DOI: 10.1534/genetics.106.055921

# A Simple Linear Regression Method for Quantitative Trait Loci Linkage Analysis With Censored Observations

**Carl A. Anderson,*,†,[1] Allan F. McRae† and Peter M. Visscher*,†**

**Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, Scotland EH9 3JT and* †*Genetic Epidemiology Group, Queensland Institute of Medical Research, Brisbane, Australia 4029*

Manuscript received January 17, 2006
Accepted for publication April 14, 2006

ABSTRACT

Standard quantitative trait loci (QTL) mapping techniques commonly assume that the trait is both fully observed and normally distributed. When considering survival or age-at-onset traits these assumptions are often incorrect. Methods have been developed to map QTL for survival traits; however, they are both computationally intensive and not available in standard genome analysis software packages. We propose a grouped linear regression method for the analysis of continuous survival data. Using simulation we compare this method to both the Cox and Weibull proportional hazards models and a standard linear regression method that ignores censoring. The grouped linear regression method is of equivalent power to both the Cox and Weibull proportional hazards methods and is significantly better than the standard linear regression method when censored observations are present. The method is also robust to the proportion of censored individuals and the underlying distribution of the trait. On the basis of linear regression methodology, the grouped linear regression model is computationally simple and fast and can be implemented readily in freely available statistical software.

DOMESTIC animals and experimental species provide a unique resource for the understanding of quantitative genetic variation. Quantitative trait analysis of experimental crosses has provided many important insights into the genetics of complex traits (Morgante and Salamini 2003; reviewed in Andersson and Georges 2004). Several genes underlying quantitative genetic variation have been identified in the fields of animal and crop science, many of which have significant commercial potential (*e.g.*, Jeon *et al.* 1999; Nezer *et al.* 1999; Frary *et al.* 2000; Fridman *et al.* 2000; Grisart *et al.* 2002).

Most current quantitative trait loci (QTL) mapping techniques utilize an interval-mapping approach first put forward by Lander and Botstein (1989). The approach places a hypothetical trait locus at fixed incremental positions (for example, every 1–2 cM) along a map of known marker positions and tests for its effect on the trait using information from flanking markers. For a given location the basic linear model is

$$y_{ij} = m_j + e_{ij},$$

where $y_{ij}$ is the trait value for individual $i$ with genotype $j$, $m_j$ is the mean effect of genotype $j$, and $e_{ij}$ is random error ($e_{ij} \sim N(0, \sigma_e^2)$). The genotype of an individual at the position being tested is rarely known so the probability of an individual being each of the possible genotypes is calculated from the available marker information. Lander and Botstein (1989) implement their method using a maximum-likelihood approach. The maximum-likelihood method takes into account heterogeneous variances within marker classes to estimate genotype probabilities. The model parameters are estimated under both the null (no QTL) and alternative (with QTL) hypotheses. An advantage of maximum likelihood is that it uses all of the available observations on marker genotypes and trait values. The disadvantage of maximum likelihood is that it is computationally intensive and usually requires specialized software.

An alternative method, least-squares regression, uses expected genotype probabilities calculated from flanking markers rather than the more complex approximation via maximum likelihood (Haley and Knott 1992). For this approach least-squares linear regression is used to estimate the effect of genotype on the trait of interest at each test position along the genome. The asymptotic equivalence of least-squares regression with maximum-likelihood interval mapping has been shown through simulation (Haley and Knott 1992) and by theoretical calculations of power (Rebai *et al.* 1995). The least-squares approach has been shown to be robust to deviations from normality in all but the most extreme situations (Visscher *et al.* 1996; Rebai 1997). Kao (2000) and Knott (2005) review the differences between maximum-likelihood and regression QTL-mapping methods.

[1]*Corresponding author:* Genetic Epidemiology Group, Queensland Institute of Medical Research, 300 Herston Rd., Brisbane, Australia 4006. E-mail: carl.anderson@qimr.edu.au

Time-to-event traits are often nonnormally distributed and show a right-skewed distribution of trait values across all individuals. Additionally, time-dependent traits often include censored observations, which occur when the true time of the event is unknown. End-of-study censoring arises when the event of interest has not occurred by the end of the study period. Within-study censoring arises if an individual is lost to follow-up during the course of the study. The loss of information due to censoring results in lower statistical power, where the greater the proportion of censoring the lower the statistical power. Some of this power can be recovered by modeling censored individuals in the statistical analysis; however, standard QTL-mapping techniques typically do not account for this.

The field of survival analysis utilizes special methods to make better use of the information provided by censored observations and to better account for the nonnormal distribution of the trait values. Traditionally, proportional hazard regression models are used to model survival traits. These methods assume that if there are two individuals, $a$ and $b$, with $p$ time-independent covariate values in vectors $\mathbf{Z_a}$ and $\mathbf{Z_b}$, respectively, the ratio of their hazards is given by

$$\frac{h(t\,|\,\mathbf{Z_a})}{h(t\,|\,\mathbf{Z_b})} = \frac{h_0(t)\exp[\sum_{k=1}^{p}\beta_k Z_{ak}]}{h_0(t)\exp[\sum_{k=1}^{p}\beta_k Z_{bk}]} = \exp\left[\sum_{k=1}^{p}\beta_k(Z_{ak}-Z_{bk})\right],$$

where $h(t\,|\,\mathbf{Z}_i)$ is the hazard for individual $i$ at time point $t$, $h_0(t)$ is the baseline hazard function, and $\beta_i$ is the coefficient for the effect of the $i$th covariate. As time dependence is included only in the baseline hazard, the ratio of the hazards of two individuals at any time point is a constant and therefore the hazards are proportional (Klein and Moeschberger 1997). Cox (1972) proposed a semiparametric proportional hazards model that can be used to model survival data without prespecifying the distribution of the baseline hazard. This method is widely used and has been shown to be both robust and powerful. Parametric proportional hazard models also exist, which assume that the survival times follow a given distribution (for example, Weibull). Under the correct baseline hazard distribution, parametric models are more powerful than the equivalent nonparametric or semiparametric methods. However, when using real data, the true underlying distribution of the baseline hazard is unknown. For this reason, Cox proportional hazards regression remains the method of choice for most simple survival analyses. Moreno *et al.* (2005) compared the Weibull and Cox proportional hazards models to a more conventional QTL-mapping method that ignored the nature of the survival data and found that when analyzing survival trait data the proportional hazards models have greater power.

A drawback of both proportional hazards methods is that they are computationally intensive for complex models. Models with many covariates, some of which may be time dependent, can take extensive periods of time to analyze. Several computationally intensive approaches have been proposed (*e.g.*, Symons *et al.* 2002; Epstein *et al.* 2003; Diao *et al.* 2004; Diao and Lin 2005; Pankratz *et al.* 2005). In the presence of censored observations, the mapping of QTL for survival traits in line crosses can be carried out using the methods of Symons *et al.* (2002), Diao *et al.* (2004), or Diao and Lin (2005). When considering QTL mapping for survival traits in outbred populations the variance component-based methods of Epstein *et al.* (2003) or Pankratz *et al.* (2005) are appropriate. All of these methods are yet to be incorporated into general, widely used genome-analysis packages.

Here we demonstrate a novel grouped linear regression method for the analysis of survival data that is computationally simple and robust and can be implemented in standard statistical packages. The method is compared to the classical Cox and Weibull proportional hazards approaches and to a standard linear regression method that ignores censoring status. We demonstrate its relative power and robustness via simulation and discuss the advantages of this simplified method compared to those currently available.

## METHODS

The Cox proportional hazards model has been widely adopted as the method of choice for survival analyses. However, when analyzing survival data with many tied or grouped observations, or when analyzing a large data set, the Cox model becomes computationally intensive. Grouped survival information is defined as noncontinuous survival time data. Prentice and Gloeckler (1978) extended the popular Cox model for the analysis of grouped survival data. The method correctly models grouped survival data but still remains computationally intensive for large data sets. We propose a grouped approximation for continuous survival data where failure times are partitioned into a number of time periods and suggest a linear regression model for the analysis of the grouped data. The aim of this method is to simplify the analysis of continuous survival data leading to reduced computation time and an increased ability to analyze models with greater complexity. The survival of each individual through these arbitrary time periods is coded using a series of conditional survival indicator variables, similar to that used by Madgwick and Goddard (1989) to predict breeding values in dairy cattle using lactation period survival data. Rather than adopt the maximum-likelihood approach of Prentice and Gloeckler (1978) for parameter coefficient estimation we suggest a computationally efficient and robust linear regression method. The simplicity and

**TABLE 1**

**Example of the grouped linear regression group coding algorithm for two time periods**

| Individual | Survival | Survival record(s) |
|---|---|---|
| 1 | Censored during the first time period | NA |
| 2 | Event occurs during the first time period | 1 |
| 3 | Survives the first time period and the event occurs during the second | 0 1 |
| 4 | Survives the first time period and then is censored during the second | 0 |

Individual $i$ survives interval $j$, survival record is $x_{ij} = 0$. Individual $i$ experiences the event during interval $j$, survival record is $x_{ij} = 1$ and there are no further survival records for the remaining intervals. An individual is censored during a particular interval and then that individual has no survival record for the current or subsequent intervals.

efficiency of the model should allow the analysis to be carried out quickly on large data sets using standard statistical packages.

**Grouped linear regression method:** Survival times are sorted in chronological order, regardless of censoring status or genotype, and separated into a predefined number of groups or time periods. The survival record for individual $i$ during time period $j$ is given by $x_{ij}$. If individual $i$ survives interval $j$, then the corresponding survival record is $x_{ij} = 0$. If individual $i$ experiences the event during interval $j$, then the survival record is $x_{ij} = 1$ and there are no further survival records for the remaining intervals. If an individual is censored during a particular interval then that individual has no survival record for the current or subsequent intervals (Table 1). For any given time period, the survival record represents the conditional probability that an individual survives the current time period given that the individual survived to the start of that time period. For any individual the survival records for each group are therefore independent observations. The linear model used in the regression analysis is given by

$$x_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_k t_{ik} + \beta_g g_i + \varepsilon_{ij},$$

where $x_{ij}$ is the survival record for individual $i$ during time period $j$, the $\beta$ terms are the estimated regression coefficients, $t_{ik}$ is an indicator variable for time period $k$ that takes a value of 1 in time period $j$ and 0 otherwise, $g_i$ is the genotype for individual $i$, and $\varepsilon_{ij}$ is the random error for individual $i$ during time period $j$. The terms $\beta_0 + \sum \beta_k t_{ik}$ model the baseline hazard and $\beta_g g_i$ estimates the genotypic effect on the hazard. Only $p - 1$ coefficients of $t_i$ can be estimated, as there are no noncensored individuals surviving the last time period. The significance of the effect of genotype is estimated using standard regression methodology. The resulting $F$-statistic (with 1 and $n - p - 1$ d.f., where $n$ is the number of individuals) is transformed to an approximate likelihood-ratio test statistic (LRT) using the formula provided by Baret *et al.* (1998),

$$\mathrm{LRT} = n \times \log_e\left(1 + \left(\frac{1}{n - p - 1}\right)F\right),$$

to allow comparisons with the maximum-likelihood test statistic of the proportional hazards methods.

**Simulation of data:** Extensive simulations were carried out using the statistical software package R (R Development Core Team 2005). We simulated genotypic data at a single locus for individuals from a backcross between two fully inbred lines. Marker data were generated at a single QTL locus with possible alleles $q$ and $Q$, thus assigning an individual the genotypes $qq$ or $Qq$ with equal probability. Phenotypic data were simulated assuming a fully penetrant QTL at the marker locus. Phenotypic data were drawn from a number of distributions (Table 2).

The baseline hazard function for data drawn from a Weibull distribution is given by

**TABLE 2**

**Distributions used to simulate data**

| Model | Distribution | Genotype $qq$ | | Genotype $Qq$ | | Probability density function |
|---|---|---|---|---|---|---|
| | | Shape ($\rho$) | Scale ($\lambda_{qq}$) | Shape ($\rho$) | Scale ($\lambda_{Qq}$) | |
| 1 | Weibull | 2.00 | 10.00 | 2.00 | 9.05 | $f(t) = \frac{\rho}{\lambda}\left(\frac{t}{\lambda}\right)^{\rho-1} e^{-(t/\lambda)^{\rho}}$ |
| 2 | Exponential | — | 10.00 | — | 9.16 | $f(t) = \lambda e^{-(\lambda t)}$ |
| 3 | Gamma | 3.65 | 2.19 | 3.65 | 2.43 | $f(t) = \frac{1}{\lambda^{\rho}\Gamma(\rho)} \times t^{\rho-1}\exp^{-(t/\lambda)}$ |
| 4 | Gamma | 0.50 | 10.00 | 0.50 | 7.50 | |

$$h_0(t) = \frac{\rho}{\lambda}\left(\frac{t}{\lambda}\right)^{\rho-1}.$$

Thus, the ratio of the hazards for genotypes *qq* and *Qq* is

$$\frac{h(t \mid Qq)}{h(t \mid qq)} = \left(\frac{\lambda_{qq}}{\lambda_{Qq}}\right)^{\rho},$$

which is independent of time ($t$) and thus satisfies the assumptions of a proportional hazards model. When considering model 1 (Weibull), individuals with genotype *Qq* are at an increased risk of ~22% when compared to the risk for genotype *qq* (for $\rho = 2$ and $\lambda_{qq} = 10$, $\lambda_{Qq} = 9.05$). The mean of a Weibull distribution is given by $\lambda \times \Gamma(1 + \rho^{-1})$, where $\Gamma()$ is the gamma function. Thus, the simulated effect of the genotype on the hazard is equivalent to a mean difference in survival time of $(10 - 9.05)\Gamma(1.5) = 0.84$ (0.18 standard deviation units). The shape and scale parameters of model 2 (exponential) were calculated to give the same ratio of hazards as that of the Weibull parameters used in model 1 but from a more highly skewed distribution. The parameters of model 3 (gamma) were chosen to approximate the means and variances of the Weibull distributions used in model 1. To test the relative robustness of the methods when phenotypic data are drawn from highly skewed distributions we simulated data from two gamma distributions where the gamma shape parameter was 0.5. Shape parameters were chosen to approximate the hazard ratio used in model 1. The proportion of censored individuals, $P_c$, was varied across simulations. For the generation of censored observations the method of Diao *et al.* (2004) and Diao and Lin (2005) was implemented. Letting $T_i$ be the survival time of the $i$th individual, $C_i$ is the censoring time and $I(C_i)$ is an indicator variable giving the censoring status (0 = censored, 1 = uncensored) for individual $i$. Censoring times ($C_i$) were drawn from a uniform distribution between $0 < x \leq 1$ and multiplied by a constant $\tau$. Diao *et al.* (2004) and Diao and Lin (2005) use a trial-and-error approach to obtain $\tau$. We calculated the value of $\tau$ via numerical integration to provide a given proportion of censored observations (see appendix). If $C_i$ was less than $T_i$ then the individual was classified as censored ($I(C_i) = 0$) and the censoring time was entered into subsequent analyses. Individuals with greater survival times are more likely to be censored. The censoring method creates both "within-study" and "end-of-study" censoring.

**Analysis of simulated data:** Two methods were used to group the observations (both censored and uncensored) into time periods to investigate the robustness of the grouped linear regression method to the grouping mechanism. The first grouping method (A) groups the individuals into $k$ groups such that an equal number of observations (either censored or uncensored) occur in each time period. The standard error of the group means is approximately equal when using this method. The second grouping method (B) groups the individuals into $k$ groups such that an equal proportion of individuals, denoted by $s$, survive each time period. Within-group variances are approximately equal when using this method. For both methods the last time period contained all remaining individuals not previously allocated a time period. For example, with 1000 individuals and five time periods, grouping mechanism A creates five groups of 200 individuals, while grouping method B with $s = 0.5$ creates groups of 500, 250, 125, 63, and 62 individuals. The change in mean test statistic in relation to the number of time periods and the proportion of individuals surviving each time period was investigated for each grouping mechanism.

**Power comparisons:** We compared the power of the grouped linear regression method to that of the Cox and Weibull proportional hazards models and the standard linear regression least-squares approach (ignoring censoring status). The inclusion of the standard linear regression method allows some approximation of the power to be gained by including the censored observations correctly in the regression model. Grouping method B was used to separate the continuous survival times into groups for the grouped linear regression method. Values were chosen for the grouping parameters that approximately maximized the power of the grouped linear regression method. The methods were contrasted by comparing the mean test statistic produced for each mode of analysis given phenotypic data drawn from the same underlying distribution. To ensure that this was an unbiased comparison of methods, using each mode of analysis we carried out 1000 simulations of 1000 individuals under the null hypothesis and checked for deviations from a chi-square distribution with 1 d.f. This was tested using a one-tailed Kolomonov–Smirnov test. Where a significant deviation from a chi-square 1-d.f. distribution was seen for a given phenotypic distribution, we compared the analysis methods via empirical $P$-values. To calculate the empirical thresholds 10,000 replicates of 1000 individuals were simulated under the null hypothesis. The proc.time( ) function in R was used to obtain run times for the model-fitting step of each mode of analysis.

## RESULTS

**Grouping method:** Tables 3–5 show the effect of altering the group survival proportion ($s$) and the number of groups ($k$) on the grouped linear regression mean test statistic. The shape and scale parameters of model 1 (Weibull) were used in the simulations. Some combinations of $s$ and $k$ are impossible with a sample size of 1000 individuals as the number of individuals in a group falls rapidly at low values of $s$, thus limiting the possible number of groups ($k$). For the range of groups we simulated, the mean test statistic approximately increased with the number of time periods into which

**TABLE 3**

**Grouping method A: grouped linear regression mean test statistic**

| Censoring proportion ($P_c$) | No. of groups ($k$) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0 | 1.94 | 7.70 | 8.24 | 8.76 | 9.42 | 9.28 | 9.55 | 9.32 | *9.87* |
| 0.1 | 1.91 | 6.72 | 7.14 | 8.26 | 8.44 | 8.48 | 8.91 | *8.98* | 8.80 |
| 0.5 | 3.13 | 5.22 | 5.04 | 5.67 | 5.68 | 5.49 | 5.83 | 5.80 | *5.84* |

Phenotypes are drawn from model 1 (Weibull) distributions ($\rho = 2$, $\lambda_{qq} = 10$, $\lambda_{Qq} = 9.05$). Numbers in italics depict the maximum mean test statistic for the given censoring proportion.

the data were grouped. This result was seen for both grouping methods. An increase in mean test statistic was seen when comparing grouping method B to grouping method A. This relationship held only if the optimal group survival proportion was determined correctly for grouping method B. However, this may not always be possible, in which case adopting a grouping method in which each group contains an equal number of individuals (A) leads only to a small reduction in mean test statistic. When the same proportion of individuals survives each group (grouping method B) the lowest possible group survival rate, given the number of time periods and sample size, gave a reasonable approximation to maximize the mean test statistic. For the power comparisons grouping method B was adopted, where an equal proportion of individuals ($s = 0.6$) survived each of the groups ($k = 10$). These parameters will not maximize the mean test statistic in all situations, but provide a good approximation to the optimal parameters in the situations tested and good practical guidelines for other studies of similar size.

**Power comparisons:** The distribution of test statistics for all four modes of analysis was shown to be chi square under the null hypothesis for model 1 (Figure 1). This supports the use of the mean test statistic as an unbiased parameter for the comparison of methods. Figure 2 shows the mean test statistic for all four modes of analysis with varying percentages of censored survival times. In the absence of censoring the four analysis methods have approximately equal power. The standard linear regression model has a slightly reduced mean test statistic when compared to the survival analysis methods. This difference is likely due to the nonnormal distribution of the survival times. As the percentage of censored observations is increased, the standard linear regression mean test statistic falls rapidly. In comparison, that of the grouped linear regression and Cox and Weibull proportional hazard models decreases at a slower and comparable rate. The relative and actual run times from the model-fitting step of each analysis method are shown in Table 6. The procedure time for fitting 100 models showed the grouped linear regression method to be almost 5 times quicker than the Cox proportional hazards method and >12 times quicker than the Weibull proportional hazards method.

As expected, when simulating data under the null hypothesis from model 2 (exponential) no significant deviations were detected from a chi-square distribution with 1 d.f. When comparing the relative mean test statistics from model 2 (exponential), similar relationships to those under model 1 (Weibull) were observed (Figure 3). Given that an exponential distribution is a

**TABLE 4**

**Grouping method B: grouped linear regression mean test statistic with no censoring**

| Group survival proportion ($s$) | No. of groups ($k$) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0.1 | 5.38 | *7.09* | — | — | — | — | — | — | — |
| 0.2 | 4.91 | *8.50* | — | — | — | — | — | — | — |
| 0.3 | 3.85 | 9.02 | *9.39* | — | — | — | — | — | — |
| 0.4 | 2.85 | 8.85 | 9.72 | 10.18 | *10.24* | — | — | — | — |
| 0.5 | 1.94 | 8.13 | 9.54 | 10.15 | 10.66 | *10.78* | — | — | — |
| 0.6 | 1.21 | 7.17 | 8.68 | 9.62 | 9.97 | 10.25 | 10.69 | *10.76* | 10.74 |
| 0.7 | 0.65 | 6.18 | 7.41 | 8.56 | 9.03 | 9.90 | 9.99 | 10.36 | *10.43* |
| 0.8 | 0.35 | 4.52 | 5.71 | 6.94 | 7.72 | 8.05 | 8.77 | *9.48* | 9.45 |
| 0.9 | 0.10 | 2.89 | 3.56 | 4.55 | 5.23 | 5.76 | 6.18 | 6.83 | *6.93* |

Phenotypes are drawn from model 1 (Weibull) distributions ($\rho = 2$, $\lambda_{qq} = 10$, $\lambda_{Qq} = 9.05$). Numbers in italics depict the maximum mean test statistic for the given group survival proportion ($s$). Some combinations of $s$ and $k$ are impossible with a sample size of 1000 (—).

**TABLE 5**

**Grouping method B: grouped linear regression mean test statistic with 50% censoring**

| Group survival proportion ($s$) | No. of groups ($k$) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0.1 | 3.46 | *3.99* | — | — | — | — | — | — | — |
| 0.2 | 3.23 | *5.04* | — | — | — | — | — | — | — |
| 0.3 | 2.66 | 5.06 | *5.29* | — | — | — | — | — | — |
| 0.4 | 2.16 | 4.81 | 5.42 | 5.34 | *5.62* | — | — | — | — |
| 0.5 | 1.78 | 4.72 | 5.04 | 5.45 | 5.69 | *5.97* | — | — | — |
| 0.6 | 1.61 | 3.97 | 4.65 | 5.23 | 5.23 | 5.69 | 5.62 | 5.63 | *5.81* |
| 0.7 | 1.29 | 3.26 | 3.93 | 4.37 | 5.13 | 5.32 | 5.48 | *5.83* | 5.76 |
| 0.8 | 1.10 | 2.34 | 3.23 | 3.68 | 3.91 | 4.32 | 4.65 | 5.13 | *5.28* |
| 0.9 | 1.07 | 1.51 | 1.93 | 2.38 | 2.61 | 2.79 | 3.09 | 3.39 | *3.80* |

Phenotypes are drawn from model 1 (Weibull) distributions ($\rho = 2$, $\lambda_{qq} = 10$, $\lambda_{Qq} = 9.05$). Numbers in italics depict the maximum mean test statistic for the given group survival proportion ($s$). Some combinations of $s$ and $k$ are impossible with a sample size of 1000 (—).

special case of a Weibull distribution, when $\rho = 1$, it is not surprising to find that the parametric Weibull proportional hazards method is of equal power when compared to the Cox proportional hazards model.

When phenotypic data were simulated under the null hypothesis using model 3 (gamma), a significant deviation from a chi-square distribution ($P = 0.012$) was seen when using the Weibull proportional hazards model (Figure 4). This result was not unexpected as the Weibull proportional hazards model fits a Weibull distribution to the distribution of survival times, which in this case follow a gamma distribution. To make unbiased comparisons between the four analysis methods we calculated empirical $P$-values for each method of analysis (Figure 5). When the phenotypes contained no censored observations the best-performing model was standard linear regression. This is not surprising as a gamma distribution with a shape ($\rho$) of 3.65 is not highly skewed. The worst-performing model under no censoring was the Weibull proportional hazards model. Again,

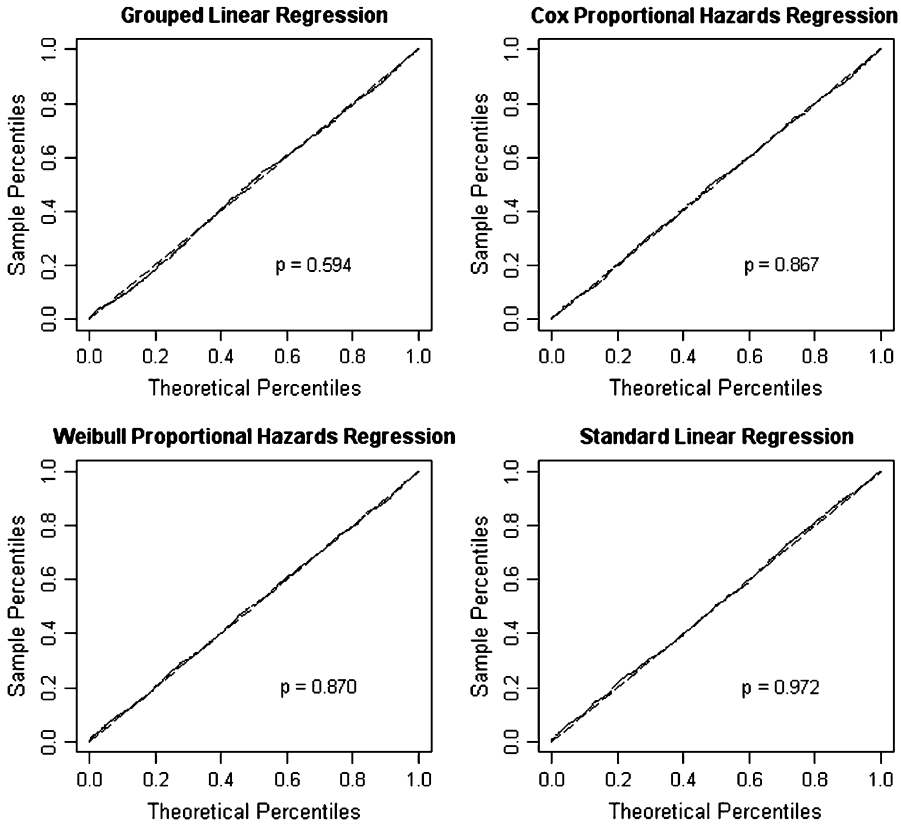


FIGURE 1.—Quantile–quantile plots for phenotypes simulated from model 1 (Weibull) under the null hypothesis. Theoretical percentiles were calculated from a chi-square distribution with 1 d.f. Sample percentiles were calculated empirically by simulating 1000 replicates of 1000 individuals under the null hypothesis. The dashed line denotes the perfect relationship between the sample and theoretical quintiles. $P$-values were calculated from a one-tailed Kolomonov–Smirnov test.

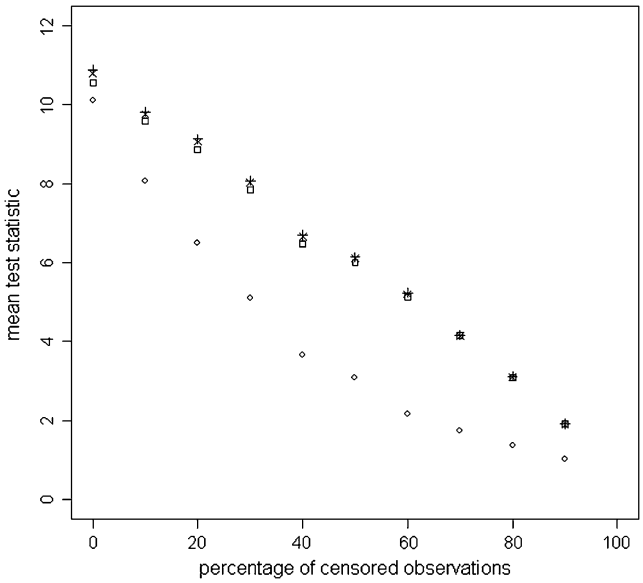


FIGURE 2.—Mean test statistics for phenotypes simulated from model 1 (Weibull) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportional hazards regression; +, Weibull proportional hazards regression; ◇, standard linear regression.

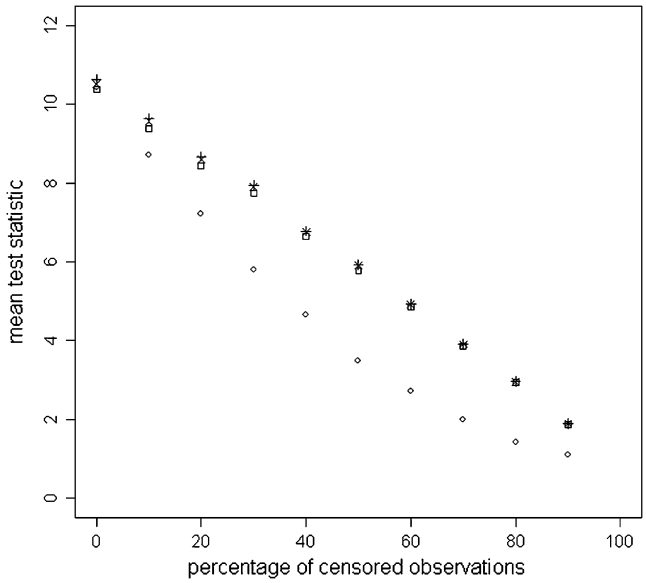


FIGURE 3.—Mean test statistics for phenotypes simulated from model 2 (exponential) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportional hazards regression; +, Weibull proportional hazards regression; ◇, standard linear regression.

this is most likely due to the incorrect parameterization of the baseline hazard. As the proportion of censoring is increased the three survival analysis methods outperform the standard linear regression method. The grouped linear regression method shows the same power as the Cox proportional hazards method.

The gamma distributions used for model 3 were not highly skewed. However, under the highly skewed gamma distributions of model 4 we see a similar pattern. When analyzing data simulated under the null hypothesis the Weibull proportional hazards model again gives test statistics that are not distributed as a chi-square distribution with 1 d.f. ($P = 0.0007$). With no censoring in the sample simulated under the alternative hypothesis the Weibull proportional hazards method is the least powerful mode of analysis (Figure 6). Again, the grouped linear regression approach is of equal power to the Cox proportional hazards model, regardless of the censoring proportion. When simulated phenotypes include censored observations, the least powerful method was the standard linear regression method.

**TABLE 6**

**Time taken to fit 100 models using R (v. 2.1.1)**

| | Relative procedure time | Actual procedure time (sec) |
|---|---|---|
| Standard linear regression | 1 | 0.22 |
| Cox proportional hazards regression | 7.2 | 1.59 |
| Weibull proportional hazards regression | 19 | 4.18 |
| Grouped linear regression[a] | 1.5 | 0.33 |

Procedure times were calculated using the proc.time() function in R.

[a] Number of groups ($k$) =10, group survival proportion ($s$) = 0.6.

## DISCUSSION

We have demonstrated the relative power of a novel grouped linear regression method for mapping QTL using censored survival time data. This method is not only as powerful as the techniques currently available but is also robust. To further check the robustness of the method we simulated two alternative censoring mechanisms. Both methods simulated only within-study censoring. No differences in the relative powers of the methods were observed (results not shown).

The grouped linear regression method is not only as robust and powerful as the Cox proportional hazards method but also computationally much faster. For genomewide scans of several thousand test positions and many potential models, the savings in computation time would be considerable. The reduction in computation time would be further appreciated when carrying out permutation testing or bootstrapping. For example, consider a genome of 3000 cM with linkage analysis carried out at intervals of 1 cM. If one was to carry out 1000 permutation tests on the sample then the grouped linear regression method, assuming that the relative times are the same as those shown in our simulations, would be 10.5 hr faster than the Cox model and 32 hr faster than the Weibull proportional hazards method.

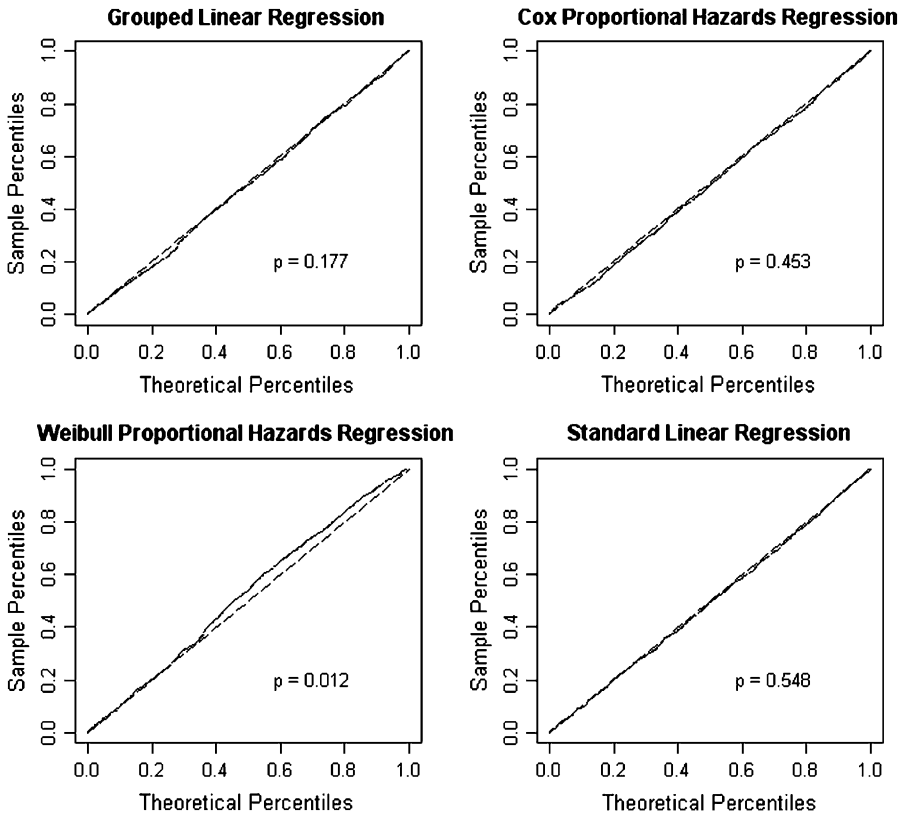


FIGURE 4.—Quantile–quantile plots for phenotypes simulated from model 3 (gamma) under the null hypothesis. Theoretical percentiles were calculated from a chi-square distribution with 1 d.f. Sample percentiles were calculated empirically by simulating 1000 replicates of 1000 individuals under the null hypothesis. The dashed line denotes the perfect relationship between the sample and theoretical quintiles. *P*-values were calculated from a one-tailed Kolomonov–Smirnov test.

The exact savings in computation time will vary as the computation times given here for each model are estimates and will vary between software packages. We expect the number of individuals in a study, the number of groups into which the continuous survival times are split, and the proportion of tied observations in the sample to have an effect on the relative run times of the various methods. In this study we have directly compared methods that all utilize the expected genotype probabilities at a given point when testing for a QTL at that location. Therefore, in terms of computation time, all these models have an advantage over those models that fit a more complete, and computationally complex, maximum-likelihood mixture model approach, such as the methods of SYMONS *et al.* (2002), DIAO *et al.* (2004), or DIAO and LIN (2005).

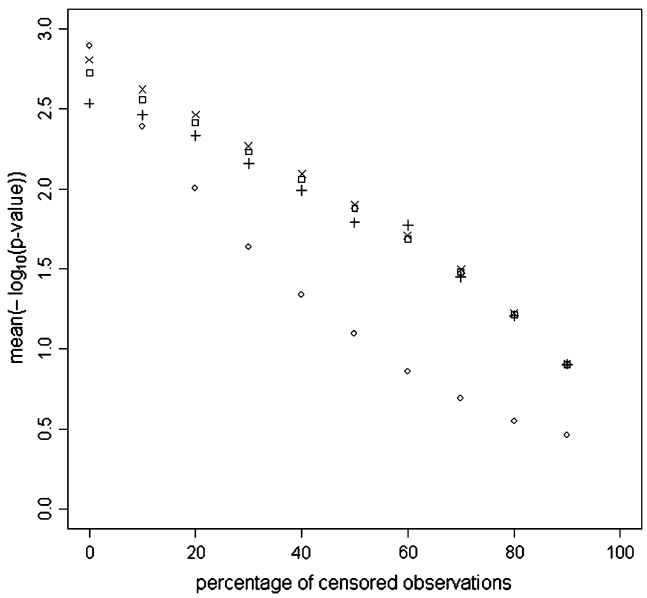


FIGURE 5.—Empirical power [shown via the mean $-\log_{10}$ (*P*-value)] for phenotypes simulated from model 3 (gamma) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportional hazards regression; +, Weibull proportional hazards regression; ◇, standard linear regression.

Recently, MORENO *et al.* (2005) compared the power of the Weibull and Cox proportional hazards methods to standard Gaussian methods for mapping QTL in survival traits. Empirical survival times for an $F_2$ population were sampled from a real data set collected on salmonella resistance in mice. Data were transformed differently for each analysis method. Differing proportions of additive/dominance effects and censoring proportions were simulated. When censored observations were included in the sample, MORENO *et al.* (2005) report a significant difference in power between the proportional hazard models and standard linear regression methods for all simulation sets. This is consistent with our findings. MORENO *et al.* (2005) also report a significant difference in power between the proportional hazards methods and standard QTL mapping methods when all observations are fully observed. This relationship was most significantly observed with an

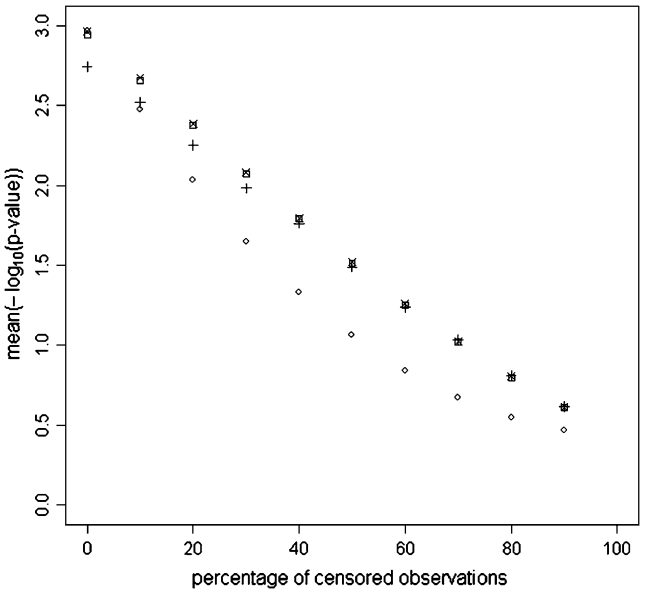


FIGURE 6.—Empirical power [shown via the mean $-\log_{10}$ (*P*-value)] for phenotypes simulated from model 4 (gamma) distributions with varying proportions of censoring. □, grouped linear regression; ×, Cox proportional hazards regression; +, Weibull proportional hazards regression; ◇, standard linear regression.

additive effect of 0.30 and an absence of dominance effects. Proportional hazards methods approximately provided a power of 0.60 while standard QTL mapping methods provided a power of 0.42 at the 95% level. MORENO *et al.* (2005) report that for larger additive effect sizes, and in the presence of dominance, the difference between the proportional hazards and standard QTL mapping methods disappears. When we simulated data using model 1 (Weibull) we noted a slight reduction in the mean test from the standard linear regression method when compared to the survival analysis methods. This reduction was much less marked than that reported by MORENO *et al.* (2005). When we simulated uncensored data from model 2 (exponential) we detected no significant differences in terms of mean test statistic when comparing the standard linear regression method to the survival analysis methods used here. Furthermore, when data were simulated from a gamma distribution with no censoring, the standard linear regression method was shown to have the most power at the 95% level. Reducing the effect of the QTL on the hazard to 3% (increased risk for genotype *Qq* when compared to that of *qq*) allowed us to compare more directly with the simulations performed by MORENO *et al.* (2005). The reduction in QTL effect size dropped the power to detect linkage to ~0.6 at the 95% level, similar to the power achieved by MORENO *et al.* (2005). No significant difference was observed between the power of proportional hazards and that of standard methods. We changed the shape (ρ) of the Weibull distributions from which the uncensored phenotypic data were simulated. We again found no difference in the power to detect linkage using either a proportional hazards framework or a standard QTL-mapping procedure when simulating from Weibull distributions with shape $\rho = 4$ or $\rho = 6$. Due to the way in which MORENO *et al.* (2005) simulated and transformed phenotypic data it is difficult to further examine the difference between the two study findings.

The grouped linear regression method uses binary survival indicators similar to those used by MADGWICK and GODDARD (1989). A grouped approach was natural for their data set as their data consisted of survival through a series of different lactation periods, which are predefined, biologically relevant time periods. A similar method was adopted by MEUWISSEN *et al.* (2002) to estimate breeding values for functional survival, in a simulated dairy cattle data set. The authors compared both a linear and a logistic regression method to a proportional hazards model and found no significant difference in estimated breeding value. In this study, we have developed a grouped linear regression method for survival traits with a continuous distribution. We have shown that if a sufficient number of time periods are chosen then little is lost in the way of power by grouping the data. The conditional survival probabilities (group survival indicators) are directly related to the hazard for a particular interval. Asymptotically, with many time intervals and a large number of observations per group, the conditional survival probabilities, scaled by the probability of survival until that time, are simply discrete versions of the continuous hazard. Just as the hazard function is a continuous approximation of a discrete observation (survival or death at a particular point in time), so the grouped approximation is a discrete approximation of a continuous distribution. With the grouped linear regression model the effects on the hazard are additive, whereas the usual assumption of proportional hazard models is that the effects act in a multiplicative manner. If the intervals are chosen such that the conditional survival probabilities in different time intervals are the same then the multiplicative and additive models converge.

The power of the grouped linear regression method was maximized, via simulation, prior to the comparative power analysis. However, the gain in power achieved by this is small. Our analyses demonstrate that most nonextreme values of *s* and *k* closely approximate the power provided by the optimal values. While it would be possible to carry out this optimization step before analyzing real data, it could be of relatively little benefit and time consuming. Thus, the robustness of the grouped linear regression method to the chosen number of time periods and group survival probability is encouraging.

Current mapping methods require specialized software for genomewide linkage analysis. The grouped linear regression method uses standard linear regression

methodology and thus can be implemented in many of the widely available statistical packages, including the freely available R that we used here. Expanding the grouped linear regression method to a genomewide level is straightforward. The ability to analyze genomewide marker data for linkage in freely available and easy to use packages is significant, especially if this can be done with little or no sacrifice in power.

In this study we simulated a backcross population; however, the extension of the method for other line crosses is relatively simple. Furthermore, it should be possible to extend the method to more complex situations such as the mapping of QTL with potentially censored data in outbred populations. Current methods for the mapping of QTL using censored data in general outbred populations are limited. Unlike in inbred lines where the QTL effect is fixed in both populations, not all individuals in an outbred population will segregate a given QTL. Furthermore, unlike fully inbred lines, each individual has a different genetic background effect. A random-effects QTL model based upon multiple 0/1 indicator variables would naturally fit into a linear mixed-model framework and would allow QTL analyses in general pedigrees when a proportion of observations are censored.

In summary, we have described a computationally efficient and fast method for the analysis of continuous survival data. The grouped regression method is of equal power when compared to other available methods and is robust to changes in censoring proportion and mechanism and to the underlying distribution of the phenotype.

We thank David Duffy, Bill Hill, and Mike Goddard for helpful discussion. We also thank two anonymous reviewers for helpful and constructive comments on an earlier version of the manuscript. C.A.A. is supported by a Medical Research Council postgraduate studentship.

## LITERATURE CITED

Andersson, L., and M. Georges, 2004 Domestic-animal genomics: deciphering the genetics of complex traits. Nat. Rev. Genet. **5:** 202–212.

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

Cox, D. R., 1972 Regression models in life-tables. J. R. Stat. Soc. Ser. B **34:** 187–220.

Diao, G., and D. Y. Lin, 2005 Semiparametric methods for mapping quantitative trait loci with censored data. Biometrics **61:** 789–798.

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

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

Frary, A., T. C. Nesbitt, S. Grandillo, E. Knaap, B. Cong *et al.*, 2000 fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science **289:** 85–88.

Fridman, E., T. Pleban and D. Zamir, 2000 A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484bp within an invertase gene. Proc. Natl. Acad. Sci. USA **97:** 4718–4723.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford *et al.*, 2002 Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. **12:** 222–231.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

Jeon, J. T., O. Carlborg, A. Tornsten, E. Giuffra, V. Amarger *et al.*, 1999 A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. Nat. Genet. **21:** 157–158.

Kao, C.-H., 2000 On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics **156:** 855–865.

Klein, J. P., and M. L. Moeschberger, 1997 *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York.

Knott, S. A., 2005 Regression-based quantitative trait loci mapping: robust, efficient and effective. Philos. Trans. R. Soc. B **360:** 1435–1442.

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

Madgwick, P. A., and M. E. Goddard, 1989 Genetic and phenotypic parameters of longevity in Australian dairy cattle. J. Dairy Sci. **72:** 2624–2632.

Meuwissen, T. H. E., R. F. Veerkamp, B. Engel and S. Brotherstone, 2002 Single and multitrait estimates of breeding values for survival using sire and animal models. Anim. Sci. **75:** 15–24.

Moreno, C. R., J. M. Elsen, P. Le Roy and V. Ducrocq, 2005 Interval mapping methods for detecting QTL affecting survival and time-to-event phenotypes. Genet. Res. **85:** 139–149.

Morgante, M., and F. Salamini, 2003 From plant genomics to breeding practice. Curr. Opin. Biotechnol. **14:** 214–219.

Nezer, C., L. Moreau, B. Brouwers, W. Coppieters, J. Detilleux *et al.*, 1999 An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. Nat. Genet. **21:** 155–156.

Pankratz, V. S., M. De Andrade and T. M. Therneau, 2005 Random-effects Cox proportional hazards model: genetic variance components methods for time-to-event data. Genet. Epidemiol. **28:** 97–109.

Prentice, R. L., and L. A. Gloeckler, 1978 Regression analysis of grouped survival data with application to breast cancer data. Biometrics **34**(1): 57–67.

R Development Core Team, 2005 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna (http://www.R-project.org).

Rebai, A., 1997 Comparison of methods of regression interval mapping in QTL analysis with non-normal traits. Genet. Res. **69:** 69–74.

Rebai, A., B. Goffinet and B. Mangin, 1995 Comparing power of different methods for QTL detection. Genetics **51:** 87–99.

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

Visscher, P. M., C. S. Haley and S. A. Knott, 1996 Mapping QTLs for binary traits in backcross and F2 populations. Genet. Res. **68:** 55–63.

Communicating editor: J. B. Walsh

## APPENDIX: τ CENSORING

Here we derive a value for the τ parameter used in the censoring method. For a general distribution of times to events, $f(T)$, we want to find a value τ such that a random uniform variable between 0 and τ is less than a random value for $T$ with a probability $P_c$. The general solution to this problem is

$$P(T > C) = E_T(P(T > C \mid C)) = \int_0^\infty f(T)P(T > C \mid C)dT = P_c,$$

where $C$ is the random censoring time. The probability $P(T > C|C)$ is calculated noting that all values of $T$ above $\tau$ are censored. Thus,

$$P(T > C \mid C) = \begin{cases} \int_0^T \frac{1}{\tau} dC = \frac{T}{\tau}; & T < \tau \\ 1; & T \geq \tau. \end{cases}$$

It follows that

$$\begin{aligned} P(T > C) &= \int_0^\tau f(T) T dT + \int_\tau^\infty f(T) dT \\ &= 1 - \int_0^\tau f(T) dT + \frac{1}{\tau} \int_0^\tau T f(T) dT = P_c. \end{aligned}$$

For the cases of the Weibull and gamma examined in this study, the value of $\tau$ is solved by numerical integration.