Interval Mapping Methods to Detect QTL on Survival Data.

C. R. Moreno[a], J.M. Elsen[a], P. Le Roy[b], V. Ducrocq[b].
[a] INRA, Station d'Amélioration Génétique des Animaux, BP27, 31326 Castanet-Tolosan Cedex, France
[b] INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy-en-Josas, France

Summary
Quantitative Trait Loci (QTL) are usually looked for using classical interval mapping methods which assume that the trait follows a normal distribution. However, these methods cannot take into account the characteristics of most survival data such as non normal distribution and presence of censored data. In this paper, we propose two new QTL detection approaches which allow to consider censored data. One interval mapping method uses a Weibull model (W) which is popular to model survival trait and the other uses a Cox model (C) which avoids making any assumption on the trait distribution.
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' where censored data are analysed as uncensored data). An adequate mathematical transformation was used for parametric methods (G, G' and W).
When no data were censored, the three methods gave similar results. However, when some data were censored, G had a power of QTL detection but also accuracy of QTL location and of QTL effects, which decreased considerably with censoring, particularly when censoring is at a fixed date. Considering G', this decrease with censoring is also observed but it is low. Censoring had a negligible effect on results obtained with W and C methods.

1. INTRODUCTION

QTL (Quantitative Trait Loci) detection methods are used to look for chromosomal regions having an effect on production traits of interest. This type of analysis has two main aims. In selection programs, information about markers linked to a QTL can be considered (Boichard *et al*., 2000). From a more fundamental point of view, detected chromosomal regions can be used to look 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, the use of classical QTL detection methods induces a low power of detection and a bias in the estimate of effects and position of the QTL. Interval mapping methods have been proposed to analyse discrete traits (Kadarmideen *et al.*, 2000), but none applies to survival data.

Survival data are positive random variables (called failure time here after) describing in some sense the length of the interval between a point of origin and an end point. Survival analysis takes into account distribution forms (often far from the 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 (missing data) or censored data are incorrectly considered as uncensored. To estimate fixed effects, proportional hazard models are classically used in survival analysis. They can be parametric such as the Weibull regression model (Kalbfleisch and Prentice, 1980) or semi parametric such as the so-called Cox semi-parametric model (Cox, 1972). In the present paper, these two types of models were used to look for QTL based on 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 are variable.

2. MODEL DEFINITIONS

In this section of the paper, the QTL detection methods are first developed for inbred crosses. The methods are presented in two parts. First, the general form of the likelihood and the classical expression for the contribution of one observation to the likelihood are described for a classical interval mapping method using a Gaussian model. With this method, it must be underlined that only uncensored data can be legitimately included. Therefore, censored data were excluded from the analysis (G) or were illegitimately assumed 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, uncensored and censored data were included.

(a) General expression of the likelihood.
In an F2 population, considering that individuals are produced by heterozygote parents, each animal, k, has 4 possible QTL genotypes (1,1), (1,2), (2,1), (2,2), denoted as g=1,…, 4. As described by Lander & Botstein (1989), the general form of the likelihood at a chromosomal location x, can be written as :

$$\Delta^x = \prod_k \left\{ \sum_g p(d_k^x = g \mid M_k) \cdot l(k \mid g) \right\} \quad (1)$$

where $p(d^x_k=g|M_k)$ is the probability that animal k has genotype g conditional to its flanking marker information. The contribution to the likelihood $l(k|g)$ of the observation k depends on the assumed distribution of the trait (y_k). Let Ω (Ω=1, …,N) represent the list of uncensored (Ω_{unc}=1,…, N_{unc}) and censored observations k (Ω_{cens}=N_{unc}+1,…, N_{cens}): $\Omega=\Omega_{unc}+\Omega_{cens}$.

Thereafter, we considered three alternatives corresponding to the Gaussian, Weibull or Cox models.

(b) Interval mappping method using a Gaussian model: G and G'.
In (1), the contribution l(k|g) of animal 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: $k \in \Omega_{unc}$. So, the contribution to the likelihood is:

$$l(k \in \Omega_{unc} \mid g) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_k - \mu - x_k'\beta - qtl_g)}{\sigma} \right)^2 \right] \quad (2)$$

where y_k is the trait (failure time) of individual k, μ is the mean, σ is the standard deviation, β is the $(n_c \times 1)$ vector of covariate effects, n_c the number of levels of covariate effects, x_k' is the k^{th} row of the (N_{unc}, n_c) incidence matrix 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 dominant effects, respectively.
We distinguished two different Gaussian approaches: G considered alone uncensored information in the likelihood ($k \in \Omega_{unc}$) and G' considered all information ($k \in \Omega$, so the censored observations were assumed as uncensored observations).

(c) Interval mapping methods using Weibull and Cox survival models: W and C.
Survival analyses allow properly to consider censored observations. These analyses generally assume a random (i.e., non informative) censoring (Kalbfleisch & Prentice, 1980). Some useful definitions of functions are recalled here. If t represents the actual failure time, f(t) is the density function, S(t) is the survivor function and h(t) is the hazard function representing the rate at which failure occurs at time t (Kalbfleisch & Prentice, 1980).

In a parametric Weibull regression model, the hazard function is:
$h(t_k)=\lambda\rho(\lambda t_k)^{\rho-1}.\exp(x_k'\beta)$, 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 failure time which can be written as the product of the hazard function and the survival 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:

$$\Delta = \prod_k \left[h(t_k)\right]^{\delta_k} \times \left[S(t_k)\right] \quad (3)$$

where $\delta_k=1$ if $k \in \Omega_{unc}$ and $\delta_k=0$ if $k \in \Omega_{cens}$.
Therefore, the contribution of the animal 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:

$$l(k \in \Omega \mid g) \propto \left[h(y_k)\right]^{\delta_k} \times \left[S(y_k)\right] \propto \left[\rho.{y_k}^{\rho-1}\left(\exp\left(\rho \log\lambda + x_k'\beta + qtl_{kg}\right)\right)\right]^{\delta_{ijk}} \times \exp\left[-y_k^{\rho}\left(\exp\left(\rho\log\lambda + x_k'\beta + qtl_{kg}\right)\right)\right] \quad (4)$$

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

The Cox model allows the estimate of the regression coefficients in β making no assumption about the form of $h_0(t_{[k]})$. 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 $h_0(t_{[k]})$. In this expression, only uncensored observations have a non zero contribution but 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 of the exact likelihood function when the baseline hazard function is assumed to be piecewise constant. Peto's version of the Cox's partial likelihood is:

$$\Delta = \prod_{k\in\Omega_{unc}}\left[\frac{\exp(x_k'\beta)}{\sum_{k_d\in R(t_k)}\exp(x_{k_d}'\beta)}\right] \quad (5)$$

where $R(t_k)$ is the list 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 and no longer over all distinct failure times in Cox (1972).
In the interval mapping model, there are four terms corresponding to the different genotypes of each individual k_d. Then to obtain an expression equivalent to (5), the terms in the denominator must be weighted by the probability that animal k has genotype g conditional to its marker information ($p(d^x_{kd}=g_d \mid M_{kd})$. By analogy with (5), the contribution of animal k to the likelihood in interval mapping using a Cox model (C) can be written as:

$$l(k \in \Omega_{unc} \mid g) \approx \frac{\left(\exp\left(x_k'\beta + qtl_g\right)\right)}{\left[\left(\sum_{k_d\in R(t_k)}\sum_{g_d} p(d^x_{k_d}=g_d \mid M_{k_d})\cdot\exp\left(x_{k_d}'\beta + qtl_{g_d}\right)\right)\right]} \quad (6)$$

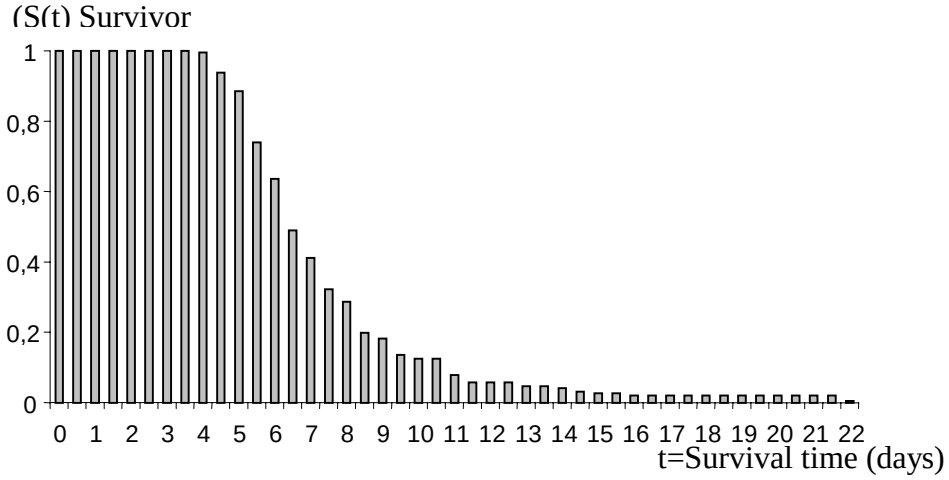
where $R(t_k)$ is the list of individuals at risk at time $t_k=y_k$.

3. DATA AND SIMULATIONS

(a) Experimental Design.

Data were simulated following the structure of a published experimental design (Sebastiani *et al.*, 1998). One hundred ninety one F2 animals were produced using two inbred mouse lines. Their survival times were measured after inoculation with a pathogen bacteria, *Salmonella thiphimurium*. All animals died at the end of the experiment, so no data were censored. Sebastiani *et al.* (1998) used two approaches to look 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 QTL located in similar regions and having similar effects.
Here, data were simulated based on the failure time 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.

Figure 1: Survivor distribution of experimental data used to produce simulations.



The failure time distribution of this design (figure 1) was used as the basal survival data (the simulation process is explained in the next section). A QTL effect and a censoring process were added to this basal distribution as described in the following section. Marker genotypes of chromosome 1 were used. This chromosome had the longest typed region.
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 corresponding to the end of the experiment. The censoring at random dates allows to consider 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 inoculation dates). Five scenarios of censoring were considered: uncensored data, 20% and 40% of censored records at random dates, 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, 0.3 or 0.5 and either no dominant effect (dom=0) or a fully dominant effect (dom=add). Then for each scenario of censoring, six situations with different values of dominant and additive effects were considered (500 simulations each time). Therefore under H1, a total number of 20 situations were studied.

(b) 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 $y_k=t_k$ (k=1,..,n) be the *observed failure times* and $T_{[1]}<T_{[2]}<\ldots<T_{[i]}<\ldots<T_{[m]}$ the *ordered distinct failure times* ($m\leq n$).

First, the survival function S(t), which is the probability to be alive at time t, was estimated using the Kaplan-Meier estimator (Kaplan and Meier, 1958):

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

where $n_{[i]}$ is the number of animals known to be alive just prior to time $T_{[i]}$, $d_{[i]}$ 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: $S_0(t)=S_{KM}(t)$.

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

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

The generation of simulated records was realised in two steps: firstly the choice of a QTL genotype and secondly, the choice of a survival time value. For each animal, the probability of the four QTL genotypes was calculated conditional to flanking marker information: $p(d^x_k=g|M_k)$ (see (1)). Upon using the three probabilities $p(d^x_k=1|M_k)$, $p(d^x_k=2 \text{ or } 3|M_k)$ and $p(d^x_k=4|M_k)$, a QTL genotype was drawn from a trinomial distribution. Knowing the genotype g, the survival function value was drawn from a [0,1] uniform distribution. The value of the observed survival time (t_k) was obtained by inversing the survival function conditional to genotype g, $t_k= S^{-1}(t_k|g)$.

In almost all cases, the simulated t_k values did not correspond exactly to an originally observed value. Therefore, to get realistic t_k values, a linear interpolation between the original observed t_k values or 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 a true parameter distribution.

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

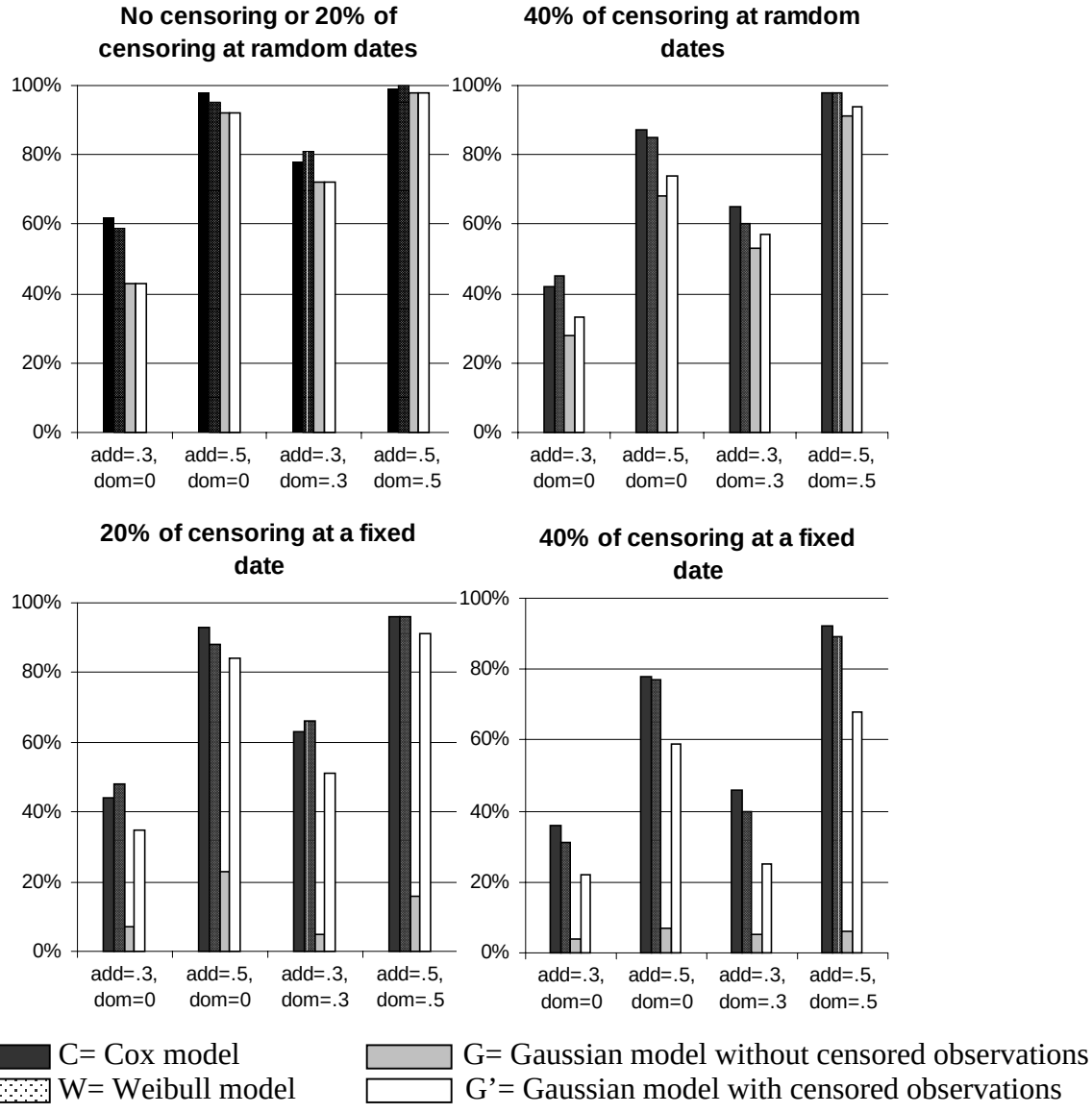
(d) Computational techniques.
Simulated data were transformed to perform the analysis with G, G' and W. With G and G', a logarithmic transformation was used to partly normalise the data. With W, a translation of the data was necessary because there is no failure observations between days 0 and 4 (Cox and Oakes, 1984). The best transformation was found to be t*=t-3.9 for failure time.
The likelihood function was maximised using a quasi-Newton algorithm implemented as a NAG subroutine (E04JYF) for the three methods.

4. RESULTS

The G, G', W and C were computed to estimate maximum likelihood ratio tests (LRT) for 4 situations under H0 and 20 situations under H1. The power of these methods and the estimates of the additive effect, dominant effect and location of QTL were compared.

(a) Comparison of QTL detection power of G, G', W and C.
In figure 2, the evolution of the power of the 4 methods is presented as a function of the QTL effects for the five situations of censoring. Without censoring and with 20% of censoring at random dates, the three methods have a similar power, whatever the simulated QTL effects. In these situations, the power of all methods becomes close to 100%, even though the power of G and G' decreases slightly for small values of QTL effects.
With 40% of censoring at random dates, C and W have similar power and are more powerful than G and G'. Under a Gaussian model, taking into account censored data is slightly more powerful than missing them.
With censoring at a fixed date, C and W appear equivalent. Otherwise, the power of G becomes very low and the larger the dominant effect is, the lower the power is. This trend increases when censoring rate increases, that is when more censored data are excluded from the G analysis. In the extreme case of 40% of censoring at a fixed date and additive and dominant effect equal to .5, the power of G is less than 10%, when the power of C and W is more than 90%. G' has a power 10 to 20% lower than C and W when there are 40% of censoring at fixed date or 20% of censoring at fixed date and an additive effect equal to .3.

Figure 2: QTL detection power with G, G', W and C methods, as a function of simulated QTL effects for the five situations of censoring (QTL have an additive effect a and a dominant effect d).



(b) Comparison of QTL location estimated with G, G', W and C.

Most estimated locations tend 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

increased. However, the bias decreased and the accuracy increased when QTL effects increased. The trend is slightly higher than previously for G'. Equivalent results were obtained with G when no censoring or censoring at random dates was applied. However, the latter method had an accuracy that decreased and a bias that increased with the proportion of censoring at a fixed date, particularly for situations where a dominant QTL effect was simulated.

(c) Comparisons of additive and dominant QTL effects estimated with G, G', W and C.
With W and C, estimates of QTL effects were similar between the different situations of censoring. The estimates were only slightly biased or were unbiased.
Comparing G (or G') and W standardised estimates, the values were slightly under-estimated for G and G' when censoring was not applied or was at random dates. A different situation was found when censoring was at a fixed date: the G estimates of the effects were two or three times smaller than the W estimates. The accuracy was also considerably affected. G' show more steady results.

6. DISCUSSION CONCLUSION

Considering that an adequate transformation is used in the parametric models (logarithmic transformation in G or G' and translation transformation in W), the selection of the model did not appear critical when censoring was not applied or was at random at least in the example considered here. In this situation, the G, G', W and C approaches obtained similar results: detection power, accuracy and bias of the estimates. When censoring at a fixed date was applied on the data, the situation changed. The G approach was strongly affected by censoring at a fixed date. This tendency increased when there was a dominant QTL effect. In the latter case, extreme data, which were censored, were the most informative ones for estimating QTL effects.
In the analysed situations where censoring at a fixed date was applied (for example, due to the end of the study) a classical method, such as G, was not at all adequate. An alternative is to use a Gaussian model, assuming censored data as uncensored. In this case, results were close to but lower than C and W.
Therefore, the use of QTL methods taking into account the characteristics of survival traits is very attractive for the study of such traits as genetic resistance to a disease and longevity in animal populations. This approach can be applied for example to detect QTL related with scrapie incubation time in sheep, the length of production life or time until mastitis occurrence in ruminants or the length of competitive life of sport horses.

References

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.

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. & Oakes, D. (1984). Analysis of survival data. Chapman and Hall, London, 201p.

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

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 quantiatitve trait locus mapping for polygenic binary traits using generalized and regression interval mappping 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: John Wiley & Sons.

Kaplan, E. & Meier, P. (1958). Non parametric 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.

Leroy, 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. Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, University of New England, Armidale, 11-16 January 1998, vol. 26, Australia, 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.