Mapping time-to-death quantitative trait loci in a mouse cross with high survival rates

Karl W. Broman[1], Victor L. Boyartchuk[2] and William F. Dietrich[2]
[1]Department of Biostatistics, Johns Hopkins University, and
[2]Howard Hughes Medical Institute/Department of Genetics, Harvard Medical School

Technical Report MS00-04
Department of Biostatistics, Johns Hopkins University

5 May 2000

We describe a method for the interval mapping of quantitative trait loci (QTLs) for a time-to-death trait when a considerable proportion of individuals fail to die. This work was motivated by a mouse cross for susceptibility to *Listeria* infection, in which 30% of the mice recovered from infection, and the other 70% died within 240 hours, with an average (standard deviation, SD) time-to-death of 106 (33) hours. We consider a single-QTL model in which a mouse with genotype g at the QTL has probability p_g of recovering from the infection, and, if the mouse dies, its log time-to-death follows a normal distribution with mean μ_g and SD σ. We describe the calculation of three lod scores: LOD(p), to test the hypothesis that the p_g are equal; LOD(μ), to test the hypothesis that the μ_g are equal; and LOD(p,μ), to test the combined hypothesis that both the p_g and the μ_g are constant in g.

Introduction

The inbred mouse strains BALB/cByJ and C57BL/6ByJ show striking differences in their susceptibility to *Listeria monocytogenes* infection. Intravenous infection with 2–5$\times 10^4$ cfu of *Listeria* leads invariably to death within 72 hours for BALB/cByJ animals, while similarly infected C57BL/6ByJ mice survive indefinitely. We sought to map the quantitative trait loci (QTLs) contributing to this difference (Boyartchuk et al., submitted). Among 116 F2 intercross progeny infected with 3×10^4 cfu *Listeria*, 70% died within 240 hours, with an average (SD) time-to-death of 106 (33) hours; the other 30% recovered from the infection. Traditional interval mapping, which relies on a normally distributed trait, is thus not appropriate. Right-censored observations, in which time-to-death is only known to be greater than some value, are frequently seen in survival data. However, the times-to-death for the surviving mice in this cross are not really right-censored: the observed times-to-death were not truncated due to the limited length of the study, but rather the mice recovered from infection to die considerably later from other causes. Thus, this cross presents a novel quantitative trait requiring the development of new methods for QTL mapping.

We consider a single-QTL model in which a mouse with genotype g at the QTL has probability p_g of surviving the infection and, if it dies, its log time-to-death is drawn from a normal distribution with mean μ_g and SD σ. We calculate three LOD scores at each marker and at steps of 1 cM between markers: LOD(p), to test the hypothesis that the p_g are equal; LOD(μ), to test the hypothesis that the μ_g are equal; and LOD(p,μ), to test the combined hypothesis that the p_g and the μ_g are constant in g.

In the genome scan for loci contributing to variation in susceptibility to *Listeria*, we identified three loci, on chromosomes 1, 5, and 13, giving significant LOD scores. These loci were observed to display different modes of action. The locus on chromosome 13 showed effects on both the p_g and the μ_g, while the locus on chromosome 5 affected largely the p_g, and the locus on chromosome 1 affected largely the μ_g.

Material and Methods

Data

Each of 120 age-matched (9 weeks of age) female BALB/cByJ $\times$ C57BL/6ByJ intercross (CB6F2/ByJ) mice were infected by intravenous injection with 300 μl of 3×10^4 cfu *Listeria monocytogenes* (strain 1040s), and were monitored to discover their time of death to within eight hours. All moribund animals were recorded as dead. Animals surviving past 240 hours were considered recovered from the disease. Four mice were excluded from the

Address for correspondence: Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205. E-mail: kbroman@jhsph.edu

analysis for technical reasons, such as under or overdosing of the animals. Genomic DNA was extracted from all CB6F2/ByJ animals. MapPairs primer sets (Research Genetics) for genetic mapping were selected based on an allelic difference of at least 8bp. The initial genetic markers on which the CB6F2/ByJ individuals were genotyped covered all autosomes with a spacing of ∼20 cM. Additional markers were added to regions showing linkage in a preliminary analysis.

Statistical methods

We consider a set of n F_2 intercross progeny derived from inbred mouse strains with genotypes CC and BB. Let g_i (= CC, BC, or BB) denote the genotype of mouse i at a putative QTL. Let z_i = 1 or 0 according to whether or not mouse i survived, and let y_i denote its log time-to-death (left undefined in the case that $z_i = 1$). We assume that the (g_i, y_i, z_i) are mutually independent, that $\Pr(z_i = 1|g_i = g) = p_g$, and that $y_i|(g_i = g) \sim \text{normal}(\mu_g, \sigma)$. To simplify some of the equations, let $x_{ig} = 1$ if $g_i = g$ and 0 otherwise.

In the case of complete genotype data at the putative QTL, the log likelihood for the parameters $\theta = (p_g, \mu_g, \sigma)$ is the following:

$$L(\theta) = \sum_i \sum_g z_i x_{ig} \log p_g + (1 - z_i) x_{ig} \log(1 - p_g) + (1 - z_i) \log f(y_i; \mu_g, \sigma)$$

where $f(y; \mu, \sigma)$ is the density of a normal distribution with mean μ and SD σ. Maximum likelihood estimates (MLEs) for the parameters may be obtained in closed form:

$$\hat{p}_g = \frac{\sum_i x_{ig} z_i}{\sum_i x_{ig}}$$

$$\hat{\mu}_g = \frac{\sum_i y_i x_{ig} (1 - z_i)}{\sum_i x_{ig} (1 - z_i)}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_i \sum_g (y_i - \hat{\mu}_g)^2 x_{ig} (1 - z_i)}{n}}$$

In words, $\hat{p}_g$ is the proportion of mice with genotype g that survived infection, $\hat{\mu}_g$ is the average of the log time-to-death among the non-survivor mice that had genotype g, and $\hat{\sigma}$ is the usual pooled estimate of the SD.

When there is missing genotype data at the putative QTL (especially for positions between markers), we follow the approach of Lander and Botstein (1989) and use a form of the EM algorithm (Dempster et al. 1977). For each individual, we calculate $q_{ig} = \Pr(g_i = g|\boldsymbol{m}_i)$ (where $\boldsymbol{m}_i$ is the multipoint marker data for mouse i), under the assumption of no crossover interference, with the use of the hidden Markov model (HMM) technology of Baum et al. (1971). We used an incomplete penetrance model for the genotypes, as described by Lincoln and Lander (1992), in order to account for possible genotyping errors, with an assumed genotyping error rate of $\epsilon = 0.001$. (However, this was found to have little effect on the results.)

To begin the EM algorithm, we form initial estimates $\hat{\theta}^{(0)} = (\hat{p}_g^{(0)}, \hat{\mu}_g^{(0)}, \hat{\sigma}^{(0)})$ using the equations for the case of complete genotype data (above), replacing x_{ig} with the probabilities q_{ig}. The EM algorithm is then performed iteratively. At iteration $(k + 1)$, we calculate

$$w_{ig}^{(k+1)} = \Pr(g_i = g|y_i, z_i, \boldsymbol{m}_i, \hat{\theta}^{(k)})$$

$$= \begin{cases} \frac{q_{ig} \hat{p}_g^{(k)}}{\sum_g q_{ig} \hat{p}_g^{(k)}} & \text{if } z_i = 1 \\ \frac{q_{ig}(1 - \hat{p}_g^{(k)}) f(y_i; \hat{\mu}_g^{(k)}, \hat{\sigma}^{(k)})}{\sum_g q_{ig}(1 - \hat{p}_g^{(k)}) f(y_i; \hat{\mu}_g^{(k)}, \hat{\sigma}^{(k)})} & \text{if } z_i = 0 \end{cases}$$

New estimates $\hat{\theta}^{(k+1)} = (\hat{p}_g^{(k+1)}, \hat{\mu}_g^{(k+1)}, \sigma^{(k+1)})$ are then formed using the equations for the case of complete genotype data, this time using $w_{ig}^{(k+1)}$ in place of x_{ig}.

We calculate the MLEs and the maximum log likelihood under four hypotheses: H_0: the p_g and the μ_g are equal; H_1: the p_g are equal but the μ_g vary; H_2: the μ_g are equal but the p_g vary; and H_3: the p_g and μ_g each vary. Let L_i be the maximum log (base 10) likelihood under hypothesis H_i. From these, we calculate three LOD scores: $\text{LOD}(\mu) = L_3 - L_2$, to test the hypothesis that the μ_g are equal; $\text{LOD}(p) = L_3 - L_1$, to test the hypothesis that the p_g are equal; and $\text{LOD}(p, \mu) = L_3 - L_0$, to test the combined hypothesis that the p_g and μ_g are constant in g.

The calculation of the MLEs under the unrestricted hypothesis H_3 is described above. The corresponding log likelihood $L(\theta)$ is

$$\sum_i \log \left\{ \sum_g q_{ig} p_g^{z_i} (1 - p_g)^{(1 - z_i)} f(y_i; \mu_g; \sigma)^{(1 - z_i)} \right\}$$

Under H_0, that the p_g and μ_g are constant in g, things are considerably simpler. We have the likelihood

$$L(\theta) = \log p_0 \sum_i z_i + \log(1 - p_0) \sum_i (1 - z_i) + \sum_i \log(1 - z_i) f(y_i; \mu_0; \sigma_0)$$

And thus we obtain the following MLEs

$$\hat{p}_0 = \frac{1}{n} \sum_i z_i$$

$$\hat{\mu}_0 = \frac{\sum_i y_i (1 - z_i)}{\sum_i (1 - z_i)}$$

$$\hat{\sigma}_0 = \sqrt{\frac{\sum_i (y_i - \hat{\mu}_0)^2 (1 - z_i)}{n}}$$

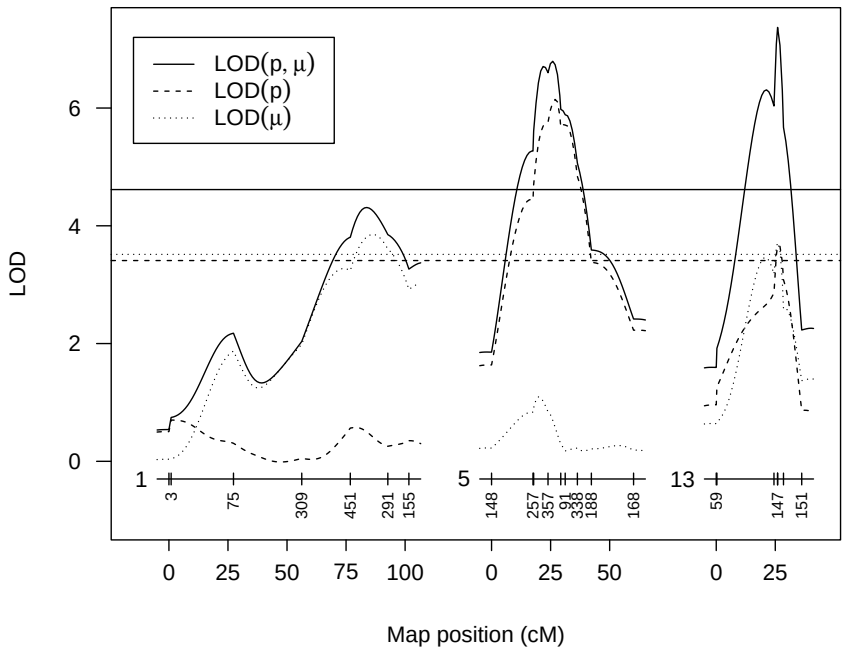

Figure 1: LOD curves from the interval mapping of the *Listeria* survival trait for chromosomes 1, 5, and 13. Horizontal lines indicate 5% genomewide significance thresholds (estimated by permutation tests).

For the hypothesis H_1, that the p_g are equal, the MLE of the common p_0 is the same as that under H_0, while the MLEs of μ_g and σ are obtained by the EM algorithm, with the modification that in the calculation of the $w_{ig}^{(k+1)}$, $\hat{p}_0$ is used in place of the $\hat{p}_g^{(k)}$. Similarly, for the hypothesis H_2, that the μ_g are equal, the MLEs of σ and the common μ_0 are simply those calculated under H_0, while the MLEs of the p_g are obtained by the EM algorithm, with $\hat{\mu}_0$ and $\hat{\sigma}_0$ used in place of $\hat{\mu}_g^{(k)}$ and $\hat{\sigma}$ in the calculation of the $w_{ig}^{(k+1)}$.

The three LOD scores, and maximum likelihood estimates under the four models, were calculated at each marker and at 1 cM steps between markers. (The likelihood and MLEs under the hypothesis H_0 need to be calculated only once.) Note that in the case of complete genotype data, the likelihoods for p and μ separate, and, as a consequence, $\text{LOD}(p, \mu) = \text{LOD}(p) + \text{LOD}(\mu)$. In the more common case of at least partially missing data, this no longer holds.

Empirical genomewide significance thresholds were calculated by permutation tests (Churchill and Doerge 1994), using 1000 permutations.

Genotype-specific averages of the time-to-death (displayed in Table 1, below) were estimated on the log scale and then converted back into the original scale: $e^{\bar{y}}$, where $\bar{y}$ is the average $\log_e$ time-to-death. Estimated standard errors were calculated as $e^{\bar{y}}\ \hat{\text{SE}}(\bar{y})$.

Results

Empirical 5% significance thresholds, based on 1000 permutations, were estimated to be 3.41, 3.51, and 4.62, for LOD(p), LOD(μ), and LOD(p,μ), respectively.

Our analysis revealed significant loci on chromosomes 1, 5, and 13, near markers D1Mit291, D5Mit357, and D13Mit147. LOD curves for these chromosomes are displayed in Figure 1.

These loci were observed to have different modes of action. The c13 locus showed effects on both the p_g and the μ_g, while the c5 locus affected largely the p_g, and the c1 locus affected largely the μ_g. The C57BL/6ByJ allele at the c13 locus was approximately dominant and was associated with both an increase in the probability of survival and in the average time-to-death. At the c5 locus, the alleles appeared codominant, and they influenced the probability of survival in the opposite direction: individuals with genotype CC at D5Mit357 showed a greater chance of survival than those with genotype BB.

Table 1 displays the joint effects of the loci at D5Mit357 and D13Mit147. (Note that complete genotype data are available for these two markers.) These results are surprising given the phenotypes of the parental lines. The BALB/cByJ mice (genotype CC) all died with 72 hours, and the C57BL/6ByJ mice all survived; however, 3/8 of the mice that had genotype CC at both loci survived, and 5/6 of the mice that had genotype BB at both loci died. This suggests the presence of additional

Table 1: Effects of *Listeria* susceptibility loci on survival and average time-to-death of infected animals

		D13Mit147							
		CC		BC		BB		Overall	
		$\hat{p}_g$	(n)	$\hat{p}_g$	(n)	$\hat{p}_g$	(n)	$\hat{p}_g$	(n)
	CC	0.38	(8)	0.61	(18)	1.00	(4)	0.60	(30)
D5Mit357	BC	0.04	(23)	0.52	(23)	0.33	(9)	0.29	(55)
	BB	0.00	(13)	0.00	(12)	0.17	(6)	0.17	(31)
	Overall	0.09	(44)	0.43	(53)	0.42	(19)	0.30	(116)

$\hat{p}_g$, proportion of animals with genotype g which survived the infection; n, number of animals with indicated genotype; C indicates the BALB/cByJ allele, and B indicates the C57BL/6ByJ allele

		D13Mit147							
		CC		BC		BB		Overall	
		$\hat{\mu}_g$	(SE)	$\hat{\mu}_g$	(SE)	$\hat{\mu}_g$	(SE)	$\hat{\mu}_g$	(SE)
	CC	106	(20)	125	(9)	—		117	(10)
D5Mit357	BC	92	(5)	121	(12)	106	(6)	102	(5)
	BB	84	(2)	112	(11)	100	(9)	97	(5)
	Overall	91	(3)	118	(6)	103	(5)	102	(3)

$\hat{\mu}_g$, mean log time-to-death (returned to the original scale) of non-surviving animals; SE, estimated standard error

loci (of which the c1 locus is one example).

Discussion

We have described a method for mapping QTLs for a time-to-death trait when many individuals fail to die. This approach is appropriate when the survivor mice have truly recovered from the initial infection, so that their undefined times-to-death cannot be treated as right-censored.

The presence of such survivor individuals precludes the use of traditional interval mapping. While nonparametric approaches, such as that described by Kruglyak and Lander (1995), are an option, such approaches are not ideal for situations in which many individuals share a common trait value. Our approach allows separate estimates of the average time-to-death and the probability of survival, and so may reveal the separate effects of a locus on the timing of death and the chance of survival.

In our analysis method, the time-to-death is assumed to follow a lognormal distribution. One might instead consider a semi-parametric approach, for example, with use of the Cox proportion hazards model (Cox 1972), in which the hazard function for mice with genotype CC at a putative QTL is left unspecified (the baseline hazard), while the hazard functions for the other two genotypes are proportional to the baseline hazard. (The hazard function is defined as $h(x) = f(x)/[1 - F(x)]$, where $f(x)$ is the density for the time-to-death distribution, and $F(x)$ is the corresponding cumulative distribution function.) While such an approach would eliminate the reliance on the log-normal assumption, it would likely have a minimal effect on the results.

References

Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann Math Stat 41:164–171

Boyartchuk VL, Broman KW, Mosher RE, D'Orazio S, Starnbach M, Dietrich WF (2000) Multigenic control of *Listeria monocytogenes* susceptibility in mice. Submitted

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

Cox DR (1972) Regression models and life tables. J Roy Stat Soc B 34:187–220

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

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

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lincoln SE, Lander ES (1992) Systematic detection of errors in genetic linkage data. Genomics 14:604–610