Theor Appl Genet (2011) 122:855–863
DOI 10.1007/s00122-010-1491-6

ORIGINAL PAPER

Statistical optimization of parametric accelerated failure time model for mapping survival trait loci

Zhongze Piao · Xiaojing Zhou · Li Yan · Ying Guo · Runqing Yang · Zhixiang Luo · Daniel R. Prows

Received: 10 March 2010 / Accepted: 4 November 2010 / Published online: 25 November 2010

Abstract Most existing statistical methods for mapping quantitative trait loci (QTL) are not suitable for analyzing survival traits with a skewed distribution and censoring mechanism. As a result, researchers incorporate parametric and semi-parametric models of survival analysis into the framework of the interval mapping for QTL controlling survival traits. In survival analysis, accelerated failure time (AFT) model is considered as a de facto standard and fundamental model for data analysis. Based on AFT model, we propose a parametric approach for mapping survival traits using the EM algorithm to obtain the maximum likelihood estimates of the parameters. Also, with Bayesian information criterion (BIC) as a model selection criterion, an optimal mapping model is constructed by choosing specific error distributions with maximum likelihood and parsimonious parameters. Two real datasets were analyzed by our proposed method for illustration. The results show that among the five commonly used survival distributions, Weibull distribution is the optimal survival function for mapping of heading time in rice, while Log-logistic distribution is the optimal one for hyperoxic acute lung injury.

Introduction

Survival traits, which are usually defined as the length of time between two events, have been observed broadly in nature, such as heading times in plants and failure times in animals. Yet, as survival traits with long right tails do not follow the normal distribution and is often subject to censoring, the existing statistical methods for mapping quantitative trait loci (QTL) have difficulties to analyze such traits appropriately. Some methodologies of survival analysis, including the cure-rate model, parametric and semi-parametric models, have been developed sequentially. In the parametric and semi-parametric models, the Cox

Communicated by M. Sillanpaa.

Z. Piao
Crop Breeding and Cultivation Research Institute, Shanghai Academy of Agricultural Sciences, Shanghai 201106, People's Republic of China

X. Zhou · Y. Guo
Department of Mathematics, Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China

R. Yang
College of Animal Science and Veterinary Medicine, Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China

R. Yang (✉)
School of Agriculture and biology, Shanghai Jiaotong University, Shanghai 200240, People's Republic of China
e-mail: runqingyang@sjtu.edu.cn

L. Yan
College of Information Technology, Heilongjiang Bayi Agricultural University, Daqing 163319, People's Republic of China

Z. Luo
Rice Research Institute, Anhui Academy of Agricultural Sciences, Hefei 230036, People's Republic of China

D. R. Prows
Division of Human Genetics, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA

proportional hazard model (PHM) or the accelerated failure time model (AFT) is the natural choice for regressing phenotypes onto flanking markers when the primary phenotypes belong to survival time (Cheng and Tzeng 2009).

Symons et al. (2002) characterized the QTL effects on the failure time with a PHM and estimated the model parameters and computed LOD scores by a variant of the EM algorithm (Lipsitz and Ibrahim 1998). The PHM with a Weibull baseline hazard function was used to formulate the effects of QTLs on the failure time (Diao et al. 2004). In the case of such a spike in the phenotype distribution, Broman (2003) used a two-part parametric model and a nonparametric approach based on the Kruskal–Wallis test for QTL mapping. Fine et al. (2004) proposed nonparametric estimates for genetic effects of quantitative trait loci, which complemented the rank tests of Kruglyak and Lander (1995). Based on the PHM, Diao and Lin (2005) proposed efficient likelihood-based inference measures and developed semi-parametric statistical methods for mapping survival trait loci. Fang (2006) investigated a simple and efficient approach to estimating QTL parameters through partial likelihood function. In outbred populations, the variance-component based methods of Epstein et al. (2003) or Pankratz et al. (2005) are implemented for mapping QTL of survival traits.

In the survival analysis, the AFT model has an intuitive physical interpretation for real-life examples, as it directly expresses the failure time, rather than the probability as in the PHM, and therefore would be an important alternative to the PHM (Jin et al. 2003; Ma and Bechinski 2009). The AFT model has a simple expression as it relates the logarithm of the failure time linearly to the covariates (Cox and Oakes 1984; Kalbfleisch and Prentice 2002). It can also reduce the potential error amplifications from linking models with different structures. Therefore, in contrast to PHM, the AFT model can be an alternative approach to mapping survival traits. Cheng and Tzeng (2009) proposed parametric and semi-parametric methods based on accelerated failure time models for interval mapping. However, the computation burden of their approach, which is based on the likelihood derived by Diao et al. (2004) to estimate model parameters, is very heavy. Meanwhile, extensive simulations have revealed that the parametric estimators may be more efficient in determining the effect and location of QTL, although parametric estimators may have obvious bias when selecting the incorrect error distribution. In contrast, the semiparametric inference is robust to the error distribution. There is no apparent difference in statistical power of QTL detection between parametric and semiparametric estimations.

Similar to the PHM, the AFT model describes the relationship between survival probabilities and a set of covariates. For each error distribution in AFT model, there is a corresponding survival distribution (Qi 2009). The members of the AFT model class include the exponential AFT model, Weibull AFT model, log-logistic AFT model, log-normal AFT model, and gamma AFT model. So, it is necessary to determine which model is optimal in practical mapping. The objectives of this paper are (1) to formulate a parametric model for mapping survival trait loci based on AFT model and give the EM algorithm of maximum likelihood estimation for QTL parameters within the framework of interval mapping, (2) to demonstrate the effect of baseline survival model selection on mapping survival trait loci via computer simulation experiments, and (3) to optimize the model by selecting a survival function with a higher goodness of fit and parsimonious parameters for two real datasets.

Methods

Genetic model

We only take a BC population as an example to describe our mapping method. However, the method can be easily extended to other experimental populations, such as F_2 intercrosses, recombination inbred lines, and four-way crosses. All n individuals were observed for survival time (T: t_1, t_2, …, t_n) and genotyped for markers with a known linkage map. Assume that a single QTL flanked by any two adjacent markers M_k and M_{k+1} is responsible for the traits of interest and specify that the QTL multiplicatively act on the failure time T or additively, on log$T$; then, the AFT model for mapping survival trait loci can be described by

$$y_i = \mu + z_i a + \sigma\varepsilon_i, \tag{1}$$

where $y_i = \log t_i$ for the ith individual, μ is population mean, z_i is the indicator variable of QTL genotypes (1 for QQ and -1 for Qq), a is the additive effects of QTL, σ is scale parameter, and ε_i is a random error which is assumed to have a particular distribution. Note that the AFT models are named for the distribution of T rather than the distribution of ε_i or log$T$ (Qi 2009).

Maximum likelihood estimation

The survival function of T can be expressed in terms of the survival function of ε_i

$$\begin{aligned} S(t_i) &= P(T \geq t_i) \\ &= P(\log T \geq \log t_i) \\ &= P(\mu + z_i a + \sigma\varepsilon_i \geq \log t_i) \\ &= P\left(\varepsilon_i \geq \frac{\log t_i - \mu - z_i a}{\sigma}\right) \\ &= S_{\varepsilon_i}\left(\frac{\log t_i - \mu - z_i a}{\sigma}\right) \end{aligned} \tag{2}$$

As survival function and density function have the relationship as $f(t) = -S'(t)$, we can obtain the density function for survival time as

$$f(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{\log t_i - \mu - z_i a}{\sigma}\right) \tag{3}$$

where, f_{ε_i} is the density function of the random variable ε_i. Assume that there are some right-censoring records in the observed survival times, let C_i be censoring time for the ith individual and $\Delta_i = I(T_i \le C_i)$, where, $\Delta_i = 1$ when T_i is fully observed (uncensored); otherwise $\Delta_i = 0$. For the ith individual, then, the survival density function can be formulated as

$$\varphi(t_i) = f(t_i)^{\Delta_i} S(t_i)^{1-\Delta_i} \tag{4}$$

All the survival density functions for the uncensored record and the records with different censoring types are given in Appendix. Given the two QTL genotypes, The density function $\varphi(t_i)$ will have two types, denoted as $\varphi(t_i|QQ)$ and $\varphi(t_i|Qq)$, respectively. Let $p_i(QQ)$ and $p_i(Qq)$ be the conditional probabilities of the QTL genotypes given the flanking markers M_k and M_{k+1}; then a mixture model (Broman 2003) can be formed as

$$\varphi(t_i) = p_i(QQ)\varphi(t_i|QQ) + p_i(Qq)\varphi(t_i|Qq) \tag{5}$$

Suppose that the trait values are independent to each other, then the likelihood of the data $t_1, t_2, \ldots, t_n$, is the product of independent mixture models for n individuals. That is, given the survival time (T) and marker information (M),

$$L(\Omega|T,M) = \prod_{i=1}^{n} \varphi(t_i) \tag{6}$$

where the vector $\Omega = (\mu, a, \sigma, \theta, \delta)^{\mathrm{T}}$ with θ being the parameters in baseline hazard function and δ being the scanning position. The log-likelihood is then

$$\begin{aligned}&\log L(\Omega|T,M)\\ &= \sum_{i=1}^{n} \log[p_i(QQ)\varphi(t_i|QQ) + p_i(Qq)\varphi(t_i|Qq)]\end{aligned} \tag{7}$$

The derivative of Eq. (7) is

$$\begin{aligned}&\frac{\partial}{\partial\Omega}\log L(\Omega|T,M)\\ &= \sum_{i=1}^{n} \frac{p_i(QQ)\frac{\partial}{\partial\Omega}\varphi(t_i|QQ) + p_i(Qq)\frac{\partial}{\partial\Omega}\varphi(t_i|Qq)}{p_i(QQ)\varphi(t_i|QQ) + p_i(Qq)\varphi(t_i|Qq)}\\ &= \sum_{i=1}^{n} \frac{p_i(QQ)\varphi(t_i|QQ)\frac{\partial}{\partial\Omega}\log\varphi(t_i|QQ) + p_i(Qq)\varphi(t_i|Qq)\frac{\partial}{\partial\Omega}\log\varphi(t_i|Qq)}{p_i(QQ)\varphi(t_i|QQ) + p_i(Qq)\varphi(t_i|Qq)}\\ &= \sum_{i=1}^{n}\left[p_i^*(QQ)\frac{\partial}{\partial\Omega}\log\varphi(t_i|QQ) + p_i^*(Qq)\frac{\partial}{\partial\Omega}\log\varphi(t_i|Qq)\right],\end{aligned}$$

where

$$\begin{aligned}p_i^*(QQ) &= \frac{p_i(QQ)\varphi(t_i|QQ)}{p_i(QQ)\varphi(t_i|QQ) + p_i(Qq)\varphi(t_i|Qq)}\\ p_i^*(Qq) &= \frac{p_i(Qq)\varphi(t_i|Qq)}{p_i(QQ)\varphi(t_i|QQ) + p_i(Qq)\varphi(t_i|Qq)}\end{aligned} \tag{8}$$

which are the posterior probabilities of QTL genotypes QQ and Qq for the ith individual. We then implement the EM algorithm (Dempster et al. 1977) to solve the maximum likelihood estimations of Ω. The iteration steps are described below:

1. Choose initial values $\Omega^{(0)} = \left(\mu^{(0)}, a^{(0)}, \sigma^{(0)}, \theta^{(0)}, \delta\right)^{\mathrm{T}}$ for $\Omega = (\mu, a, \sigma, \theta, \delta)^{\mathrm{T}}$.
2. Calculate the posterior probabilities $p_i^*(QQ), p_i^*(Qq)$ given the initial values.
3. Solve for $\frac{\partial}{\partial\Omega}\log L(\Omega|T,M) = 0$ to get the estimates of Ω, denoted as $\Omega^{(1)} = \left(\mu^{(1)}, a^{(1)}, \sigma^{(1)}, \theta^{(1)}, \delta\right)^{\mathrm{T}}$. In practical computation, the simplex algorithm implemented with function 'fminsearch' in the MatLab can be used to obtain the solution for the nonlinear and complicated equations.
4. Replace the initial parameters $\Omega^{(0)} = (\mu^{(0)}, a^{(0)}, \sigma^{(0)}, \theta^{(0)}, \delta)^{\mathrm{T}}$ by $\Omega^{(1)} = \left(\mu^{(1)}, a^{(1)}, \sigma^{(1)}, \theta^{(1)}, \delta\right)^{\mathrm{T}}$ and go back to step (2) for the next iteration.
5. Continue the iterations until a criterion of convergence is reached. At the convergence, the values of the parameters are the maximum likelihood (ML) solutions, denoted by $\hat{\Omega} = \left(\hat{\mu}, \hat{a}, \hat{\sigma}, \hat{\theta}, \delta\right)^{\mathrm{T}}$.

Significance test

To test the significance of the QTL effect, a likelihood ratio statistic is used. Substituting the above maximum likelihood estimates to Eq. (6), we first obtain the log-likelihood value under the full model as $L_1\left(\hat{\mu}, \hat{a}, \hat{\sigma}, \hat{\theta}, \delta|T,M\right)$ and then evaluate the log-likelihood function under the null model (reduced model) so that $a = 0$ is used in place of a, denoted by $L_0\left(\hat{\mu}_0, \hat{\sigma}_0, \hat{\theta}_0, \delta|T,M\right)$. Note that $\hat{\mu}_0$, $\hat{\sigma}_0$ and $\hat{\theta}_0$ are different from $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\theta}$ because the former are estimated by maximizing

$$\begin{aligned}\log L(\mu, \sigma, \theta, \delta|T,M) = &\sum_{i=1}^{n} \Delta_i \log\left[\frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{\log t_i - \mu}{\sigma}\right)\right]\\ &+ \sum_{i=1}^{n} (1-\Delta_i)\log\left[S_{\varepsilon_i}\left(\frac{\log t_{iR} - \mu}{\sigma}\right)\right]\end{aligned} \tag{9}$$

the log-likelihood function under the reduced model.

Table 1 The commonly used error distributions and corresponding survival distributions

Error		Survival time	
Error distribution	Density function	Survival distribution	Survival function
Extreme value (1 parameters)	$\exp(\varepsilon - \exp(\varepsilon))$ $(\sigma = 1)$	Exponential	$\exp(-\lambda t)$
Extreme value (2 parameters)	$\exp(\varepsilon - \exp(\varepsilon))$	Weibull	$\exp[-(\lambda t)^{\gamma}]$
Normal	$\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-\varepsilon^2}{2}\right)$	Log-normal	$1 - \Phi(\gamma \log \lambda t)$
Log-gamma	$\frac{\exp(k\varepsilon - \exp(\varepsilon))}{\Gamma(k)}$ $(\sigma = 1)$	Gamma	$1 - \frac{\int_0^{\lambda t} x^{k-1}e^{-x}dx}{\Gamma(k)}$
Logistic	$\frac{\exp(\varepsilon)}{(1+\exp(\varepsilon))^2}$	Log-logistic	$\frac{1}{1+(\lambda t)^{\gamma}}$

The likelihood ratio test statistic is defined as

$$\mathrm{LOD} = -2\log_{10}\frac{L_0\left(\hat{\mu}_0, \hat{\sigma}_0, \hat{\theta}_0, \delta|T, M\right)}{L_1\left(\hat{\mu}, \hat{a}, \hat{\sigma}, \hat{\theta}, \delta|T, M\right)}$$

To determine the significance of the LOD test, we use the critical threshold generated by permutation tests (Churchill and Doerge 1994). By repeatedly shuffling the relationships between marker genotypes and phenotypes, a series of the maximum LODs are calculated, from the distribution of which the critical threshold is obtained.

By the same way, we calculate the LOD statistic at each locus over the genome (by spacing of 1 or 2 cM) and plot the profile for LODs against the linkage map distance. The linkage map position corresponding to a peak of the LOD plot will be determined as the maximum-likelihood estimate of the QTL location.

Model selection

In a parametric model based on AFT model, many distribution functions may be available to describe the error item in model (1). Each error distribution corresponds to a survival distribution. The commonly used error distributions and the corresponding survival distributions are summarized in Table 1. For a practical data set, we should know which distribution function is optimal to model the error distribution. Also, the selection of an optimal error distribution function is crucial to improve the power of interval mapping for survival traits and estimation precision of QTL parameters. In statistics, there are several criteria for model selection although none of them perform the best under all circumstances (Burnham and Anderson 2002). Here, we used BIC (Schwarz 1978) as the model selection criterion of the optimal error distribution function with parsimonious parameters. The BIC is defined as

$$\mathrm{BIC} = -2\ \log\ L\left(\hat{\Omega}|T, M\right) + \mathrm{dimension}(\Omega|T, M)\log(n),$$

where $\hat{\Omega}$ is the maximum likelihood estimation of Ω under the reduced model, $\mathrm{dimension}(\Omega|T, M)$ represents the number of independent parameters under this model, and n is the number of observations. The optimal error distribution function is the one that displays the minimum BIC value.

Simulations

On a single chromosome of length 100 cM, 11 equally spaced markers were simulated for a backcross population with sample size of 150 and 300. A single QTL was placed at position 25 cM between markers 3 and 4. Genetic effect of the QTL was assumed to be 0.10 and 0.15. A scaled parameter σ was assumed to be 0.5. Survival times were generated from the model (1) based on Weibull distribution.

For each case, 100 replicated were simulated. The means and SD of the estimates and statistical powers of QTL detection were recorded. The critical values of the test statistic used to declare statistical significance will vary according to different models. We simulated 500 additional samples under the null model with $a = 0$ based on Weibull distribution to determine the critical values.

We analyzed the simulated data by using mapping models based on five survival distributions used in this study. Parameter estimates and statistical powers of QTL detection obtained with interval mapping were listed in Table 2. In general, the performance of mapping model based on Weibull distribution has a higher accuracy and precision in parameter estimation and statistical power as compared with the other four survival distributions. Gamma and exponential distributions seemed to be able to fit the simulated data well because their shapes are similar to that of Weibull distribution. The estimated precision of parameters and statistical power of QTL detection, as expected, increase with the QTL effect and sample size increased for each survival distribution.

Based on AFT model with Weibull distribution as baseline distribution, we took right censoring proportions to three levels of 10, 20 and 30% under sample size of 300 to investigate the influence of censored records on mapping survival time locus. Results were listed in Table 3. Overall, the censoring remarkably impacted the accuracy and

Table 2 Parameter estimates (standard deviations) and statistical powers obtained with interval mapping based on different survival distributions for the simulated data from Weibull distribution

Sample size	True effect	$a = 0.10$					$a = 0.15$				
	Distribution	Weibull	Gamma	Log-logistic	Log-normal	Exponential	Weibull	Gamma	Log-logistic	Log-normal	Exponential
150	μ	5.01 (0.05)	3.36 (0.44)	4.77 (0.05)	4.71 (0.05)	4.88 (0.05)	4.99 (0.05)	3.30 (0.43)	4.78 (0.05)	4.73 (0.05)	4.89 (0.04)
	a	0.14 (0.02)	0.14 (0.02)	0.16 (0.02)	0.17 (0.02)	0.14 (0.02)	0.18 (0.03)	0.17 (0.03)	0.18 (0.03)	0.19 (0.03)	0.16 (0.03)
	σ or k	0.49 (0.03)	3.68 (0.13)	0.34 (0.03)	0.62 (0.05)	–	0.49 (0.03)	3.70 (0.12)	0.34 (0.03)	0.62 (0.05)	–
	Position	28.26 (7.95)	28.70 (8.47)	29.24 (11.43)	31.48 (12.04)	28.97 (8.48)	26.39 (5.23)	26.64 (6.69)	27.03 (6.00)	28.23 (8.97)	26.89 (6.59)
	Power (%)	43	40	29	23	41	83	81	72	62	80
300	μ	5.00 (0.04)	3.17 (0.27)	4.77 (0.04)	4.71 (0.04)	4.88 (0.03)	4.99 (0.03)	3.12 (0.29)	4.76 (0.04)	4.71 (0.04)	4.88 (0.03)
	a	0.11 (0.02)	0.12 (0.02)	0.13 (0.02)	0.13 (0.02)	0.11 (0.02)	0.16 (0.03)	0.15 (0.03)	0.16 (0.03)	0.16 (0.03)	0.14 (0.03)
	σ or k	0.50 (0.02)	3.73 (0.08)	0.35 (0.02)	0.64 (0.04)	–	0.50 (0.02)	3.74 (0.09)	0.35 (0.02)	0.64 (0.04)	–
	Position	28.75 (7.83)	28.93 (7.37)	28.96 (6.28)	29.31 (7.14)	29.10 (7.46)	25.54 (3.19)	25.88 (3.99)	25.63 (4.05)	25.87 (4.28)	25.82 (4.17)
	Power (%)	77	73	61	52	72	100	97	96	92	96

precision of QTL parameter estimations and the statistical power of QTL detection. As the censoring proportion increased, accuracy and precision of QTL parameter estimations got increased, whereas the statistical power got significantly decreased.

To prove that BIC is reasonable to select the best survival distribution in mapping survival locus, we simulated 1,000 samples based on Weibull distribution, then, respectively, fitted these samples with the five used survival distributions and calculated the corresponding BIC values. As a result, Weibull distribution was chosen as the best model in the proportion of 94% at simple size of 150, whereas in the proportion of 100% at simple size of 300. Likewise, we, respectively, used the four other survival distributions to generate samples and showed that in most cases, BIC is able to correctly choose the survival distribution used to simulate samples. At sample size of 300, BIC can accurately select each used survival distribution, whereas at sample size of 150, the selection success rates are 93% for exponential distribution, 95% for log-normal distribution, 90% for gamma distribution and 92% for log-logistic distribution.

Examples

Heading time of rice

F_{10} recombinant inbred line (RIL) derived from the hybrids of Dasanbyeo (a Korean tongil type rice) × TR22183 (a Chinese japonica variety) had been designed to locate QTL for heading time in rice. Heading times were measured without censoring and 208 SSR and STS markers were genotyped on 162 RILs. A linkage map was constructed on the basis of the RIL population, spanning a total length of 1,437.5 cM.

We, respectively, substitute five commonly used error distributions listed in Table 1 into maximum likelihood estimation for model parameters. By calculating and comparing the BIC values, an optimal survival function can be determined to fit the heading time. It can be seen from Table 4 that the Weibull distribution has the smallest BIC value in all survival distributions and therefore this error distribution is chosen to be fitted into AFT model for mapping heading time.

Under Weibull distribution, the profile for LODs of scanning loci on entire genome is depicted in Fig. 1. The genome-wide empirical critical value is obtained with 1,000 permutation tests and determined as 3.038 at the significant level of 5%. In total, there are four peaks in the LOD profile that exceed the threshold, which provide the evidence of QTL for controlling heading time. These QTLs are mapped on chromosomes 6, 9, 12 and 17, respectively

Table 3 Parameter estimates (standard deviations) and statistical powers obtained with interval mapping based on Weibull distribution under three levels of censoring proportions

Effect	Censoring proportion (%)	μ	α	σ	Position	Power
$\alpha = 0.10$	10	5.06 (0.03)	0.12 (0.02)	0.46 (0.02)	28.11 (6.20)	59
	20	5.12 (0.04)	0.12 (0.02)	0.42 (0.01)	28.45 (7.61)	33
	30	5.18 (0.04)	0.13 (0.03)	0.37 (0.01)	29.05 (7.92)	13
$\alpha = 0.15$	10	5.06 (0.03)	0.15 (0.03)	0.46 (0.02)	24.61 (3.34)	91
	20	5.13 (0.03)	0.16 (0.03)	0.42 (0.02)	25.54 (3.59)	67
	30	5.19 (0.03)	0.18 (0.03)	0.37 (0.02)	27.12 (4.41)	36

Table 4 BIC values under different survival distributions for heading time in rice

Distribution	Log*L*	Number of parameters	BIC
Exponential	−140.2	1	285.494
Weibull	−113.1	2	236.388
Log-normal	−266.8	2	543.788
Gamma	−135.6	2	281.388
Log-logistic	−136.4	2	282.988

(Fig. 1). Table 5 tabulates parameter estimates of the detected QTLs and survival function. Using the estimated parameters, two survival curves for heading times are drawn for each QTL (Fig. 2). From the differences in curve shape between the two genotypes, we can find that all the four detected QTLs display a substantial effect on survival density distribution for heading time in rice, in which the expected value of *qq* genotype is higher than that of *QQ* genotype for the first, third and fourth QTLs, except for the second QTL. Additionally, mapping analysis based on Weibull distribution can detect all QTLs detected with other survival distributions (Results not shown).

Acute lung injury survival time of mice

A mouse model system for mapping QTL of hyperoxic acute lung injury (HALI) survival has been established by crossing sensitive B (C57BL/6J) strain mice and significantly more resistant S (129X1/SvJ) strain mice to HALI mortality (Prows et al. 2007a, b). The F_1 lines were first generated by mating B females to S males (B.S) and S females to B males (S.B). The four possible F_2 crosses were systematically bred through BS × BS, BS × SB, SB × BS, and SB × SB (female F1 listed first) intercross mating schemes, and a total of 840 F_2 mice were phenotyped for survival time in hours and genotyped for 97 polymorphic microsatellite markers distributed over the genome, including the X chromosome. The logarithms of raw survival times were adjusted for the effects of each system environment factor due to dam, sire, and sex.

The BIC values for the survival distributions are listed in Table 6. It can be seen that different survival distributions give notably different BIC values. In general, the Log-logistic distribution has the best performance among all the five distributions, followed by the gamma and Weibull distributions.

Figure 3 plots the profile of LOD test statistics over the genome under Log-logistic distribution. The genome-wide empirical critical threshold for significance declaration is 3.494 at the significant level of 5%, which is obtained by using permutation tests with 1,000 replicates. Four significant QTLs were identified on chromosomes 1, 4, and 15. Parameter estimates of the QTLs under Log-logistic function are listed in Table 7. In the F_2 population, with these parameter estimates, we can draw three survival curves

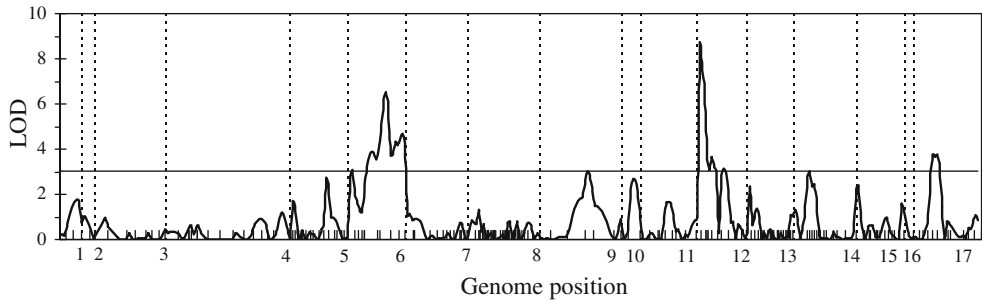

Fig. 1 The profile of LOD test statistics from interval mapping with Weibull distribution for heading time in rice. The genome-wide threshold value is given as a *horizontal* reference line. Linkage groups are separated by the *vertical dotted lines* and marker positions are indicated by the *ticks* on the horizontal axis

Table 5 Parameter estimation of the QTLs detected with Weibull distribution for heading time in rice

QTL no	Chr.- position	Marker interval	μ	σ	a	LOD
1	6-60.1	S03136–RM468	0.083	4.556	−0.051	6.497
2	9-76.7	S06018–S06033	0.091	4.566	0.034	3.056
3	12-5.5	RM25–RM72	0.082	4.561	−0.056	8.657
4	17-29.9	S12038–RM277	0.091	4.552	−0.037	3.753

Table 6 BIC values under different survival distributions for acute lung injury survival time of mice

Distribution	Log$L$	Number of parameters	BIC
Exponential	−874.3	1	1,755.3
Weibull	−236.2	2	485.9
Log-normal	−594.6	2	1,202.7
Gamma	−145.1	2	303.7
Log-logistic	−126.3	2	266.1

corresponding to QTL genotypes (Fig. 4). The comparison of the three survival curves for each QTL indicates that generally all the four detected QTLs lead to the change of survival density, and the differences in the shape of survival density function among the three QTL genotypes are minor for third and fourth QTLs. As compared with other survival distributions, mapping analysis based on Log-logistic distribution also identified more QTLs than those detected using other survival distributions (Results not shown).

Discussion

Based on AFT model, we have developed a parametric approach to interval mapping for survival traits, including the derivation of EM algorithm for maximum likelihood estimation of model parameters using the specified error distributions and the construction of optimal mapping model by model selection procedures. The optimal mapping strategy may not only increase the detecting power of survival trait loci and estimation precision of QTL parameters, but also provides more sensible and interpretable results than semi-parametric approaches. This is due to the choice for appropriate survival distribution and the

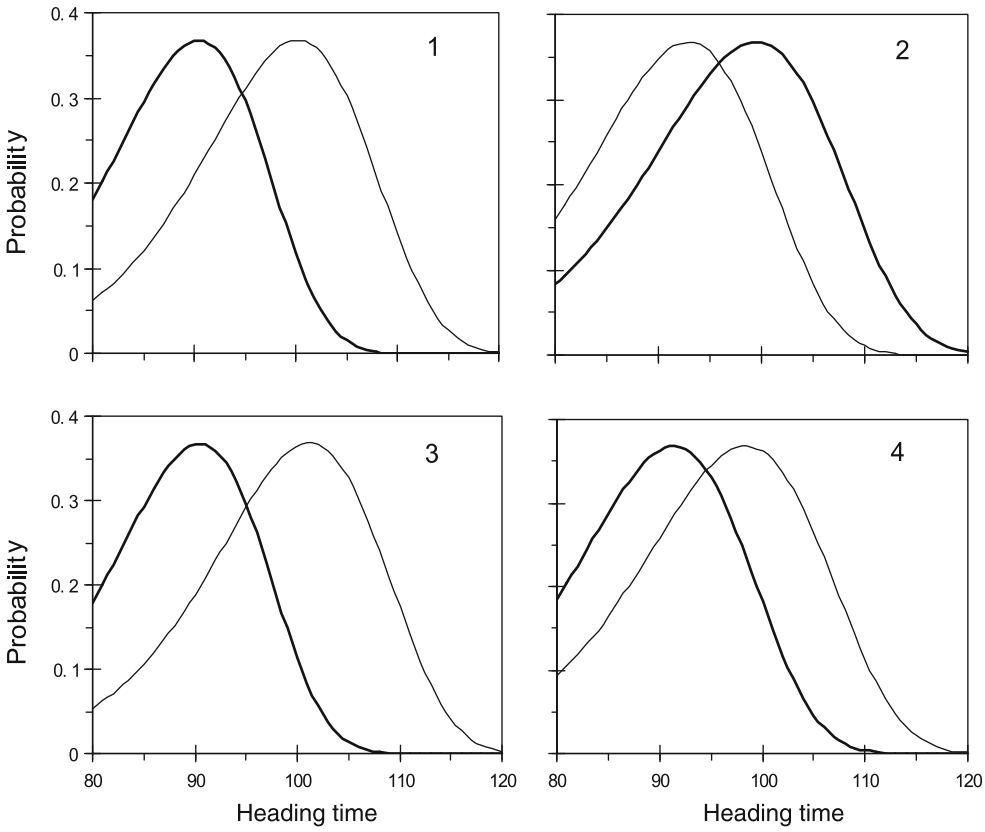

Fig. 2 The survival density curves of two QTL genotypes for 4 detected QTLs (Marked by 1, 2, 3 and 4) drawn according to $f(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{\log t_i - \mu - z_i a}{\sigma}\right)$ with $z_i = 1$ for QQ genotype and $z_i = -1$ for qq genotype. In each plot, the *thick solid* and *thin solid lines* are for QQ and qq genotypes, respectively

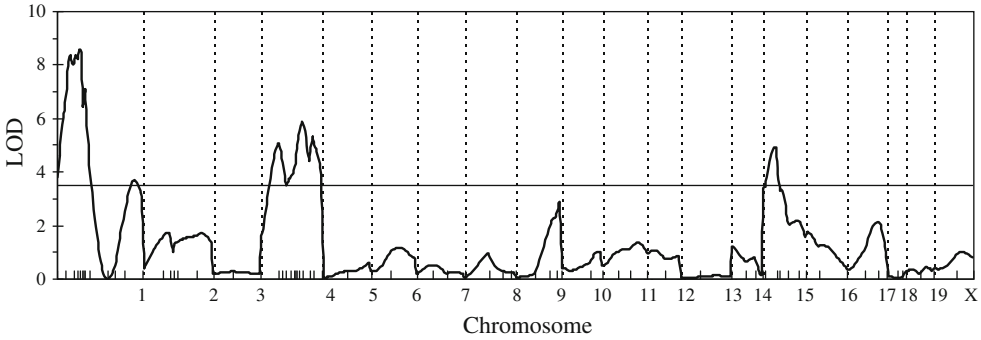

Fig. 3 The profile of LOD test statistics obtained with the interval mapping based on Log-logistic distribution for HALI survival time in mice. The *horizontal reference line* is the empirical critical value. Linkage groups are separated by the *vertical dotted lines,* and marker positions are indicated by the *ticks* on the horizontal axis

Table 7 Parameter estimation of the QTLs detected with Log-logistic distribution for acute lung injury survival time of mice

QTL no	Chr.- position	Marker interval	μ	σ	a	d	LOD
1	1-24.0	D1Mit478–D1Mit214	0.154	4.830	−0.084	−0.011	8.593
2	1-84.0	D1Mit34–D1Mit361	0.155	4.859	0.039	−0.076	3.711
3	4-43.7	D4Mit146–D4Mit308	0.156	4.813	0.067	0.022	5.859
4	15-10.6	D15Mt175–D15Mit5	0.156	4.814	0.065	0.010	4.904

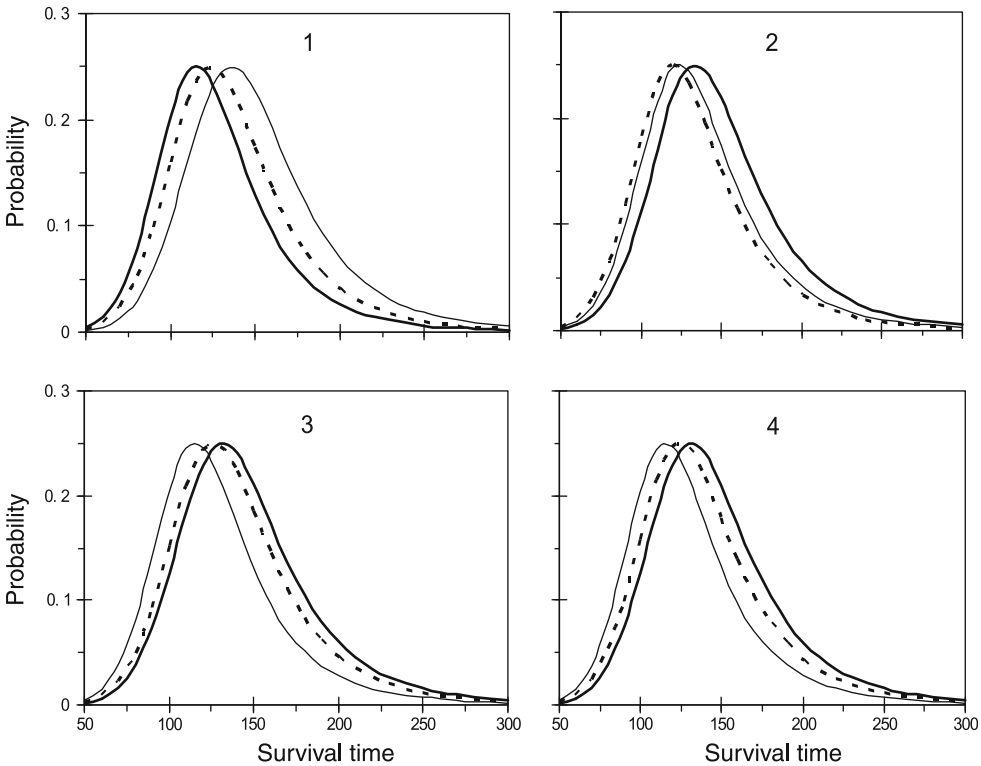

Fig. 4 The survival density curves of three QTL genotypes for four detected QTLs (Marked by 1, 2, 3 and 4) drawn according to $f(t_i) = \frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{\log t_i - \mu - z_i a - w_i d}{\sigma}\right)$. Where, d is dominance effect and $w_i$is indicator variable for dominance effect. $z_i = 1$ and $w_i = 0$ for QQ genotype, $z_i = -1$ and $w_i = 0$ for qq genotype and $z_i = 0$ and $w_i = 1$ for Qq genotype. In each plot, the *thick solid, thin solid,* and *dashed lines* are for QQ, qq, and Qq genotypes, respectively

estimated procedures of survival distribution (Cheng and Tzeng 2009; Diao and Lin 2005; Fang 2006). Survival traits have the properties of non-normal distributions and censoring mechanism because of random loss to follow-up, failures from competing causes, or the limited experimental time. Usually, censoring mechanisms are classified into three types: right censored, interval censored, and left censored (Kalbfleisch and Prentice 2002). In this study, a joint survival density function is used to accommodate the information from censored records. If survival distribution is specified as Log-normal one, then we can treat censored records as missing variable and estimate them with Monte Carlo sampling (Sillanpää and Hoti 2007).

Except for the five commonly used survival distributions adopted in this study, other distributions can also be considered, such as the generalized F and the generalized

gamma distributions. It will depend on goodness of fit to real dataset to determine which distribution is optimal for fitting baseline hazard function. The exponential, Weibull and Log-logistic distributions are often used, because these distributions have closed expressions for tail area probabilities and simple formulas for survivor and hazard functions. Although Log-normal and gamma distributions are generally less convenient in computation, they are still applied frequently. For convenience to apply the method, we have developed a Matlab computer program implementing model selection for mapping survival trait loci. This program can be easily implemented using R/qtl in R software by means of MATLAB R-link package.

Acknowledgments This preparation of work is partially supported by the National Natural Science Foundation of China (30972077) and Key Basic Research Project in Shanghai (10JC1413900). We would like to thank Dr. Annie Lin for her suggestions and helps.

Appendix

Let C_{iL} and C_{iR} be the left and right censoring times, respectively, for the ith subject. The observation on the trait value of the ith subject consists of four possible components: $t_{iL} = \min(T_i, C_{iL})$, $t_{iR} = \max(T_i, C_{iR})$, $\Delta_{iL} = I(T_i > C_{iL})$ and $\Delta_{iR} = I(T_i \le C_{iR})$, where $I(\cdot)$ is the indicator function. For mapping QTL using survival data with the censored records, we should replace the survival density function in the maximum likelihood estimation by a general form:

$$f(t_i)^{I(\cdot)} S(t_i)^{1-I(\cdot)} = \begin{cases} \frac{1}{\sigma t_i} f_{\varepsilon_i}\left(\frac{\log t_i - \mu - z_i a}{\sigma}\right) & \text{for uncensored} \\ 1 - S_{\varepsilon_i}\left(\frac{\log t_{iL} - \mu - z_i a}{\sigma}\right) & \text{for left censored} \\ \left[1 - S_{\varepsilon_i}\left(\frac{\log t_{iL} - \mu - z_i a}{\sigma}\right)\right] S_{\varepsilon_i}\left(\frac{\log t_{iR} - \mu - z_i a}{\sigma}\right) & \text{for interval censored} \\ S_{\varepsilon_i}\left(\frac{\log t_{iR} - \mu - z_i a}{\sigma}\right) & \text{for right censored} \end{cases}$$

References

Broman KW (2003) Mapping quantitative trait loci in the case of a spike in the phenotype distribution. Genetics 163(3):1169–1175

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Spinger, New York

Cheng JY, Tzeng S (2009) Parametric and semiparametric methods for mapping quantitative trait loci. Comput Stat Data Anal 53:1843–1849

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138(3):963–971

Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, London

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

Diao G, Lin DY (2005) Semiparametric methods for mapping quantitative trait loci with censored data. Biometrics 61(3):789–798

Diao G, Lin DY, Zou F (2004) Mapping quantitative trait loci with censored observations. Genetics 168(3):1689–1698

Epstein MP, Lin X, Boehnke M (2003) A tobit variance-component method for linkage analysis of censored trait data. Am J Hum Genet 72(3):611–620

Fang Y (2006) A note on QTL detecting for censored traits. Genet Sel Evol 38(2):221–229

Fine JP, Zou F, Yandell BS (2004) Nonparametric estimation of the effects of quantitative trait loci. Biostatistics 5(4):501–513

Jin Z, Lin DY, Wei LJ, Ying Z (2003) Rank-based inference for the accelerated failure time model. Biometrika 90:341–353

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. Wiley, New York

Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. Genetics 139(3):1421–1428

Lipsitz SR, Ibrahim JG (1998) Estimating equations with incomplete categorical covariates in the Cox model. Biometrics 54(3):1002–1013

Ma ZS, Bechinski EJ (2009) Accelerated failure time (AFT) modeling for the development and survival of Russian wheat aphid, *Diuraphis noxia* (Mordvilko). Popul Ecol 51:543–548

Pankratz VS, de Andrade M, Therneau TM (2005) Random-effects Cox proportional hazards model: general variance components methods for time-to-event data. Genet Epidemiol 28(2):97–109

Prows DR, Hafertepen AP, Gibbons WJ Jr, Winterberg AV, Nick TG (2007a) A genetic mouse model to investigate hyperoxic acute lung injury survival. Physiol Genomics 30(3):262–270

Prows DR, Hafertepen AP, Winterberg AV, Gibbons WJ Jr, Liu C, Nick TG (2007b) Genetic analysis of hyperoxic acute lung injury survival in reciprocal intercross mice. Physiol Genomics 30(3):271–281

Qi JZ (2009) Comparison of proportional hazards and accelerated failure time models. Dissertation, University of Saskatchewan

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

Sillanpää MJ, Hoti F (2007) Mapping quantitative trait loci from a single-tail sample of the phenotype distribution including survival data. Genetics 177(4):2361–2377

Symons RC, Daly MJ, Fridlyand J, Speed TP, Cook WD, Gerondakis S, Harris AW, Foote SJ (2002) Multiple genetic loci modify susceptibility to plasmacytoma-related morbidity in E(mu)-v-abl transgenic mice. Proc Natl Acad Sci USA 99(17):11299–11304