $\boldsymbol{\theta}$ is given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{g} A^{\Delta_i}\{Y_i\} \exp\bigl(\Delta_i \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g} - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}}\bigr) \pi_{i,g} \right\}.$$

Let

$$L_{i,g}(\boldsymbol{\theta}) = A^{\Delta_i}\{Y_i\} \exp\bigl(\Delta_i \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g} - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}}\bigr) \pi_{i,g}$$

and

$$l_{i,g}(\boldsymbol{\theta}) = \Delta_i \bigl(\log A\{Y_i\} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g} \bigr) - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}}.$$

Then, the first and second derivatives of $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are given by

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \sum_{g} p_{i,g}(\boldsymbol{\theta}) \frac{\partial l_{i,g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} = \sum_{i=1}^{n} \sum_{g} p_{i,g}(\boldsymbol{\theta}) \left\{ \frac{\partial^2 l_{i,g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} + \left(\frac{\partial l_{i,g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{\otimes 2} \right\} - \left\{ \sum_{g} p_{i,g}(\boldsymbol{\theta}) \frac{\partial l_{i,g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{\otimes 2},$$

where for a column vector $\mathbf{a}, \mathbf{a}^{\otimes 2}$ denotes the matrix $\mathbf{a}\mathbf{a}^{\mathrm{T}}$. Thus, we only need to calculate the first and second derivatives of $l_{i,g}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:

$$\frac{\partial l_{i,g}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \Delta_i \mathbf{g} - A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}} \mathbf{g},$$

$$\frac{\partial l_{i,g}(\boldsymbol{\theta})}{\partial \gamma_j} = \Delta_i I\bigl(Y_i = T_{(j)}\bigr) / \gamma_j - e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}} I\bigl(Y_i \geq T_{(j)}\bigr), \quad j = 1, \ldots, L,$$

$$\frac{\partial^2 l_{i,g}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = -A(Y_i) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}} \mathbf{g}^{\otimes 2},$$

$$\frac{\partial^2 l_{i,g}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \gamma_j} = -I\bigl(Y_i \geq T_{(j)}\bigr) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}} \mathbf{g}, \quad j = 1, \ldots, L,$$

$$\frac{\partial^2 l_{i,g}(\boldsymbol{\theta})}{\partial \gamma_j^2} = -\Delta_i I\bigl(Y_i = T_{(j)}\bigr) / \gamma_j^2, \quad j = 1, \ldots, L,$$

$$\frac{\partial^2 l_{i,g}(\boldsymbol{\theta})}{\partial \gamma_j \partial \gamma_k} = 0, \quad j, k = 1, \ldots, L \quad \text{and} \quad j \neq k.$$

The observed information matrix is the negative of the second derivatives of $l(\boldsymbol{\theta})$, i.e., $\mathbf{I}(\boldsymbol{\theta}) = -(\partial^2 l(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}})$. Thus, a consistent estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by the inverse of the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}_n$, i.e., $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\theta}}_n) = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n)$.

Appendix B

Asymptotic Properties of the NPMLE and Likelihood-Ratio Statistics

Consistency. To describe the consistency of the NPMLE, we denote the supremum norm on the interval $[0, \tau]$ by $\| \cdot \|_\infty$, and the Euclidean norm by $\| \cdot \|_2$. In the sequel, we use the notation P_n and G_n for the empirical distribution and the empirical process of the observations. Furthermore, we use P_0 for the expectation under the true measure of the observations. We shall prove that the NPMLE $(\hat{\boldsymbol{\beta}}_n, \hat{A}_n)$ is consistent; that is $\|\hat{A}_n - A_0\|_\infty$ and $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2$ converge almost surely to 0 as $n \to \infty$. We outline the proof as follows: the first step is to show that $\hat{A}_n(\tau)$ does not diverge to infinity. A natural approach is to show that the form of $\hat{A}_n$ forces itself to be bounded almost surely. Indeed, because $\boldsymbol{\beta}_0$ is in a compact set, there exist constants $B_u > 0$ and $B_l > 0$ such that

$$\frac{1}{B_u} \sum_{i=1}^{n} \frac{\Delta_i I(Y_i \leq t)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)} < \hat{A}_n(t) < \frac{1}{B_l} \sum_{i=1}^{n} \frac{\Delta_i I(Y_i \leq t)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)},$$

which implies that $\hat{A}_n$ is uniformly bounded almost surely. Once $\hat{A}_n(\tau)$ is proved to be bounded almost surely, Helly's selection theorem can be used to prove the existence of a convergent subsequence of $\hat{\boldsymbol{\theta}}_n$. The second step is to show that any such convergent subsequence of $\hat{\boldsymbol{\theta}}_n$ must converge to $\boldsymbol{\theta}_0$. The idea is to characterize the limit of a subsequence of $\hat{\boldsymbol{\theta}}_n$ by using the fact that $\log L_n(\hat{\boldsymbol{\beta}}_n, \hat{A}_n) - \log L_n(\boldsymbol{\beta}_0, \bar{A}_0) \geq 0$ for finite n, where

$$\bar{A}_0(t) = \sum_{i=1}^{n} \frac{\Delta_i I(Y_i \leq t)}{\sum_{j=1}^{n} I(Y_j \geq Y_i) \sum_{g} p_{i,g}(\boldsymbol{\theta}_0) \exp\bigl(\boldsymbol{\beta}_0^T \mathbf{g}\bigr)},$$

which uniformly converges to $A_0(t)$ almost surely. This will entail $l(\boldsymbol{\beta}^*, A^*) - l(\boldsymbol{\beta}_0, A_0) = 0$, where $l(\cdot)$ is the limit of $\log L_n(\cdot)$ under P_0 and $(\boldsymbol{\beta}^*, A^*)$ is the limit of the convergent subsequence of $(\hat{\boldsymbol{\beta}}_n, \hat{A}_n)$. Note that $l(\boldsymbol{\beta}^*, A^*) - l(\boldsymbol{\beta}_0, A_0)$ is the minus Kullback–Leibler information. As in Chen and Little (1999), we can prove that the model is identifiable; that is,

$$\frac{(dA(Y))^{\Delta} \sum_{g} \exp\bigl(\Delta \boldsymbol{\beta}^{\mathrm{T}} \mathbf{g} - A(Y) e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{g}}\bigr) \pi_g}{(dA_0(Y))^{\Delta} \sum_{g} \exp\bigl(\Delta \boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{g} - A_0(Y) e^{\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{g}}\bigr) \pi_g} = 1$$

almost surely implies that $(\boldsymbol{\beta}, A) = (\boldsymbol{\beta}_0, A_0)$. The identifiability of the model implies that $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $A^* = A_0$. Because every subsequence of n contains a further subsequence for $(\hat{\boldsymbol{\beta}}_n, \hat{A}_n)$ which converges uniformly to $(\boldsymbol{\beta}_0, A_0)$, we have the convergence of the entire sequence.

Asymptotic normality. We shall show that $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ converges in distribution to a normal distribution with mean 0 and covariance matrix $\tilde{\mathbf{I}}_0^{-1}$, where $\tilde{\mathbf{I}}_0$ is the efficient information matrix. We shall prove this result by verifying the four conditions in Theorem 3.3.1 of Van der Vaart and Wellner (1996). Define $\mathrm{H} = \{(\mathbf{h}_1, h_2) \,|\, \mathbf{h}_1 \in R^2, h_2$ has bounded variation], which is endowed with a norm $\|h\|^2 = \mathbf{h}_1^{\mathrm{T}} \mathbf{h}_1 + \int_0^\tau h_2^2 dA_0$

for $h = (\mathbf{h}_1,\, h_2) \in H$. The score operator for $(\boldsymbol{\beta}, A)$ takes the form $\dot{l}(\boldsymbol{\beta}, A)(h) = \mathbf{h}_1^{\mathrm{T}}\dot{\mathbf{l}}_1(\boldsymbol{\beta}, A) + \dot{l}_2(\boldsymbol{\beta}, A)(h_2)$, where $\dot{\mathbf{l}}_1(\boldsymbol{\beta}, A) = \Delta E_{(\boldsymbol{\beta},A)}(\mathbf{G}) - A(Y)E_{(\boldsymbol{\beta},A)}(\mathbf{G}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{g}})$ and $\dot{l}_2\,(\boldsymbol{\beta}, A)(h_2) = \Delta h_2(Y) - \int_0^Y h_2\, dAE_{(\boldsymbol{\beta},A)}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{G}})$. Here, $E_{(\boldsymbol{\beta},A)}$ is the expectation under the conditional distribution of the QTL genotype G given in (4). It is known that the class of uniformly bounded functions of bounded variations is a Donsker class. Applying Theorem 2.10.6 of Van der Vaart and Wellner (1996) and using the fact that $(\hat{\boldsymbol{\beta}}_n, \hat{A}_n)$ maximizes L_n, we can show that $P_n[\dot{l}(\hat{\boldsymbol{\theta}}_n)(h)] = 0$ and $P_0[\dot{l}(\boldsymbol{\theta}_0)(h)] = 0$ for all $h \in H$, which verifies the fourth condition of Van der Vaart and Wellner (1996). Again by the Donsker class argument, we verify the second condition that is $G_n[\dot{l}(\boldsymbol{\theta}_0)]$ converges in distribution to a tight Gaussian process on $l^{\infty}(H)$ with mean 0 and covariance process

$$\mathrm{Cov}(\phi(h), \phi(\tilde{h})) = \mathbf{h}_1^{\mathrm{T}}\boldsymbol{\sigma}_1(\tilde{h}) + \int_0^{\tau} h_2\sigma_2(\tilde{h})\, dA_0,$$

where $\sigma = (\boldsymbol{\sigma}_1, \sigma_2)$ is the information operator in the form of

$$\begin{aligned}\boldsymbol{\sigma}_1(h) = {} & P_0\big[\Delta h_2 E_{(\boldsymbol{\beta}_0,A_0)}\big(\Delta\mathbf{G} - A_0\mathbf{G}e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{G}}\big)\big] \\ & - P_0\left[\int_0^Y h_2\, dA_0 E_{(\boldsymbol{\beta}_0,A_0)}\big(e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{G}}\big) E_{(\boldsymbol{\beta}_0,A_0)}\right. \\ & \qquad \left. \times \big(\Delta\mathbf{G} - A_0(Y)\mathbf{G}e^{\boldsymbol{\beta}_0\mathbf{G}}\big)\right] \\ & + P_0\big[\big(E_{(\boldsymbol{\beta}_0,A_0)}\big(\Delta\mathbf{G} - A_0(Y)\mathbf{G}e^{\boldsymbol{\beta}_0\mathbf{G}}\big)\big)^{\otimes 2}\mathbf{h}_1\big]\end{aligned}$$

and

$$\begin{aligned}\sigma_2(h)(u) = {} & h_2(u)P_0\big[I(Y \geq u)E_{(\boldsymbol{\beta}_0,A_0)}(e^{\boldsymbol{\beta}_0\mathbf{G}})\big] \\ & - \int_0^u h_2(t)\, dA_0(t) P_0\big[I(Y \geq u)E_{(\boldsymbol{\beta}_0,A_0)}(e^{2\boldsymbol{\beta}_0\mathbf{G}})\big] \\ & - P_0\big[\Delta I(Y \geq u)E_{(\boldsymbol{\beta}_0,A_0)}(e^{\boldsymbol{\beta}_0\mathbf{G}})h_2(Y)\big] \\ & + P_0\left[I(Y \geq u)\big(E_{(\boldsymbol{\beta}_0,A_0)}(e^{\boldsymbol{\beta}_0\mathbf{G}})\big)^2 \int_0^Y h_2\, dA_0\right] \\ & + P_0\big[I(Y \geq u)\big(\mathrm{Cov}_{(\boldsymbol{\beta}_0,A_0)} \\ & \qquad \times \big(e^{\boldsymbol{\beta}_0\mathbf{G}}, \Delta\mathbf{h}_1^{\mathrm{T}}\mathbf{G} - A_0\mathbf{h}_1^{\mathrm{T}}\mathbf{G}e^{\boldsymbol{\beta}_0\mathbf{G}}\big)\big)\big].\end{aligned}$$

Applying the Donsker class argument and using the consistency of the NPMLE, we prove that

$$\sup_{h\in H} |G_n\dot{l}(\hat{\boldsymbol{\theta}}_n)(h) - G_n\dot{l}(\boldsymbol{\theta}_0)(h)| \overset{a.s.}{\to} 0,$$

which implies that the first condition holds. By writing σ as the sum of a continuous invertible operator and a compact operator as in Murphy, Rossini, and Van der Vaart (1997) and using the similar argument as in the proof of the identifiability of the model to prove σ is one to one, we can prove that the information operator $\sigma : \mathrm{H} \to \mathrm{H}$ is continuously invertible by Theorem 4.25 of Rudin (1973). Thus, all the four conditions in Theorem 3.3.1 of Van der Vaart and Wellner (1996) are satisfied. Write the inverse of σ as $\tilde{\sigma} = (\tilde{\boldsymbol{\sigma}}_1, \tilde{\sigma}_2)$. Then, $n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in distribution to a tight Gaussian process ψ on $l^{\infty}(\mathrm{H})$ with mean 0 and covariance process

$$\mathrm{Cov}(\psi(h), \psi(h')) = \mathbf{h}_1^{\mathrm{T}}\tilde{\boldsymbol{\sigma}}_1(h') + \int_0^{\tau} h_2\tilde{\sigma}_2(h')\, dA_0.$$

Setting h_2 to 0 and $\mathbf{h}_1$ to each of the unit vectors, $\mathbf{e}_i$, we have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = n^{-1/2}\sum_{i=1}^n \tilde{\mathbf{I}}_0^{-1}\tilde{\mathbf{l}}_0(X_i) + o_P(1), \qquad \text{(B.1)}$$

where $\tilde{\mathbf{I}}_0^{-1} = (\tilde{\boldsymbol{\sigma}}_1(\mathbf{e}_1, 0), \tilde{\boldsymbol{\sigma}}_1(\mathbf{e}_2, 0))$, and $\tilde{\mathbf{l}}_0 = \dot{\mathbf{l}}_1(\boldsymbol{\theta}_0) - \dot{l}_2(\boldsymbol{\theta}_0)(\mathbf{h}^*)$, which is the efficient score function for $\boldsymbol{\beta}$. Here, $\mathbf{h}^*$ is a two-dimensional vector of functions defined by

$$\mathbf{h}^* = -\tilde{\mathbf{I}}_0 \begin{pmatrix} \tilde{\sigma}_2(\mathbf{e}_1, 0) \\ \tilde{\sigma}_2(\mathbf{e}_2, 0) \end{pmatrix}.$$

The matrix $\tilde{\mathbf{I}}_0$ can also be shown to be equal to the covariance matrix of $\tilde{\mathbf{l}}_0$ and is the efficient information matrix for $\boldsymbol{\beta}$. The foregoing result and equation (B.1) imply that $n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ converges in distribution to $N(0, \tilde{\mathbf{I}}_0^{-1})$.

Likelihood-ratio statistics. Theorem 3.1 of Murphy and Van der Vaart (2000), together with the consistency of the NPMLE, yields that

$$\begin{aligned}LR = {} & (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}}\sum_{i=1}^n \tilde{\mathbf{l}}_0(X_i) - \frac{1}{2}n(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^{\mathrm{T}}\tilde{\mathbf{I}}_0(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ & + o_P\big(n^{1/2}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 + 1\big)^2.\end{aligned}$$

In view of the above equation and (B.1) and the fact that $n^{1/2}\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2$ is bounded in probability,

$$LR = n^{-1}\left(\sum_{i=1}^n \tilde{\mathbf{l}}_0^{\mathrm{T}}(X_i)\right)\tilde{\mathbf{I}}_0^{-1}\left(\sum_{i=1}^n \tilde{\mathbf{l}}_0(X_i)\right) + o_P(1), \qquad \text{(B.2)}$$

which implies that LR converges in distribution to a χ^2 distribution with 2 df.