DOI: 10.1534/genetics.108.099028

Multiple-Interval Mapping for Quantitative Trait Loci With a Spike in the Trait Distribution

Wenyun Li* and Zehua Chen†,1

**School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, People's Republic of China 510275 and* †*Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546*

Manuscript received November 24, 2008
Accepted for publication February 16, 2009

ABSTRACT

For phenotypic distributions where many individuals share a common value—such as survival time following a pathogenic infection—a spike occurs at that common value. This spike affects quantitative trait loci (QTL) mapping methodologies and causes standard approaches to perform suboptimally. In this article, we develop a multiple-interval mapping (MIM) procedure based on mixture generalized linear models (GLIMs). An extended Bayesian information criterion (EBIC) is used for model selection. To demonstrate its utility, this new approach is compared to single-QTL models that appropriately handle the phenotypic distribution. The method is applied to data from Listeria infection as well as data from simulation studies. Compared to the single-QTL model, the findings demonstrate that the MIM procedure greatly improves the efficiency in terms of positive selection rate and false discovery rate. The method developed has been implemented using functions in R and is freely available to download and use.

MANY statistical methods for mapping quantitative trait loci (QTL) have been developed for traits with regular distributions. These include single-interval mapping (Lander and Botstein 1989), marker-based regression (Cowen 1989; Moreno-Gonzalez 1992), composite-interval mapping (Jansen 1993; Jansen and Stam 1994; Zeng 1994), multiple-interval mapping (Kao *et al.* 1999; Kao and Zeng 2002), and methods especially for binary traits (see, *e.g.*, Xu and Atchley 1996; Visscher *et al.* 1996; Xu *et al.* 1998; Yi and Xu 2000; McIntyre *et al.* 2001).

Recently, Broman (2003) considered the traits with distribution having a spike, *i.e.*, a mixture of a regular distribution and a single-point mass. This type of trait is common in survival analysis and tumor studies (see, *e.g.*, Boyartchuk *et al.* 2001; Hunter *et al.* 2001). Broman (2003) studied several single-QTL methods. The common feature of these methods is that putative QTL are considered one at a time. The single-QTL methods can be efficient for identifying QTL-bearing chromosomes. But if they are used to identify individual QTL, there is a potential to commit a high false discovery rate due to the existence of spurious genotype correlations between loci not in linkage disequilibrium (LD) with QTL and those in LD with QTL.

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.108.099028/DC2.

[1]*Corresponding author:* Department of Statistics and Applied Probability, National University of Singapore, 3 Science Dr. 2, Singapore 117546. E-mail: stachenz@nus.edu.sg

A natural alternative to single-QTL methods is to consider multiple QTL simultaneously. In this article, we consider a multiple-interval mapping (MIM) procedure based on mixture generalized linear models (GLIM) for traits with the spike feature. An EM algorithm for the mixture GLIM and a forward procedure using an extended Bayesian information criterion (EBIC) (see Chen and Chen 2008) are developed. The MIM procedure is illustrated with the Listeria data (Boyartchuk *et al.* 2001) that were analyzed by Broman (2003), using the single-QTL methods mentioned above. Simulation studies are carried out to compare the MIM procedure with the single-QTL methods.

METHODS

For simplicity, we consider backcross designs without loss of generality. Let the marker genotypes of an interval be coded by x as follows: $x = 1$, if both markers are homozygous; $x = 2$, if the left one is homozygous and the right one is heterozygous; $x = 3$, if the left one is heterozygous and the right one is homozygous; and $x = 4$, if both markers are heterozygous. Let y_i be the trait value of individual i and x_{ij} be its genotype code on interval j. Denote by δ_{ij} the unobservable genotype of individual i at a putative QTL on interval j, where $\delta_{ij} = 1$, if the genotype is homozygous, and 0 otherwise. The probability that $\delta_{ij} = 1$ is determined by x_{ij} and r_j, where r_j is the recombination fraction between the left marker and the putative QTL of interval j. Let $p(r_j, x_{ij})$ denote this probability.

The multiple-QTL mixture GLIM: Consider any m intervals. Let $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{im})$. Assume that the conditional density function of y_i given $\boldsymbol{\delta}_i$ is

$$[1 - \pi(\boldsymbol{\delta}_i)]^{1-z_i}[\pi(\boldsymbol{\delta}_i)\phi(y_i, \boldsymbol{\delta}_i)]^{z_i}, \tag{1}$$

where $z_i = I\{y_i \neq 0\}$, $\pi(\boldsymbol{\delta}_i) = P(z_i = 1)$, and ϕ is the density function of an exponential family distribution. Then the joint density of $\{(y_i, \boldsymbol{\delta}_i), i = 1, \ldots, n\}$ is given by

$$f(\mathbf{y}, \boldsymbol{\Delta}) = \prod_{i=1}^{n}\prod_{j=1}^{m} p^{\delta_{ij}}(r_j, x_{ij})[1 - p(r_j, x_{ij})]^{1-\delta_{ij}} \times [1 - \pi(\boldsymbol{\delta}_i)]^{1-z_i}[\pi(\boldsymbol{\delta}_i)\phi(y_i, \boldsymbol{\delta}_i)]^{z_i}, \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n)$. The marginal density of $\mathbf{y}$ is obtained by summing up the second product over all possible values of the δ_{ij}'s, which gives rise to a mixture of 2^m components of form (1).

Consider the general exponential family form of $\phi(y_i, \boldsymbol{\delta}_i)$,

$$\phi(y_i, \boldsymbol{\delta}_i) = \exp\left\{\frac{y_i\theta_\mu(\boldsymbol{\delta}_i) - b_\mu(\theta_\mu(\boldsymbol{\delta}_i))}{\tau} + C(\tau, y_i)\right\},$$

where τ is a dispersion parameter common to all i, and b_μ is a monotone function related to the mean $\mu(\boldsymbol{\delta}_i)$ of the distribution by $\mu(\boldsymbol{\delta}_i) = (\partial b_\mu/\partial\theta_\mu)(\theta_\mu(\boldsymbol{\delta}_i))$. Let $g_1(\mu(\delta_i))$ be the link function that connects $\mu(\boldsymbol{\delta}_i)$ with a linear predictor $\eta_\mu(\boldsymbol{\delta}_i)$ as $\eta_\mu(\boldsymbol{\delta}_i) = g_1(\mu(\delta_i))$. If only the main effects of the QTL are considered, $\eta_\mu(\boldsymbol{\delta}_i) = \beta_{\mu 0} + \sum_{j=1}^{m}\beta_{\mu j}\delta_{ij}$. If epistasis effects among the QTL are considered, $\eta_\mu(\boldsymbol{\delta}_i) = \beta_{\mu 0} + \sum_{j=1}^{m}\beta_{\mu j}\delta_{ij} + \sum_{1\le j<k\le m}\gamma_{\mu jk}\delta_{ij}\delta_{ik}$. Similarly, let $\pi(\boldsymbol{\delta}_i)$ be related to a linear predictor $\eta_\pi(\delta_i)$ through another link function g_2. The linear predictor η_π has the same structure as η_μ. For example, in the main-effect-only model, $\eta_\pi(\boldsymbol{\delta}_i) = \beta_{\pi 0} + \sum_{j=1}^{m}\beta_{\pi j}\delta_{ij}$. A common choice for g_2 is the logistic link $g_2(\pi) = \log(\pi/(1-\pi))$.

The mixture GLIM described above forms the basis of the MIM procedure. For details on GLIM, the reader is referred to McCullagh and Nelder (1989).

The EM algorithm: In the EM algorithm, the unobservable QTL genotypes $\boldsymbol{\Delta}$ are treated as missing data. The pair $(\boldsymbol{\Delta}, \mathbf{y})$ is considered as the complete data and $\mathbf{y}$ as the incomplete data. The parameters to be estimated are $\boldsymbol{\beta}_\mu$, $\boldsymbol{\beta}_\pi$, the coefficient vectors in the two linear predictors, and $\mathbf{r}$, the vector of recombination fractions, as well as τ, the dispersion parameter. The EM algorithm alternates iteratively between an E-step and an M-step. In an E-step, the conditional expectation of the log likelihood of the complete data, $E[\log f(\mathbf{y}, \boldsymbol{\Delta}) \mid \mathbf{y}; \boldsymbol{\beta}_\mu^0, \boldsymbol{\beta}_\pi^0, \mathbf{r}^0, \tau^0]$, is computed at the most updated values of $\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\pi, \mathbf{r}, \tau$. In an M-step, the conditional expectation is maximized with respect to the parameters. Let $\theta_\pi(\boldsymbol{\delta}_i) = \ln(\pi(\boldsymbol{\delta}_i, \boldsymbol{\beta}_\pi)/(1 - \pi(\boldsymbol{\delta}_i, \boldsymbol{\beta}_\pi)))$ and $b_\pi(\theta_\pi) = \ln(1 + \exp(\theta_\pi))$. The log density of the complete data, $\log f(\mathbf{y}, \boldsymbol{\Delta})$, is expressed as follows:

$$\begin{aligned} L(\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\pi, \mathbf{r}, \tau) &= \sum_{i=1}^{n}\sum_{j=1}^{m}[\delta_{ij}\ln(p(r_j, x_{ij})) + (1 - \delta_{ij})\ln(1 - p(r_j, x_{ij}))] \\ &\quad + \frac{1}{\tau}\sum_{i=1}^{n} z_i[y_i\theta_\mu(\boldsymbol{\delta}_i, \boldsymbol{\beta}_\mu) - b_\mu(\theta_\mu(\boldsymbol{\delta}_i, \boldsymbol{\beta}_\mu))] \\ &\quad + \sum_{i=1}^{n}[z_i\theta_\pi(\boldsymbol{\delta}_i, \boldsymbol{\beta}_\pi) - b_\pi(\theta_\pi(\boldsymbol{\delta}_i, \boldsymbol{\beta}_\pi))] \\ &= L_0(\mathbf{r}) + \frac{1}{\tau}L_1(\boldsymbol{\beta}_\mu) + L_2(\boldsymbol{\beta}_\pi). \end{aligned}$$

Let $(k_1, \ldots, k_m)$ be an m-tuple with k_j's taking values 0 or 1. Define $\Delta_{ik_1\ldots k_m} = \prod_{j=1}^{m} I\{\delta_{ij} = k_j\}$. Let $\theta_{\mu k_1\ldots k_m} = \theta_\mu((k_1, \ldots, k_m), \boldsymbol{\beta}_\mu)$ and $\theta_{\pi k_1\ldots k_m} = \theta_\pi((k_1, \ldots, k_m), \boldsymbol{\beta}_\pi)$. Then $L_1(\boldsymbol{\beta}_\mu)$ and $L_2(\boldsymbol{\beta}_\pi)$ can be expressed as

$$L_1(\boldsymbol{\beta}_\mu) = \sum_{(k_1\ldots k_m)}[\theta_{\mu k_1\ldots k_m}\sum_{i=1}^{n} z_i y_i\Delta_{ik_1\ldots k_m} - b_\mu(\theta_{\mu k_1\ldots k_m})\sum_{i=1}^{n} z_i\Delta_{ik_1\ldots k_m}],$$

$$L_2(\boldsymbol{\beta}_\pi) = \sum_{(k_1\ldots k_m)}[\theta_{\pi k_1\ldots k_m}\sum_{i=1}^{n} z_i\Delta_{ik_1\ldots k_m} - b_\pi(\theta_{\pi k_1\ldots k_m})\sum_{i=1}^{n}\Delta_{ik_1\ldots k_m}],$$

where the sums are taken over all possible m-tuples $(k_1, \ldots, k_m)$. The E-step thus is reduced to the computation of the conditional expectations $E(\Delta_{ik_1\ldots k_m} \mid \mathbf{y})$, and the M-step is broken down into three separate maximization problems. For ease of notation, in what follows, we use the same notation for $\Delta_{ik_1\ldots k_m}$, δ_{ij}, and their respective conditional expectations. Since $L_0(\mathbf{r})$ is the sum of m sums, each of them involving a different position parameter, the maximization of $L_0(\mathbf{r})$ is further broken down into m maximization problems. Each of them can be solved easily by a grid-point search procedure. The maximization of $L_1(\boldsymbol{\beta}_\mu)$ and $L_2(\boldsymbol{\beta}_\pi)$ can be carried out by two separate iterated weighted least-squares procedures that we describe as follows.

Let $\mathbf{x}_{k_1\ldots k_m}$ be the row vector of the covariate values in the linear predictors with $\boldsymbol{\delta} = (k_1, \ldots, k_m)$. Let X be the matrix obtained by stacking the $\mathbf{x}_{k_1\ldots k_m}$'s one above another in lexicographical order; *i.e.*, the indexes $(k_1, \ldots, k_m)$ are in the order $(00\ldots00)$, $(00\ldots01)$, $(00\ldots10)$, $(00\ldots11)$, $\ldots$, $(11\ldots11)$. Define $W_\mu(\boldsymbol{\beta}_\mu)$ as the diagonal matrix with diagonal elements $w_{\mu k_1\ldots k_m}$ given by $\Delta^{[z]}_{k_1\ldots k_m}/[g_1'(\mu_{k_1\ldots k_m})]^2 b_\mu''(\theta_{\mu k_1\ldots k_m})$, where $\mu_{k_1\ldots k_m}$ is the mean value corresponding to $\boldsymbol{\delta} = (k_1, \ldots, k_m)$, $\Delta^{[z]}_{k_1\ldots k_m} = \sum_{i=1}^{n} z_i\Delta_{ik_1\ldots k_m}$, g_1' is the first derivative of g_1, and b_μ'' is the second derivative of b_μ. Define $\mathbf{z}_\mu(\boldsymbol{\beta}_\mu) = X\boldsymbol{\beta}_\mu + \mathbf{v}_\mu(\boldsymbol{\beta}_\mu)$, where $\mathbf{v}_\mu(\boldsymbol{\beta}_\mu)$ is the vector with its $(k_1\ldots k_m)$th component given by $g_1'(\mu_{k_1\ldots k_m})(\Delta^{[zy]}_{k_1\ldots k_m}/\Delta^{[z]}_{k_1\ldots k_m} - \mu_{k_1\ldots k_m})$, where $\Delta^{[zy]}_{k_1\ldots k_m} = \sum_{i=1}^{n} z_i y_i\Delta_{ik_1\ldots k_m}$. Similarly, define $W_\pi(\boldsymbol{\beta}_\pi)$ and $\mathbf{z}_\pi(\boldsymbol{\beta}_\pi)$ by replacing g_1, b_μ, $\boldsymbol{\beta}_\mu$, $\Delta^{[z]}$, and $\Delta^{[zy]}$ with $g_2, b_\pi, \boldsymbol{\beta}_\pi$, $\Delta^{[1]}$, and $\Delta^{[z]}$, respectively, where $\Delta^{[1]}_{k_1\ldots k_m} = \sum_{i=1}^{n}\Delta_{ik_1\ldots k_m}$. The M-step for updating $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_\pi$ is then realized by iteratively solving the following equations:

$$X'W_\mu(\boldsymbol{\beta}_\mu^{\mathrm{OLD}})X\boldsymbol{\beta}_\mu^{\mathrm{NEW}} = X'W_\mu(\boldsymbol{\beta}_\mu^{\mathrm{OLD}})\mathbf{z}_\mu(\boldsymbol{\beta}_\mu^{\mathrm{OLD}}),$$
$$X'W_\pi(\boldsymbol{\beta}_\pi^{\mathrm{OLD}})X\boldsymbol{\beta}_\pi^{\mathrm{NEW}} = X'W_\pi(\boldsymbol{\beta}_\pi^{\mathrm{OLD}})\mathbf{z}_\pi(\boldsymbol{\beta}_\pi^{\mathrm{OLD}}).$$

After $\boldsymbol{\beta}_\mu$ is updated, the dispersion parameter τ is updated by the average squared Pearson's residuals associated with L_1. The EM algorithm above is developed along the same line as that in Chen and Liu (2009).

Multiple-interval mapping procedure: The MIM procedure makes use of a model selection criterion adapted from the EBIC recently developed by Chen and Chen (2008). For a model with m intervals, the adapted criterion is given by

$$-2\ln L(\hat{\boldsymbol{\beta}}_\mu, \hat{\boldsymbol{\beta}}_\pi, \hat{\mathbf{r}}, \hat{\tau}) + \nu m \ln n + 2\ln\binom{M}{m}, 1 < \nu < 3,$$

where M is the total number of intervals under study. The number νm is considered as the effective number of unknown parameters in the model. For a model with m intervals, there are m components in each of $\boldsymbol{\beta}_\mu$, $\boldsymbol{\beta}_\pi$, and $\mathbf{r}$. But $\boldsymbol{\beta}_\pi$ does not play the same role as $\boldsymbol{\beta}_\mu$. Furthermore, only a portion of the data involve $\boldsymbol{\beta}_\mu$, and a position parameter cannot be counted fully as a free parameter in terms of its effect on the likelihood. For example, most backcross progenies have flanking markers that are either both homozygous or both heterozygous. In this case, the position of a putative QTL has little effect on the likelihood. A definite effective number, which is in fact dependent on the data, is difficult to determine. The ν adjusts the effective number according to the data. The choice of ν should be data dependent. In the discussion, some *ad hoc* rules and an outline of a data-driven approach to the choice of ν are provided.

We use EBIC as the model selection criterion because, as has been shown, it is consistent if the number of covariates under consideration is of the order $O(n^\kappa)$ for any κ where n is the sample size, but the ordinary BIC fails to be consistent if $\kappa > 0.5$ (see Chen and Chen 2008).

The MIM procedure starts with models containing only one interval. The model with the minimum EBIC is compared with the null model with no QTL at all. If the minimum is smaller than the EBIC of the null model, the interval contained in the model is selected and the procedure continues; otherwise, it stops. At a general step, suppose m intervals have already been selected. Then all the models containing these m intervals plus an additional one are assessed. The minimum EBIC of these models is compared with that of the previous model consisting of m intervals. If the minimum is still smaller, the additional interval corresponding to the minimum EBIC is selected, and the procedure continues; otherwise, it stops. In the above procedure, if an additional interval to be added is adjacent to any one already selected, it is skipped to avoid potential colinearity that might cause nonconvergence of the EM algorithm. To summarize, the procedure sequentially adds intervals to a tentative model if the EBIC of the model decreases. The procedure stops when the EBIC begins to increase. The intervals contained in the final model are taken as QTL-bearing ones.

EXAMPLE

The Listeria data of Boyartchuk *et al.* (2001) are reanalyzed to illustrate the MIM procedure. The data consist of the time to death following infection with *Listeria monocytogenes* of 116 F_2 mice from an intercross between the BALB/cByJ and C57BL/6ByJ strains and the mice's genotypes at 133 markers on 20 chromosomes. The result obtained is compared with that of the single-QTL two-part model method considered by Broman (2003).

The single-QTL two-part model method is implemented with threshold value of 4.93 for the LOD score. This threshold value is obtained by 10,000 permutation replicates. The intercross version of the MIM procedure is applied with $\nu = 2.5$ in EBIC. The exponential family distribution ϕ in the GLIM is taken as the normal distribution.

In the following, we use $[k, d]$ to denote a locus on chromosome k with a genetic distance d cM from the left end of that chromosome. The single-QTL two-part model method detects chromosomes 1, 5, and 13 as QTL-bearing ones. The loci at which the LOD score attains its maximum over each of the three chromosomes are, respectively, [1, 81], [5, 30.9], and [13, 26.16]. The MIM procedure detects chromosome 1, 2, 5, 6, 8, and 13 as QTL-bearing ones. The detected loci in the last step of the MIM procedure are [1, 81], [2, 3.5], [5, 29.0], [6, 13.0], [8, 10.0], [13, 26.5], and [13, 13.05]. These results are summarized in Table 1. The loci and the EBIC value at each step of the MIM procedure are given in Table 2. The positions of the loci are slightly different from step to step because they are reestimated at each step. We cannot judge which result is better in this example. In the next section, we evaluate these two methods by simulation studies.

SIMULATIONS

The genetic map of the mouse genome in the example section is used to generate the data in simulation studies; that is, the number and lengths of chromosomes and the number and positions of markers on each chromosome are kept the same as those in the mouse genome. The genetic map is provided in supporting information, File S1.

Backcross progenies are generated according to the genetic map. In each replicate of the simulation, 5 chromosomes are randomly chosen from the first 19

TABLE 1

Loci (genetic distance in centimorgans from the left end of each chromosome) indicating evidence of QTL detected by the single-QTL two-part model (TPM) and multiple interval-mapping (MIM) in the example

	Loci detected	
Chromosome	TPM	MIM
1	81	81
2	—	3.5
5	30.90	29
6	—	13
8	—	10
13	26.16	13.05, 26.5

chromosomes (the 20th chromosome is ignored since there are only two markers on it); on each of them, a QTL location is generated at random, and then the genotypes at these 5 QTL together with the 133 markers of 200 backcross progenies are generated. The trait values are generated under the assumption of no epistasis effect. The $\boldsymbol{\beta}_\pi$ is set at

$$\boldsymbol{\beta}_\pi = (1.03,\ 0.91,\ 0.75,\ 0.31,\ -0.49,\ -1.01)$$

and kept unchanged throughout. It is designed such that ~25% of the progenies will survive. Three settings of $\boldsymbol{\beta}_\mu$ are considered:

$$\boldsymbol{\beta}_\mu = (4.650, 1.605, 1.065, 0.585, -0.855, -1.410),$$
$$\boldsymbol{\beta}_\mu = (3.100, 1.070, 0.710, 0.390, -0.570, -0.940),$$
$$\boldsymbol{\beta}_\mu = (1.550, 0.535, 0.355, 0.195, -0.285, -0.470).$$

The survival times are generated by a normal distribution with mean equal to η_μ and variance 1. The three $\boldsymbol{\beta}_\mu$-values correspond to the survival time heritabilities 0.63, 0.43, and 0.16, respectively.

The 95% threshold value for the LOD is simulated as 3.35 by 10,000 replicates and used in all three settings. The LOD scores are calculated at grid points spaced 1 cM apart. Three values of ν, 1.5, 2, and 2.5, are used in the EBIC. The MIM procedure and the single-QTL two-part model method are applied to the same data. Their performances are assessed by positive selection rate (PSR) and false discovery rate (FDR). To make the comparison fair to the single-QTL two-part model method, we consider only whether or not a QTL-bearing chromosome is correctly identified, since single-QTL methods have been used mainly for this purpose. A chromosome is claimed as a QTL-bearing one if at least one of the selected loci falls in that chromosome. A claimed QTL-bearing chromosome is said to be a positive discovery if that chromosome does contain QTL. Otherwise, it is said to be a false discovery. The PSR and the FDR are defined as follows:

$$\text{PSR} = \frac{\text{Number of positive discoveries}}{\text{Total number of true QTL-bearing chromosomes}},$$

$$\text{FDR} = \frac{\text{Number of false discoveries}}{\text{Total number of claimed QTL-bearing chromosomes}}.$$

The simulation results over 500 replicates are given in Table 3. The findings are summarized as follows. With heritability 0.63, the MIM procedure has a much higher PSR with all the three ν-values, a lower FDR when $\nu = 2$ or 2.5, and a comparable FDR when $\nu = 1.5$. With heritability 0.43, the MIM procedure has higher PSR and lower or comparable FDR when $\nu = 2$ or 1.5 and lower FDR and comparable PSR when $\nu = 2.5$. We may claim that the MIM procedure is better than the single-QTL two-part model method when the heritability is moderate or high. However, in the case of heritability 0.16, the single-QTL two-part model method is better than the MIM procedure in terms of either PSR or FDR. An explanation is given below. The heritability considered in the simulation accounts only for the nonsurvival portion and the QTL effect on the survival proportion is fixed. Any QTL with a heritability as low as 0.16 is hard to detect no matter what approach is used. The fairly sizeable PSR in this case is mainly due to the QTL effect on the survival proportion. In the EBIC criterion of the MIM procedure, an overpenalization arises when the effect on the survival time is in fact negligible. This explains why the PSR of the MIM procedure is lower in this case. A remedy for the problem of overpenalization is discussed in the next section.

TABLE 2

The loci included and the corresponding EBIC value at each step of the MIM procedure in the example

Step	Loci included	EBIC
1	[13, 27]	153.02
2	[13, 26.5] [5, 28]	141.11
3	[13, 26.5] [5, 28] [1, 81]	138.78
4	[13, 26.5] [5, 28] [1, 81] [6, 14]	136.88
5	[13, 26.5] [5, 28] [1, 81] [6, 14] [2, 4]	136.74
6	[13, 26.5] [5, 29] [1, 81] [6, 14] [2, 3.5] [8, 8.5]	128.31
7	[13, 26.5] [5, 29] [1, 81] [6, 13] [2, 3.5] [8, 10] [13, 13.05]	124.48

TABLE 3

The average positive selection rate (PSR), false discovery rate (FDR), number of positive discoveries (NPD), and number of false discoveries (NFD) of the two approaches: the single-QTL two-part model (TPM) and the multiple-interval mapping (MIM), over 500 replicates in the simulation study

Heritability	Method	ν	PSR	FDR	NPD	NFD
0.63	TPM	—	0.614	0.018	3.072	0.056
	MIM	1.5	0.839	0.026	4.196	0.112
		2.0	0.778	0.008	3.888	0.030
		2.5	0.722	0.003	3.610	0.010
0.43	TPM	—	0.509	0.017	2.538	0.044
	MIM	1.5	0.626	0.026	3.130	0.082
		2.0	0.547	0.008	2.734	0.022
		2.5	0.471	0.002	2.354	0.004
0.16	TPM	—	0.293	0.033	1.112	0.038
	MIM	1.5	0.269	0.059	1.346	0.084
		2.0	0.224	0.054	1.118	0.064
		2.5	0.208	0.053	1.038	0.058

DISCUSSION

We have demonstrated that the MIM procedure compares favorably with the single-QTL two-part model method when being used to identify QTL-bearing chromosomes. It has the further advantage of identifying individual QTL with accurately estimated positions. We discuss some further issues in this section.

In the MIM procedure considered in the previous sections, we do not distinguish between the QTL effects on the spike probability and on the survival time. This might lead to an overpenalization of the EBIC if only one type of effect exists and hence result in a reduced power for QTL detection. The procedure can be modified such that, when a new interval is considered, two substeps are taken, one for the effect on the spike probability and the other for the effect on the survival time. Correspondingly, the term $\nu m \ln n$ in the EBIC is replaced by $\nu q \ln n$, where q counts the number of parameters of the model. When a new interval is considered, if only one type of effect is included, q increases by 1, and if both types of effect are included, q increases by 2.

In the simulation studies, we used different ν-values in EBIC. For smaller ν, the PSR is higher but the FDR is also higher, and vice versa. We give some *ad hoc* rules for the choice of ν here. First, different ν-values should be used and the results compared. It is usually the case that, if the heritability is relatively high, the results will be similar in a range of ν-values. If this is the case, the smallest ν in this range produces the highest PSR and comparable FDR compared with other values in the range and should be used for a final decision. If there is a big discrepancy among different values, the choice should be based on the purpose of the study. If the study is for confirmation, the FDR is a more serious concern, and a larger ν should be taken. If the study is a preliminary step to detect regions for further investigation, a smaller ν should be taken.

A data-driven approach based on the idea of model averaging and bootstrapping can be used. The approach is outlined as follows. Starting with a moderate ν-value, a set of claimed QTL together with their estimated effects is obtained. Then the following bootstrap-like procedure is carried out. A random number, say m^*, is generated from a Poison distribution with the mean as the number of claimed QTL. Then m^* loci, each on a different interval, are randomly selected from the genetic map and assigned as QTL, the effects of the QTL are generated using the estimated effects of the claimed QTL, and the trait values of individuals are generated by using the GLIM. Finally, the MIM procedure with different ν-values is applied to the generated data, and the positive discoveries and false discoveries are obtained by comparing the claimed QTL with the assigned QTL. This process is repeated for a large number of times. The numbers of positive discoveries and false discoveries are averaged to provide estimates for PSR and FDR for each of the ν-values. Then with the estimated PSR and FDR, the user can make a choice on the basis of a balanced consideration of PSR and FDR. A full development of the data-driven approach in more general settings is underway, which is beyond the scope of this article. We will report the general data-driven approach elsewhere.

The MIM procedure has been implemented using functions in the R package migtlm. The package will be updated soon to include a general function for the MIM procedure. The package can be downloaded from www.stat.nus.edu.sg/~stachenz.

LITERATURE CITED

Boyartchuk, V. L., K. W. Broman, R. E. Mosher, S. F. F. D'Orazio, M. N. Starnbach *et al.*, 2001 Multigenetic control of *Listeria monocytogenes* susceptibility in mice. Nat. Genet. **27:** 259–260.

Broman, K. W., 2003 Quantitative trait locus mapping in the case of a spike in the phenotype distribution. Genetics **163:** 1169–1175.

Chen, J., and Z. Chen, 2008 Extended Bayesian information criteria for model selection with large model spaces. Biometrika **95:** 759–771.

Chen, Z., and J. Liu, 2009 Mixture generalized linear models for multiple interval mapping of quantitative trait loci in experimental crosses. Biometrics (in press).

Cowen, N. M., 1989 Multiple linear regression analysis of RELP data sets used in mapping QTLs, pp. 113–116 in *Development and Application of Molecular Markers to Problems in Plant Genetics*, edited by T. Helentjaris and B. Burr. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Hunter, K.W., K. W. Broman, T. Le Voyer, L. Lukes, D. Gozma *et al.*, 2001 Predisposition to efficient mammary tumor metastatic progression is linked to the breast cancer metastasis suppressor gene *Brms1*. Cancer Res. **61:** 8866–8872.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics **136:** 1447–1455.

Kao, C. H., and Z. B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160:** 1243–1261.

Kao, C. H., Z. B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

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

McCullagh, P., and J. A. Nelder, 1989 *Generalized Linear Models*, Ed. 2. Chapman & Hall, London/New York.

McIntyre, L. M., C. Coffman and R. W. Doerge, 2001 Detection and location of a single binary trait locus in experimental populations. Genet. Res. **78:** 79–92.

Moreno-Gonzalez, J., 1992 Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques. Theor. Appl. Genet. **85:** 435–444.

Visscher, P. M., C. S. Haley and S. A. Knott, 1996 Mapping QTLs for binary traits in backcross and F-2 populations. Genet. Res. **68:** 55–63.

Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417–1424.

Xu, S., N. Yonash, R. L. Vallejo and H. H. Cheng, 1998 Mapping quantitative trait loci for complex binary traits using a heterogeneous residual variance model: an application to Marek's disease susceptibility in chickens. Genetica **104:** 171–178.

Yi, N., and S. Xu, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155:** 1391–1403.

Zeng, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: K. W. Broman

GENETICS

Supporting Information

http://www.genetics.org/cgi/content/full/genetics.108.099028/DC2

Multiple-Interval Mapping for Quantitative Trait Loci With a Spike in the Trait Distribution

Wenyun Li and Zehua Chen

DOI: 10.1534/genetics.108.099028

FILE S1

Multiple Interval Mapping for Quantitative Trait Loci with a Spike in the Trait Distribution

Wenyun Li and Zehua Chen

In this supporting information, we provide the genetic map used in the simulation studies of the paper. It is the genetic map of a mouse genome extracted from the Listeria data of BOYARTCHUK *et al.* (2001). The map contains 132 markers on 20 chromosomes. On each chromosome, any two adjacent markers form an interval. The markers form a total of 112 intervals. In the following tables, the intervals on each chromosome together with their lengths (in cM) and the positions of the flanking markers (indicated by genetic distance in cM from the left end of the chromosome) are provided.

TABLE S1

The genetic map of the mouse genome in the Listeria data

Chrom.	Interval	Inte-length	Left-marker	Right-marker
1	1	0.99675	0.00000	0.99675
1	2	23.85098	0.99675	24.84773
1	3	15.56588	24.84773	40.41361
1	4	9.58107	40.41361	49.99468
1	5	2.80552	49.99468	52.80020
1	6	17.31184	52.80020	70.11204
1	7	0.69438	70.11204	70.80642
1	8	9.81682	70.80642	80.62324
1	9	0.77299	80.62324	81.39623
1	10	3.53851	81.39623	84.93474
1	11	7.74920	84.93474	92.68394
1	12	0.95950	92.68394	93.64344
2	1	27.94171	0.00000	27.94171
2	2	19.16370	27.94171	47.10541
2	3	20.15644	47.10541	67.26185
2	4	10.13620	67.26185	77.39805
2	5	13.45825	77.39805	90.85630
3	1	32.47839	0.00000	32.47839
3	2	11.45964	32.47839	43.93803
3	3	13.65535	43.93803	57.59338
3	4	5.59202	57.59338	63.18540
3	5	7.65360	63.18540	70.83900
4	1	19.16072	0.00000	19.16072
4	2	16.16014	19.16072	35.32086
4	3	32.78230	35.32086	68.10316
5	1	6.10396	0.00000	6.10396
5	2	13.11939	6.10396	19.22335
5	3	0.32548	19.22335	19.54883
5	4	4.16831	19.54883	23.71714
5	5	1.78295	23.71714	25.50009
5	6	5.39656	25.50009	30.89665
5	7	0.00100	30.89665	30.89765

TABLE S2

The genetic map of the mouse genome in the Listeria data (Cont.)

Chrom.	Interval	Inte-length	Left-marker	Right-marker
5	8	2.00757	30.89765	32.90522
5	9	5.16285	32.90522	38.06807
5	10	5.95569	8.06807	44.02376
5	11	6.96095	44.02376	50.98471
6	1	8.18754	10.00000	18.18754
6	2	5.68464	18.18754	23.87218
6	3	7.22192	23.87218	31.09410
6	4	10.70096	31.09410	41.79506
6	5	3.35073	41.79506	45.14579
6	6	2.38411	45.14579	47.52990
6	7	3.71746	47.52990	51.24736
6	8	0.40337	51.24736	51.65073
6	9	3.65405	51.65073	55.30478
6	10	3.70510	55.30478	59.00988
6	11	0.36101	59.00988	59.37089
6	12	1.39155	59.37089	60.76244
7	1	18.78851	0.00000	18.78851
7	2	16.12211	18.78851	34.91062
7	3	6.11986	34.91062	41.03048
7	4	19.08361	41.03048	60.11409
7	5	11.97015	60.11409	72.08424
8	1	1.33987	0.00000	1.33987
8	2	10.08104	1.33987	11.42091
8	3	15.71975	11.42091	27.14066
8	4	5.84559	27.14066	32.98625
8	5	17.87739	32.98625	50.86364
9	1	4.21823	0.00000	4.21823
9	2	10.49742	4.21823	14.71565
9	3	12.60852	14.71565	27.32417
9	4	5.63227	27.32417	32.95644
9	5	12.37923	32.95644	45.33567
9	6	7.16837	45.33567	52.50404

TABLE S3

The genetic map of the mouse genome in the Listeria data (Cont.)

Chrom.	Interval	Inte-length	Left-marker	Right-marker
10	1	24.74745	0.00000	24.74745
10	2	15.96238	24.74745	40.70983
10	3	8.02021	40.70983	48.73004
10	4	12.32617	48.73004	61.05621
11	1	15.15394	0.00000	15.15394
11	2	11.26755	15.15394	26.42149
11	3	12.09996	26.42149	38.52145
11	4	3.63994	38.52145	42.16139
11	5	22.18342	42.16139	64.34481
12	1	6.17921	0.00000	6.17921
12	2	15.40130	6.17921	21.58051
12	3	7.50353	21.58051	29.08404
12	4	12.71165	29.08404	41.79569
12	5	12.66013	41.79569	54.45582
13	1	0.28675	0.00000	0.28675
13	2	10.07913	0.28675	10.36588
13	3	2.68395	10.36588	13.04983
13	4	0.00100	13.04983	13.05083
13	5	5.85801	13.05083	18.90884
13	6	2.10374	18.90884	21.01258
13	7	3.86273	21.01258	24.87531
13	8	1.28423	24.87531	26.15954
13	9	2.23316	26.15954	28.39270
13	10	0.00100	28.39270	28.39370
13	11	7.59337	28.39370	35.98707
14	1	23.90747	0.00000	23.90747
14	2	8.87932	23.90747	32.78679
14	3	12.76343	32.78679	45.55022
15	1	13.46195	0.00000	13.46195
15	2	5.32886	13.46195	18.79081
15	3	0.57392	18.79081	19.36473
15	4	4.54900	19.36473	23.91373

TABLE S4

The genetic map of the mouse genome in the Listeria data (Cont.)

Chrom.	Interval	Inte-length	Left-marker	Right-marker
15	5	1.21277	23.91373	25.12650
15	6	6.14957	25.12650	31.27607
15	7	11.69600	31.27607	42.97207
16	1	16.76684	0.00000	16.76684
16	2	9.46451	16.76684	26.23135
16	3	15.56766	26.23135	41.79901
17	1	11.72823	0.00000	11.72823
17	2	5.60704	11.72823	17.33527
17	3	21.51280	17.33527	38.84807
18	1	0.68560	0.00000	0.68560
18	2	16.29826	0.68560	16.98386
18	3	3.91604	16.98386	20.89990
19	1	16.36398	0.00000	16.36398
19	2	16.46537	16.36398	32.82935
19	3	11.66497	32.82935	44.49432
20	1	42.34593	0.00000	42.34593