Genet. Res., Camb. (2005), **85**, *pp.* 139–149. *With* 4 *figures.* © 2005 Cambridge University Press
doi:10.1017/S0016672305007366 Printed in the United Kingdom

Interval mapping methods for detecting QTL affecting survival and time-to-event phenotypes

C. R. MORENO[1]*, J. M. ELSEN[1], P. LE ROY[2] AND V. DUCROCQ[2]
[1] *INRA, Station d'Amélioration Génétique des Animaux, BP27, 31326 Castanet-Tolosan Cedex, France*
[2] *INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy-en-Josas, France*

(*Received 4 February 2004 and in revised form 22 September 2004 and 5 January 2005*)

Summary

Quantitative trait loci (QTL) are usually searched for using classical interval mapping methods which assume that the trait of interest follows a normal distribution. However, these methods cannot take into account features of most survival data such as a non-normal distribution and the presence of censored data. We propose two new QTL detection approaches which allow the consideration of censored data. One interval mapping method uses a Weibull model (W), which is popular in parametrical modelling of survival traits, and the other uses a Cox model (C), which avoids making any assumption on the trait distribution. Data were simulated following the structure of a published experiment. Using simulated data, we compare W, C and a classical interval mapping method using a Gaussian model on uncensored data (G) or on all data (G′=censored data analysed as though records were uncensored). An adequate mathematical transformation was used for all parametric methods (G, G′ and W). When data were not censored, the four methods gave similar results. However, when some data were censored, the power of QTL detection and accuracy of QTL location and of estimation of QTL effects for G decreased considerably with censoring, particularly when censoring was at a fixed date. This decrease with censoring was observed also with G′, but it was less severe. Censoring had a negligible effect on results obtained with the W and C methods.

1. Introduction

QTL (Quantitative trait loci) detection methods are used to search for chromosomal regions having an effect on traits of interest. This type of analysis has two main aims. First, information on markers linked to a QTL can be considered in selection programmes (Boichard *et al.*, 2000). From a more fundamental point of view, detected chromosomal regions can be used to search for gene(s) involved in the biological mechanisms influencing the trait under study.

Classical QTL interval mapping methods assume that traits follow a normal distribution (Lander & Botstein, 1989; Knott *et al.*, 1996; Elsen *et al.*, 1999; etc). However, traits in animals and plants are often non-normally distributed. For example, categorical data (e.g. dead or alive) and survival data (e.g. length of life) are often recorded to describe resistance to diseases. For such traits, classical QTL detection methods have a low power and a bias in estimates of effects and of position of the QTL. Interval mapping methods have been proposed to analyse discrete traits (Kadarmideen *et al.*, 2000), but these do not apply to survival data.

Survival data are positive random variables (called failure time hereafter) describing in some sense the length of the interval between a point of origin and an end-point. Survival analysis takes into account distributional forms (often far from normal distribution) and censoring (i.e. the fact that the end-point is not observed for a part of the data). When using classical interval mapping methods, either censored data are excluded from the analysis or they are incorrectly considered as not censored. To estimate fixed effects, proportional hazard models are classically used in survival analysis. They can be either parametric such as the Weibull regression model (Kalbfleisch & Prentice, 1980) or semi-parametric such as in the Cox model (Cox, 1972). In a recent study, Diao *et al.* (2004) proposed use of a Weibull model to search for QTL with an interval mapping method. In the present

* Corresponding author. e-mail: moreno@toulouse.inra.fr

paper, both Weibull and Cox models were used to search for QTL with an interval mapping method. In order to compare these methods with each other and with the classical method assuming a normal distribution, experimental data from an F2 population (Sebastiani *et al.*, 1998) were used to produce simulated data where QTL effects and percentage of censored data were variable.

2. Model definitions

In this section, the QTL detection methods are developed for inbred crosses and they are generalized to outbred crosses in the appendix. Methods are presented in two parts. First, the general form of the likelihood and the expression for the contribution of one observation to the likelihood are described for an interval mapping method using a Gaussian model. With this method and data that conform to a normal distribution, it must be emphasized that only uncensored data can be legitimately included. Therefore, censored data were excluded from the analysis (G) or were incorrectly viewed as uncensored (G′). Second, the new interval mapping methods using a Weibull model (W) and a Cox model (C) are presented. In the latter methods, both uncensored and censored data were included.

(i) *General expression for the likelihood*

In an F2 population, assuming that individuals are produced by heterozygote parents, each animal k has four possible QTL genotypes (1,1), (1,2), (2,1), (2,2), denoted as $g=1, \ldots, 4$. As described by Lander & Botstein (1989), the general form of the likelihood at a chromosomal location z can be written (supposing that observations are mutually independent) as:

$$\Delta^z = \prod_k \left\{ \sum_g p(d_k^z = g|M_k) \cdot l(k|g) \right\} \qquad (1)$$

where d_k^z is a random variable and $p(d_k^z = g|M_k)$ is the probability that individual k has genotype g conditional on its flanking marker information. The contribution to the likelihood $l(k|g)$ of the observation on individual k depends on the assumed distribution of the trait (t_k). Let Ω ($\Omega = 1, \ldots, N$) represent the set of uncensored ($\Omega_{unc} = 1, \ldots, N_{unc}$) and censored observations k ($\Omega_{cens} = N_{unc}+1, \ldots, N$): $\Omega = \Omega_{cens} \cup \Omega_{unc}$.

In the remainder, we consider three alternatives corresponding to the Gaussian, Weibull or Cox models.

(ii) *Interval mapping method using a Gaussian model: G and G′*

In (1), the contribution $l(k|g)$ of individual k with genotype g to the log-likelihood, using a classical interval mapping method (Lander & Botstein, 1989), can be easily written only for an uncensored observation: $k \in \Omega_{unc}$. Thus, the contribution to the likelihood is:

$$l(k \in \Omega_{unc}|g) = \frac{1}{\sqrt{2\pi}\sigma} \times \exp\left[-\frac{1}{2}\left(\frac{t_k - \mu - \mathbf{x}_k'\boldsymbol{\beta} - qtl_g}{\sigma} \right)^2 \right] \qquad (2)$$

where t_k is the trait (failure time) of individual k, μ is the mean, σ is the standard deviation, $\boldsymbol{\beta}$ is the ($n_c \times 1$) vector of covariate effects, n_c is the number of levels of covariate effects, $\mathbf{x}_k'$ is the kth row of the (N_{unc}, n_c) incidence matrix $\mathbf{X}$, and the QTL effect, qtl_g, is equal to $-a$ if $g=1$, d if $g=2$ or 3 and a if $g=4$, where a and d are additive and dominance effects, respectively.

We distinguished between two different Gaussian approaches: G considered only uncensored information in the likelihood ($k \in \Omega_{unc}$) – i.e. censored records are deleted – and G′ included all information ($k \in \Omega$), so the censored observations were assumed to be uncensored at censoring time.

(iii) *Interval mapping methods using Weibull and Cox survival models: W and C*

Survival analyses allow censored observations to be considered properly. Generally, a random (i.e. non-informative) censoring is assumed (Kalbfleisch & Prentice, 1980). Some useful definitions of functions are recalled here. Let t represent the actual failure time, $f(t)$ the density function, $S(t)$ the survivor function and $h(t)$ the hazard function representing the rate at which failure occurs at time t (Kalbfleisch & Prentice, 1980).

To associate covariate effects and hazard function, proportional hazard models are the most popular. These models postulate that the hazard function of an individual k is equal to the product of a baseline hazard function ($h_0(t)$) and a positive function of the covariates ($\exp(\mathbf{x}_k'\boldsymbol{\beta})$). Two families of proportional hazard models can be used. The first family comprises parametric models which use a parametric baseline function, such as the two-parameter Weibull hazard function. The second family includes semi-parametric models, such as the Cox model. Semi-parametric models require no assumption on the functional form of the baseline hazard function, so they are more flexible than parametric models.

In a parametric Weibull regression model, the hazard function is:

$$h(t_k) = \lambda\rho(\lambda t_k)^{\rho-1} \cdot \exp(\mathbf{x}_k\boldsymbol{\beta}),$$

where λ and ρ are positive Weibull parameters.

The contribution to the likelihood of an uncensored observation k ($k \in \Omega_{unc}$) is the density function at

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

failure time which can be written as the product of the hazard function and the survivor function. The contribution to the likelihood of a censored observation k ($k \in \Omega_{cens}$) is the value of the survivor function at censoring time $S(t_k)$ (Kalbfleisch & Prentice, 1980). Then the likelihood can be written:

$$\mathrm{L} = \prod_k [h(t_k)]^{\delta_k} \times [S(t_k)] \tag{3}$$

where $\delta_k = 1$ if $k \in \Omega_{unc}$ and $\delta_k = 0$ if $k \in \Omega_{cens}$.

Therefore, the contribution of the individual k with genotype g to the general form of the likelihood in the interval mapping method (expression 1) using a Weibull model (W) can be written as:

$$\begin{aligned} l(k \in \Omega|g) &\propto [h(t_k)]^{\delta_k} \times [S(t_k)] \\ &\propto \left[\rho \cdot t_k^{\rho-1}(\exp(\rho \log \lambda + \mathbf{x}'_k\boldsymbol{\beta} + qtl_g))\right]^{\delta_k} \\ &\times \exp\left[-t_k^{\rho}(\exp(\rho \log \lambda + \mathbf{x}'_k\boldsymbol{\beta} + qtl_g))\right] \end{aligned} \tag{4}$$

where $\delta_k = 1$ if $k \in \Omega_{unc}$ and $\delta_k = 0$ if $k \in \Omega_{cens}$, t_k is the failure time or censoring time of the individual k.

The Cox model allows estimation of the regression coefficients in $\boldsymbol{\beta}$ without making any assumption about the form of the baseline hazard function. The procedure developed by Cox (1972) to estimate covariate effects assumes no tie (i.e. all failure times are distinct) and relies on the definition of what he calls a partial likelihood function which is the part of the full likelihood function that does not depend on the baseline hazard function. In this expression, only uncensored observations have a non-zero contribution, and censored observations participate in the denominator of the contribution expression. When there are few ties, Peto (in the discussion of Cox, 1972) proposed an approximation, which is an expression for the exact likelihood function when the baseline hazard function is assumed to be piecewise constant (i.e. constant over each interval, after partitioning the time axes into intervals with bounds equal to observed failure times). Peto's version of the Cox's partial likelihood is:

$$\mathrm{L} = \prod_{k \in \Omega_{unc}} \left[\frac{\exp(\mathbf{x}'_k \beta)}{\sum\limits_{k_d \in R(t_k)} \exp(\mathbf{x}'_{k_d}\boldsymbol{\beta})} \right] \tag{5}$$

where $\mathrm{R}(t_k)$ is the set of censored or uncensored individuals k_d at risk at time t_k, i.e. the set of individuals known to be alive just prior to t_k. The product is over all uncensored observations rather than over all distinct failure times as in the Cox (1972) procedure.

In the interval mapping model, there are four terms for each individual k_d, one for each possible genotype. Then to obtain an expression equivalent to (5), the terms in the denominator must be weighted by the probability that individual k has genotype g conditional on its marker information ($p(d^z_{k_d} = g_d|M_{k_d})$). By analogy with (5), the contribution of individual k to the likelihood in interval mapping using a Cox model (C) can be written as:

$$l(k \in \Omega_{unc}|g) \approx \frac{(\exp(\mathbf{x}'_k\boldsymbol{\beta} + qtl_g))}{\left[\left(\sum\limits_{k_d \in R(t_k)} \sum\limits_{g_d} p(d^z_{k_d} = g_d|M_{k_d}) \cdot \exp(\mathbf{x}'_{k_d}\boldsymbol{\beta} + qtl_{g_d})\right)\right]} \tag{6}$$

where $R(t_k)$ is the list of individuals at risk at time t_k. The subscript d is used to identify terms k and g which come from the sum of the denominator of the function (6).

3. Data and simulations

(i) *Experimental design*

Data were simulated following the structure of a published experiment (Sebastiani *et al.*, 1998). One hundred and ninety-one F2 animals were produced using two inbred mouse lines. Their survival times were measured after inoculation with a pathogenic bacterium, *Salmonella typhimurium*. All animals died at the end of the experiment, so no data were censored. Sebastiani *et al.* (1998) used two approaches to search for QTL. In the first approach, data were log-transformed and analysed using an interval mapping method assuming a normal distribution. In the second approach, a Cox regression model was used to test the marker effects. When using these two methods, they found QTLs located in similar regions and having similar effects.

Here, data were simulated based on the survivor distribution and marker genotypes observed in this F2 population, in order to compare the results obtained with the four different interval mapping methods previously presented: G, G′, W and C.

The failure time distribution of this design (Fig. 1) was used as the basal survival data. A QTL effect and a censoring process were added to this basal distribution as described in the following section. Marker genotypes for chromosome 1 were used. This chromosome had the longest typed region (Fig. 2) and a mean percentage of missing genotype by marker equal to 9%.

Simulations were carried out either under the null hypothesis (no segregating QTL) or under the H1 hypothesis of one segregating QTL. Data were censored in two ways: at a fixed date or at random dates. Censoring at a fixed date mimics censoring at the end of the experiment. The censoring at random dates allows consideration of, for example, censoring due to the death of an animal not related to the disease under study or different starting dates (e.g. birth dates or

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

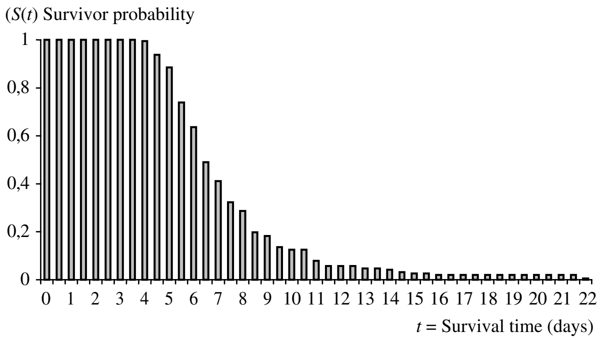

Fig. 1. Survivor distribution of the experimental data used in the simulation process.

inoculation dates). Five scenarios of censoring were considered: uncensored data, 20% and 40% of censored records at random dates, and 20% and 40% of censored records at a fixed date. Under the H0 hypothesis, 1000 simulations were performed for the five types of censoring. Under the H1 hypothesis, a single QTL was assumed at 43·5 cM on the chromosome. The QTL was given an additive effect '*add*' of 0·3 or 0·5 and either no dominance ($dom=0$) or complete dominance ($dom=add$). Then for each scenario of censoring, four situations with different values of dominance and additive effects were considered (500 simulations each time). Therefore, under H1, a total number of 20 situations were studied.

(ii) *Simulation process*

Simulated data were generated using values of *uncensored failure times* of the experimental design (Sebastiani *et al*., 1998). For an easier presentation, let t_k ($k=1, \ldots, n$) be the *observed failure times* and $T_{[1]}<T_{[2]}<\ldots<T_{[i]}<\ldots<T_{[m]}$ be the *ordered distinct failure times* ($m \leqslant n$).

First, the survival function $S(t)$, which is the probability of being alive at time t, was estimated using the Kaplan–Meier estimator (Kaplan & Meier, 1958):

$$\hat{S}_{KM}(t)=\prod_{i/T_{[i]}<t}\left(\frac{\gamma_{[i]}-\alpha_{[i]}}{\gamma_{[i]}}\right) \tag{7}$$

where $\gamma_{[i]}$ is the number of animals known to be alive just prior to time $T_{[i]}$, and $\alpha_{[i]}$ is the total number of animals dying at time $T_{[i]}$. This estimate of $S_{KM}(t)$ was considered to be the baseline survival function: $S_0(t)=\hat{S}_{KM}(t)$.

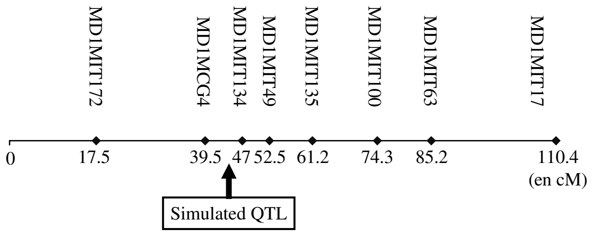

Fig. 2. Chromosomal location of markers on chromosome 1. Cumulative map distances come from the consensus map in the Mouse Genome Database (http://www.informatics.jax.org).

Considering a proportional hazard model, this baseline survival function was used to build the survivor function $S(t|g)$ conditional on each QTL genotype (g). Additive and dominance effects (*add* and *dom*) of the simulated QTL were then included as effects in the proportional hazard model, as:

$$\begin{aligned}
&S(t|g=1)=S_0(t)\\
&S(t|g=2)=S(t|g=3)=[S_0(t)]^{\exp(add+dom)}\\
&S(t|g=4)=[S_0(t)]^{\exp(2add)}
\end{aligned} \tag{8}$$

The generation of simulated records was realized in two steps: first, the choice of a QTL genotype and, second, the choice of a survival time value. For each animal, the probability of the four QTL genotypes was calculated conditional on flanking marker information: $p(d_k^z=g|M_k)$ (see 1). Using the three probabilities $p(d_k^z=1|M_k)$, $p(d_k^z=2,3|M_k)$ and $p(d_k^z=4|M_k)$, a QTL genotype was drawn from a trinomial distribution. The simulated record was then generated using the inverse-transform method (Law & Kelton, 1982). Knowing the genotype g, an ordinate U of the survival function value was drawn from a [0,1] uniform distribution. The observed survival time t_k was obtained as the value such that $S(t_k)=U$, i.e. $t_k=S^{-1}(U)$. In almost all cases, the simulated U values did not correspond exactly to an originally observed value of $S(.)$, since $S(.)$ is based on the estimates of survivor function at a specific date. Therefore, to obtain realistic t_k values, a linear interpolation between the original observed t_k values or an extrapolation beyond the smallest t_k value of $S(t_k|g)$ was applied. This approach allowed the generation of simulated records from a realistic survivor distribution without any particular assumption on the true parameter distribution.

(iii) *Definition of QTL effects*

Standardized QTL effects are classically used in simulation studies. For example, additive and dominance simulated effects equal to a fraction of the phenotypic standard deviation are chosen. Here, the use of a realistic but non-standard survivor distribution as a baseline distribution does not allow simulation of such standardized QTL effects. In the case of W or C, the value of additive and dominance simulated effects previously defined (*add* and *dom*) can be compared with the QTL estimated effects, called $\hat{a}$

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

and $\hat{d}$ in the model definition section. Unfortunately, such a direct comparison is not possible under G or G′. The only way to interpret estimates of QTL effects under G or G′ is to compare these to estimates under W. Because the Weibull model can be described as a log-linear model with a residual proportional to an extreme value distribution (Kalbfleisch & Prentice, 1980), standardized QTL effects under G and G′, $-\hat{a}/\hat{\sigma}$, were compared with standardized QTL effects under W, $\hat{a}/\hat{\rho}$.

In other words, $add=0{\cdot}5$ and $dom=0{\cdot}5$ define effects on a non-specific scale, not related to the trait variability.

(iv) *Censoring process*

Fixed or random date censoring was applied to the simulated data. When a rate of $v\%$ of censoring at a fixed date was chosen, the $v\%$ largest failure times were censored and the censoring time was set equal to the largest uncensored time. When a rate of $v\%$ of censoring at random dates was applied, records were randomly drawn from a binomial distribution ($p=v\%$) and the censoring time for record k was drawn from a $[4, t_k]$ uniform distribution (4 days being the smallest observed survival time).

(v) *Computational techniques*

Simulated data were transformed to perform the analysis with G, G′ and W. With G and G′, a logarithmic transformation was used to partly normalize the data. With W, a translation of the data was necessary because there were no failure observations between days 0 and 4 (Cox & Oakes, 1984). To choose the translation (($t \rightarrow t^* = t-\tau$) to obtain an approximate Weibull distribution, several transformations were applied. A graphical test of the adequacy of a Weibull distribution was performed. This test consists of checking whether a plot of log $[-\log(S(t))]$ against $\log(t)$ gives a straight line. The best transformation was found to be $t^*=t-3{\cdot}9$ for failure time (Fig. 3).

The likelihood function was maximized using a quasi-Newton algorithm implemented as a NAG subroutine (E04JYF) for the three methods. The empirical distribution of the likelihood ratio test statistic was generated in the same manner for each censoring situation under the null hypothesis. A significance level of 0·95 was chosen for all analyses. The empirical threshold value was defined as the 95th percentile of the empirical distribution of the likelihood ratio test statistic under H0. Under H1, the power was defined as the percentage of replicates in which the null hypothesis was rejected at the 5% significant level. The difference between two power estimates is significant (at 5%) if this difference is higher than $1.96\sqrt{u(1-u)\cdot\left(\frac{1}{m_1}+\frac{1}{m_2}\right)}$, where u is the proportion of runs above the significance threshold pooled across methods, and m_1 and m_2 are the number of runs for each method (Baret *et al.*, 1998). Powers, mean estimates of the additive effect, dominance effect and location of QTL were calculated based on the maximum likelihood estimates of all 500 H1 simulations whatever the approach used (G, G′, C or W).

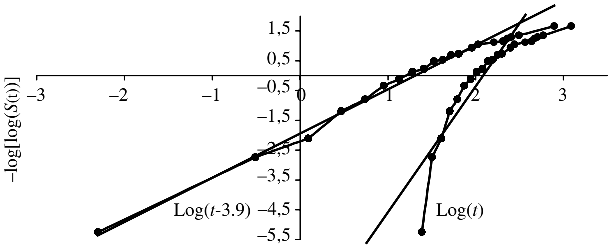

Fig. 3. Graphical test of the Weibull assumption using $t=0$ and $-3{\cdot}9$ as the origin (a straight line is synonymous with a good Weibull fit; Cox & Oakes, 1984).

4. Results

The G, G′, W and C methods were compared by considering their power, and their estimates of the additive effect, dominant effect and location of QTL.

(i) *Comparison of QTL detection power*

In Fig. 4, the differences in power among the four approaches are presented as functions of the QTL effects for the five situations of censoring. Whatever the censoring situation and the values of simulated QTL effects, the difference between the power of C and W never exceeds 6%. All these differences are non-significant, except in two cases where they are significant but weakly so.

When there is no censoring or censoring at random dates, there is no significant difference between the power of G and G′, but these approaches are less powerful than C and W. The difference between the power of the survival approaches and the Gaussian approaches is significant in 6 of 24 situations considered under no censoring or censoring at random dates. When censoring is applied at a fixed date, C and W are clearly much more powerful than G and G′ for 40% of censoring, G being always the least powerful approach. This trend increases dramatically with the rate of censoring at a fixed date and the value of the dominance QTL effect. All differences between the power of the Gaussian and the survival approaches are significant. In the extreme case of 40% of censoring at a fixed date and additive and dominance effects equal to 0·5, the difference in power of C is 86% with G and 24% with G′.

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

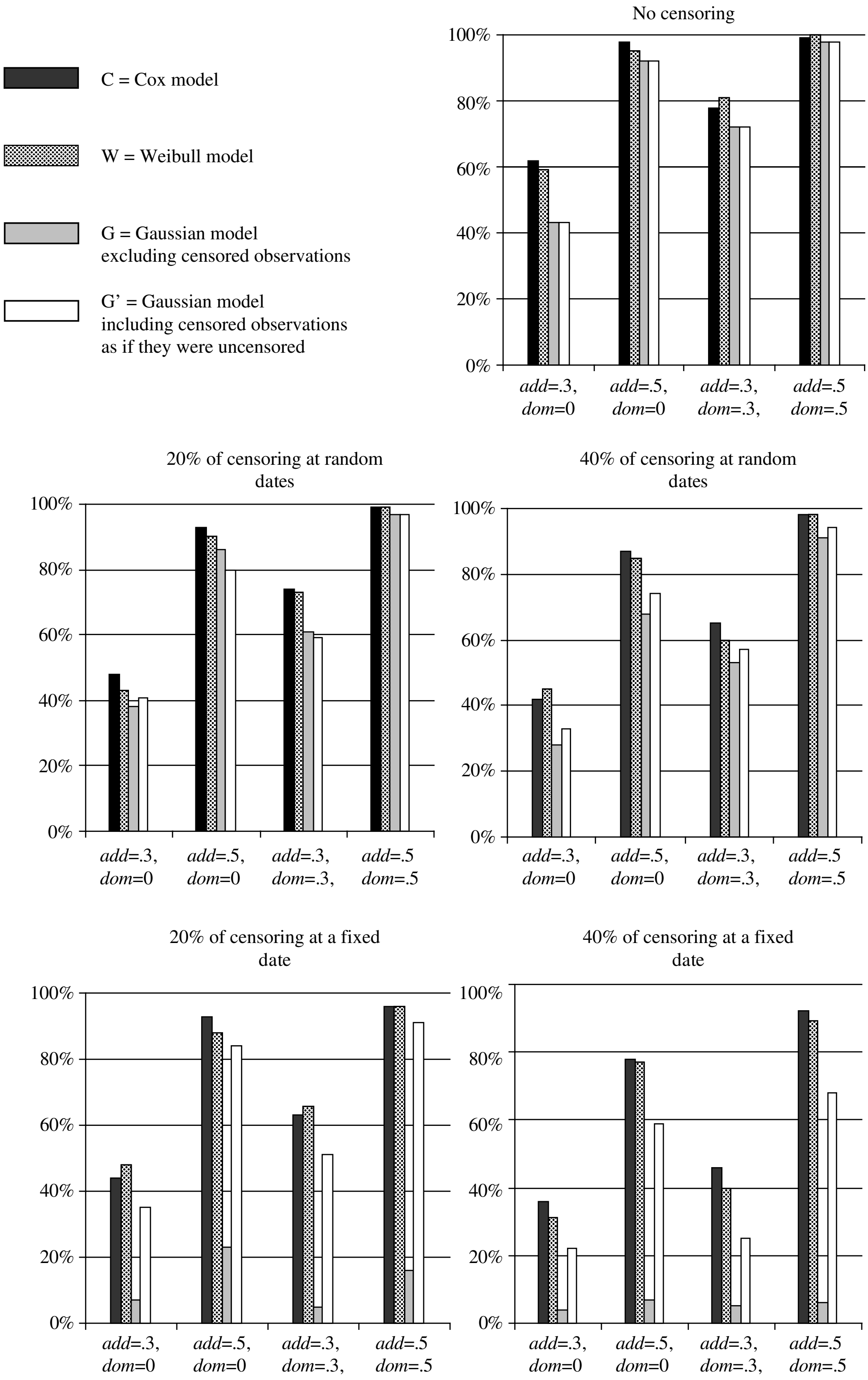

Fig. 4. QTL detection power with G, G′, W and C methods, as a function of simulated QTL effects for the five situations of censoring (the QTL has an additive effect *add* and a dominance effect *dom*).

(ii) *Comparison of QTL location estimated*

Table 1 presents means and standard deviations of QTL location for all situations simulated. Most estimated locations tended to be biased towards the centre of the chromosome. This observation is classical in interval mapping analysis (Walling *et al.*, 2002). For W and C, this bias and the accuracy of QTL location were barely influenced when censoring rate increased. However, the bias decreased and the

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

Table 1. *Mean estimates (± empirical standard deviations over 500 replicates) of QTL location for the simulated situations (true QTL location = 43·5 cM) with methods using a Gaussian model excluding censored observations (G), a Gaussian model including censored observations as though they were uncensored (G′), a Cox model (C) and a Weibull model (W)*

Model	*add*	*dom*	C	W	G	G′
			QTL location (cM)			
No censoring	0·3	0	49 ± 19	47 ± 22	49 ± 21	–
	0·5	0	43 ± 10	42 ± 11	43 ± 12	–
	0·3	0·3	45 ± 15	45 ± 16	45 ± 15	–
	0·5	0·5	43 ± 6	43 ± 7	41 ± 8	–
Random censoring = 20 %	0·3	0	50 ± 21	50 ± 23	49 ± 21	49 ± 23
	0·5	0	44 ± 11	42 ± 12	43 ± 13	45 ± 15
	0·3	0·3	46 ± 15	46 ± 17	46 ± 16	47 ± 18
	0·5	0·5	43 ± 8	42 ± 8	42 ± 9	42 ± 10
Random censoring = 40 %	0·3	0	52 ± 23	49 ± 23	53 ± 26	53 ± 26
	0·5	0	44 ± 13	43 ± 13	44 ± 16	47 ± 18
	0·3	0·3	46 ± 17	47 ± 19	46 ± 19	50 ± 22
	0·5	0·5	44 ± 10	42 ± 10	42 ± 10	44 ± 14
Fixed date censoring = 20 %	0·3	0	48 ± 20	51 ± 22	59 ± 29	52 ± 25
	0·5	0	44 ± 12	43 ± 12	56 ± 27	45 ± 14
	0·3	0·3	48 ± 17	48 ± 18	60 ± 28	49 ± 22
	0·5	0·5	44 ± 9	43 ± 8	59 ± 27	44 ± 11
Fixed date censoring = 40 %	0·3	0	52 ± 23	51 ± 23	62 ± 31	57 ± 27
	0·5	0	45 ± 15	45 ± 14	62 ± 29	46 ± 17
	0·3	0·3	50 ± 20	51 ± 21	63 ± 31	54 ± 23
	0·5	0·5	45 ± 12	45 ± 11	63 ± 30	47 ± 16

(–), when there is no censoring G and G′ models are identical.

accuracy increased when QTL effects increased. A slightly more pronounced trend was observed for G′. Similar results were obtained with G when no censoring or censoring at random dates was applied. However, with censoring at a fixed date, the accuracy of G decreased and its bias increased with the proportion of censoring, particularly for situations where a dominance QTL effect was simulated.

(iii) *Comparisons of additive and dominance QTL effects estimated*

Table 2 presents the means and standard deviations of the estimated additive QTL effects for the 20 simulated situations. Results on dominance effects are not presented in Table 2 because they followed the same trends as the estimates of the additive QTL effect. With W and C, estimates of QTL effects were similar between the different censoring situations. The estimates were only slightly biased or were unbiased.

Comparing the standardized estimates (see Section 3.iii) G or G′ and W, the values were slightly underestimated for G and G′ when censoring was not applied or was at random dates (showing that the standardized solutions in the absence of censoring are relatively consistent). A different situation was found when censoring was at a fixed date: the G estimates of the effects were 2 or 3 times smaller than the W estimates. The accuracy was also considerably affected. On the other hand, censoring at a fixed date did not greatly affect the G′ estimates.

5. Discussion

(i) *Selection of the model*

If an adequate transformation is used in the parametric models (logarithmic transformation in G or G′ and translation transformation in W), the differences between models did not appear critical when censoring was not applied or was at random, at least in the example considered here. In these situations, even though survival approaches were slightly better than the Gaussian approaches, all methods gave quite similar results, in terms of detection power, accuracy and bias of the estimates. When censoring at a fixed date was applied to the data, the situation changed. The G approach was strongly affected by censoring at a fixed date. This effect increased when there was a dominant QTL effect. In the latter case, extreme data – which were censored – were the most informative for estimating QTL effects. In the situations where censoring is at a a fixed date (for example due to the end of the study) a classical method such as G is not at all adequate. In G′, regarding censored data

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

Table 2. *Mean estimates* (± *empirical standard deviations over 500 replicates*) *of additive QTL effects with methods using a Gaussian model excluding censored observations* (*G*), *a Gaussian model including censored observations as though they were uncensored* (*G′*), *a Weibull model* (*W*) *and a Cox model* (*C*)

			Estimated additive QTL effect		Standardized estimated additive QTL effect		
			$\hat{a}$	$\hat{a}$	$\hat{a}/\hat{\rho}$	$-\hat{a}/\hat{\sigma}$	$-\hat{a}/\hat{\sigma}$
			C	W	W	G	G′
Model	*add*	*dom*	(1)	(2)	(3)	(4)	(5)
No censoring	0·3	0	0·35±0·13	0·39±0·18	0·27±0. 12	0·29±0·13	–
	0·5	0	0·55±0·11	0·60±0·14	0·39±0·09	0·47±0·10	–
	0·3	0·3	0·34±0·14	0·38±0·16	0·25±0·10	0·32±0·13	–
	0·5	0·5	0·54±0·12	0·61±0·14	0·39±0·09	0·50±0·10	–
Random censoring = 20 %	0·3	0	0·34±0·16	0·39±0·21	0·26±0·13	0·30±0·14	0·25±0·13
	0·5	0	0·57±0·13	0·63±0·16	0·40±0·10	0·47±0·11	0·40±0·11
	0·3	0·3	0·36±0·14	0·40±0·18	0·25±0·11	0·33±0·14	0·28±0·14
	0·5	0·5	0·57±0·14	0·61±0·16	0·38±0·10	0·50±0·12	0·43±0·11
Random censoring = 40 %	0·3	0	0·37±0·18	0·43±0·22	0·27±0·14	0·30±0·18	0·23±0·14
	0·5	0	0·61±0·16	0·66±0·18	0·40±0·11	0·47±0·15	0·36±0·11
	0·3	0·3	0·39±0·17	0·43±0·21	0·26±0·13	0·33±0·18	0·25±0·14
	0·5	0·5	0·61±0·17	0·67±0·19	0·39±0·11	0·50±0·13	0·39±0·11
Fixed date censoring = 20 %	0·3	0	0·33±0·14	0·33±0·16	0·20±0·10	0·14±0·15	0·26±0·12
	0·5	0	0·53±0·13	0·52±0·13	0·31±0·08	0·23±0·16	0·41±0·11
	0·3	0·3	0·33±0·15	0·31±0·14	0·19±0·09	0·11±0·18	0·23±0·14
	0·5	0·5	0·52±0·15	0·51±0·14	0·30±0·09	0·17±0·18	0·36±0·11
Fixed date censoring = 40 %	0·3	0	0·34±0·19	0·32±0·18	0·19±0·10	0·08±0·21	0·21±0·14
	0·5	0	0·53±0·16	0·52±0·15	0·31±0·09	0·12±0·20	0·34±0·11
	0·3	0·3	0·33±0·18	0·31±0·19	0·18±0·11	0·07±0·21	0·18±0·13
	0·5	0·5	0·54±0·17	0·53±0·17	0·32±0·11	0·09±0·21	0·28±0·12

add and *dom* denote the true values of simulated QTL effects. The means of the additive QTL effect estimates for C and W models (columns 1 and 2) can be compared with the true values *add*. However, the simulation process does not allow the direct comparison of the means of QTL effect estimates obtained with G or G′ methods directly with the true value *add*. But the means of standardized additive effect estimates ($-\hat{a}/\hat{\sigma}$) of G and G′ methods can be compared with the means of standardized additive effect estimates ($\hat{a}/\hat{\rho}$) of the W method. Thus columns 4 and 5 are comparable to column 3.

as though they were uncensored substantially improved the power and the estimates of location and QTL effects. However, this G′ approach is statistically incorrect and it is more affected by censoring at a fixed date than W and C. Recently, Diao *et al.* (2004) proposed a QTL mapping approach using a parametric Weibull model in QTL interval mapping methods (W_0), but they did not compare it with others by simulations. They used these methods to analyse experimental data (30 % of censored information at fixed data) previously analysed by Broman (2003) with a non-parametric method (NP) and a two-part method (2-part) and also a standard interval mapping method for only uncensored data (QT). QTL were found on chromosomes 1, 5, 13 and 15 at a significant level with at least one of these methods. W_0 found the most significant LRT for the QTL on chromosomes 5, 13 and 15 but the least LRT for QTL on chromosome 1. The result on chromosome 1 is surprising and it could be interesting to understand why this situation is so unfavourable to W_0.

(ii) *Including censoring in the likelihood for interval mapping using a Gaussian model*

In principle, it is possible to include censored records in the likelihood under the normal distribution, in a similar way to the model presented by Carriquiry *et al.* (1987). However, the likelihood expression becomes more complicated because it involves a cumulative distribution, which does not have a closed form, and computational time is increased. More importantly, the Gaussian model with censored data is very sensitive to small values of the trait (Cox & Oakes, 1984). Sorensen *et al.* (1998) proposed a Bayesian analysis of censored observations for a Gaussian mixed effects

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

model and Gibbs sampler, treating censored records as missing data. An adaptation of this approach to QTL detection could be interesting but certainly computationally demanding, particularly for the computation of the thresholds (1000 estimations under H0).

(iii) *Choice of the simulation method*

In order to compare G, G′, W and C, experimental data were used to generate simulations. Without censoring, the four methods led to similar results, showing that this simulation process did not favour any of them. However, this simulation process also had a drawback related to the difficulty of interpreting the QTL effects. To overcome this problem, data could be simulated assuming an exponential distribution. Then the estimated QTL effects and the simulated QTL effects could be compared directly with the three methods used. However, this simulation process tends to favour parametric methods (G and, above all, W which includes the exponential distribution as a specific case).

(iv) *Computational time and maximization method*

W and C have higher computational requirements than G or G′. The most time-consuming part was the likelihood maximization. With the Weibull and Cox models, the calculation of the first derivative instead of its finite difference approximation used here, might speed up the maximization process. To find the maximum likelihood estimator with the Weibull model, Diao *et al.* (2004) proposed applying the EM algorithm, which could be an interesting approach to decreasing computational requirements.

6. Conclusion

The QTL detection methods developed in this study consider Cox or Weibull survival models. When part of the data is censored at a fixed date, these methods substantially improve the power of QTL detection and the accuracy of QTL location and QTL effects, compared with classical approaches assuming a normal distribution of the uncensored data. An alternative is to use a Gaussian model, treating censored data as though they were uncensored. In this case, results were closer to, but not as good as, C and W.

Therefore, the use of QTL methods taking into account the characteristics of survival traits is attractive for the study of traits such as genetic resistance to a disease and longevity in animal populations. This approach can, for example, be applied to detect QTL related to scrapie incubation time in sheep, the length of productive life or time until occurrence of first mastitis in ruminants, or the length of competitive life of sport horses.

Appendix. Likelihood expressions for QTL mapping methods for a full-sib design

Assume now that the population consists of n sire families ($i=1, n$) with n_i mates for each sire i, ($j=1, n_i$) and n_{ij} progeny for each dam ij. Let $N=\sum_{i=1}^{n}\sum_{j=1}^{n_i} n_{ij}$ be the total number of individuals. All dams and sires are considered unrelated and a dam is assumed to be mated with only one sire. Let t_{ijk} be the failure time of ijk ($k=1, \ldots, n_{ij}$).

At a chromosomal location z, the general form of the likelihood used here was described before by Le Roy *et al.* (1998) in animal breeding designs:

$$\Delta^z = \prod_i \prod_j \left[\sum_{hd_{ij}} p(hd_{ij}|\hat{h}s_i, \mathbf{M}_i) \prod_{k\in\Omega} \times \left\{ \sum_g p(d_{ijk}^z = g|\hat{h}s_i, hd_{ij}, \mathbf{M}_i)\cdot l(ijk|g) \right\} \right] \tag{A1}$$

Where $\mathbf{M}_i$ is the vector of marker information on sire family i, and hs_i and hd_{ij} are the genotypes of markers of sire i and dam ij. $\hat{h}s_i$ corresponds to the most probable maximum hs_i conditional on $\mathbf{M}_i$. In fact, in animal breeding designs, the large number of progeny per sire allows the assumption that the sire genotype is correctly rebuilt from the marker information (Mangin *et al.*, 1999). $p(d_{ijk}^z = g)$ is the probability that individual i receives one of both grand parental segments (denoted 1 and 2) from its sire i and its dam j, where d_{ijk}^z is a random variable and $g=1, \ldots, 4$ correspond to (1,1), (1,2), (2,1), (2,2) progeny QTL genotypes, $l(ijk\in\Omega|g)$ is the contribution to the likelihood of t_{ijk} which depends on the distribution assumed. See Le Roy *et al.* (1998) for more details on the computation of the probability terms.

When using the Gaussian or Weibull parametric model (G and W), the contribution to the likelihood has an expression equivalent to the F2 design (A2 and A3). The contribution to the likelihood for the G model is:

$$l(ijk\in\Omega_{unc}|g) = \frac{1}{\sqrt{2\pi}\sigma_i} \times \exp\left[-\frac{1}{2}\left(\frac{y_{ijk} - \mu_{ij} - \mathbf{x}'_{ijk}\boldsymbol{\beta} - qtl_{ijg}}{\sigma_i} \right)^2 \right] \tag{A2}$$

where σ_i is the standard deviation of sire family i, $\mathbf{x}'_{ijk}$ is the ijkth row of the (N_{unc}, n_c) incidence matrix $\mathbf{X}$, and

$$\begin{aligned} qtl_{ijg} &= [as_i + ad_{ij}]/2 \text{ if } g=1,\\ &= [-as_i + ad_{ij}]/2 \text{ if } g=2,\\ &= [as_i - ad_{ij}]/2 \text{ if } g=3,\\ &= [-as_i - ad_{ij}]/2 \text{ if } g=4. \end{aligned}$$

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

with as_i/σ_i and ad_{ij}/σ_i being the sire and dam standardized QTL substitution effects. The other notations are the same as for the F2 design.

The contribution of individual *ijk* with genotype *g* to the likelihood for the W model is:

$$l(ijk \in \Omega | g) \propto \left[\rho \cdot y_{ijk}^{\rho-1}(\exp(\rho \log \lambda + \mathbf{x}'_{ijk}\boldsymbol{\beta} + qtl_{ijg}))\right]^{\delta_{ijk}} \times \exp\left[-y_{ijk}^{\rho}(\exp(\rho \log \lambda + \mathbf{x}'_{ijk}\boldsymbol{\beta} + qtl_{ijg}))\right] \qquad \text{(A3)}$$

where $\delta_{ijk} = 1$ if $ijk \in \Omega_{unc}$ and $\delta_{ijk} = 0$ if $ijk \in \Omega_{cens}$.

To develop an interval mapping method assuming the Cox model (C), the expression for the global likelihood must take into account both the familial structure and the order in which individuals die. The latter is necessary for constructing the set of animals at risk at any time (see expression 5). The general form of the likelihood (A1) is too complex to be used directly. Without any approximation, the complete likelihood in the interval mapping method must consider the Cox partial likelihood for each genotype combination of the global population (dam haplotypes and QTL genotype of offspring). Let *c* be one such combination ($c = 1, \ldots, N_c$). The number of combinations (N_c) is equal to the number of possible dam haplotype combinations in the population multiplied by the number of possible QTL genotype combinations of offspring conditional on the dam haplotypes of the combination *c*. In fact, the dam haplotype and the genotypes of her progeny must be compatible for each combination *c*. The probability of combination *c* can be written as:

$$p(c) = \prod_i \prod_j p(hd_{ijc} | \hat{h}s_i, \mathbf{M}_i) \times \prod_{k \in \Omega} p(d^z_{ijkc} = g | \hat{h}s_i, hd_{ijc}, \mathbf{M}_i) \qquad \text{(A4)}$$

where hd_{ijc} is the haplotype of the dam *ij* for combination *c* and d^z_{ijkc} is the QTL genotype of the individual *ijk* for combination *c*.

The global expression for the likelihood is then:

$$\Delta^z = \sum_c \left\{ [p(c)] \prod_i \prod_j \prod_{k \in \Omega unc} \times \frac{\exp(\mathbf{x}'_{ijk}\boldsymbol{\beta} + qtl_c)}{\sum_{i_d} \sum_{j_d} \sum_{k_d \in R(t_{ijk})} \exp(\mathbf{x}'_{i_d j_d k_d}\boldsymbol{\beta} + qtl_c)} \right\} \qquad \text{(A5)}$$

where $R(t_{ijk})$ is the set of censored or uncensored progeny k_d of sire i_d and dam j_d at risk at time t_{ijk}, and qtl_{ijc} is the QTL effect, corresponding to combination *c*.

The number of combinations (N_c) rapidly becomes extremely high even for a small population size. Thus the computation of (A5) is virtually impossible. If we consider:

$$e_{ijkc} = \exp(\mathbf{x}'_{ijk}\boldsymbol{\beta} + qtl_{ijc}) \qquad \text{(A6)}$$

$$E_{ijkc} = \sum_{i_d} \sum_{j_d} \sum_{k_d \in R(t_{ijk})} [\exp(\mathbf{x}'_{i_d j_d k_d}\boldsymbol{\beta} + qtl_{i_d j_d c})] \qquad \text{(A7)}$$

Then (A5) can be written as:

$$\Delta^z = \sum_c \left\{ [p(c)] \times \prod_i \prod_j \prod_{k \in \Omega unc} \frac{e_{ijkc}}{E_{ijkc}} \right\} \qquad \text{(A8)}$$

We cannot factorize this expression to obtain the general form of the likelihood (A1) because the denominator term, E_{ijkc}, depends on both the combination *c* and the familial structure. To allow such a factorization, we considered an approximation of (A8) where the denominator term (E_{ijkc}) is independent of the combination *c*. This denominator term is the weighted mean of the population at risk at time t_{ijk}.

$$\Delta^z \approx \sum_c \left\{ [p(c)] \times \prod_i \prod_j \prod_{k \in \Omega unc} \frac{e_{ijkc}}{\sum_{cd_{ijk}} [p(cd_{ijk}) E_{ijkcd_{ijk}}]} \right\} \qquad \text{(A9)}$$

where $p(cd_{ijk})$ is the contribution of each possible genotype combination (cd_{ijk}) of individuals ($i_d j_d k_d$) at risk at time t_{ijk} ($i_d j_d k_d \in R(t_{ijk})$). This probability $p(cd_{ijk})$ is:

$$p(cd_{ijk}) = \sum_{i_d} \sum_{j_d} p\left(hd_{i_d j_d cd_{ijk}} | \hat{h}s_{i_d}, \mathbf{M}_{i_d}\right) \times \sum_{k_d \in R(t_{ijk})} p\left(d^z_{i_d j_d k_d} = g | \hat{h}s_{i_d}, hd_{i_d j_d cd_{ijk}} \mathbf{M}_{i_d}\right) \qquad \text{(A10)}$$

where $hd_{i_d j_d cd_{ijk}}$ is the haplotype of dam $i_d j_d$ for the combination cd_{ijk} and $d^z_{i_d j_d k_d cd_{ijk}}$ is the QTL genotype of individual *ijk* for the combination cd_{ijk}.

We can consider genotypes within families to simplify the likelihood (A9). Thus using the general form of likelihood in equation (A4), the contribution of individual *ijk* with genotype *g* to the likelihood is as:

$$l(ijk \in \Omega_{inc} | g) \approx \frac{(\exp(\mathbf{x}'_{ijk}\boldsymbol{\beta} + qtl_{tjg}))}{\sum_{i_d} \sum_{j_d} \sum_{hd_{i_d j_d}} \left[p(hd_{i_d j_d} / \hat{h}s_{i_d}, \mathbf{M}_{i_d}) \times \left(\sum_{k_d \in R(t_{ijk})} \sum_{g_d} p(d^z_{i_d j_d k_d} = g_d | \hat{h}s_{i_d}, hd_{i_d j_d}, \mathbf{M}_{i_d}) \cdot \exp(\mathbf{x}'_{i_d j_d k_d}\boldsymbol{\beta} + qtl_{i_d j_d g_d}) \right) \right]} \qquad \text{(A11)}$$

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press

This work was partly supported by a European Grant (CT987017). The authors wish to thank D. Malo and G. Sebastiani (McGill University, Montreal) for providing data.

References

Baret, P. V., Knott, S. A. & Visscher, P. M. (1998). On the use of linear regression and maximum likelihood for QTL mapping in half-sib designs. *Genetical Research* **72**, 149–158.

Boichard, D., Grohs, C., Bourgeois, F., Cerqueira, F., Faugeras, R., Neau, A., Milan, D., Rupp, R., Amigues, Y., Boscher, M. Y. & Leveziel, H. (2000). La recherche de QTLs à l'aide de marqueurs: résultats chez les bovins laitiers. *INRA production Animale numéro hors série Génétique moléculaire: principes et applications aux populations animales*, 217–222.

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

Carriquiry, A. L., Gianola, D. & Fernando, R. (1987). Mixed model analysis of a censored normal distribution with reference to animal breeding. *Biometrics* **43**, 929–939.

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J.-J., Grisart, B., Mackinnon, M. & Georges, M. (1998). A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sibs pedigrees: application to milk production in a grand-daughter design. *Genetics* **149**, 1547–1555.

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

Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

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

Elsen, J. M., Mangin, B., Goffinet, B., Boichard, D. & Le Roy, P. (1999). Alternative models for QTL detection in livestock. 1. General introduction. *Genetics, Selection, Evolution* **31**, 213–224.

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

Kalbfleisch, J. D. & Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kaplan, E. & Meier, P. (1958). Nonparametric estimate from incomplete observations. *Journal of the American Statistical Association* **53**, 457–469.

Knott, S. A., Elsen, J. M. & Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**, 71–80.

Kruglyak, L. & Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

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

Law, A. M. & Kelton, W. D. (1982). *Simulation Modelling and Analysis*. New York: McGraw-Hill.

Le Roy, P., Elsen, J. M., Boichard, D., Mangin, B., Bidanel, J. P. & Goffinet, B. (1998). An algorithm for QTL detection in mixture full and half sib families. In *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, University of New England, Armidale, Australia, 11–16 January 1998, vol. 26, pp. 257–260.

Mangin, B., Goffinet, B., Le Roy, P., Boichard, D. & Elsen, J. M. (1999). Alternative models for QTL detection in livestock. II. Likelihood approximations and sire marker genotype estimations. *Genetics, Selection, Evolution* **31**, 225–237.

Sebastiani, G., Olien, L., Gauthier, S., Skamene, E., Morgan, K., Gros, P. & Malo, D. (1998). Mapping of genetic modulators of natural resistance to infection with *Salmonella typhimurium* in wild derived mice. *Genomics* **47**, 180–186.

Sorensen, D. A., Gianola, D. & Korsgaard, I. R. (1998). Bayesian mixed-effects model analysis of a censored normal distribution with animal breeding applications. *Acta Agriculturae Scandinavica* **48**, 222–229.

Walling, G. A., Halley, C. S., Perez-Enciso, M., Thompson, R. & Visscher, P. M. (2002). On the mapping of quantitative trait loci at marker and non marker locations. *Genetical Research* **79**, 97–106.

https://doi.org/10.1017/S0016672305007366 Published online by Cambridge University Press