Mapping Quantitative Trait Loci in the Case of a Spike in the Phenotype Distribution

Karl W. Broman
Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205
Manuscript received August 1, 2002
Accepted for publication November 26, 2002

ABSTRACT
A common departure from the usual normality assumption in QTL mapping concerns a spike in the phenotype distribution. For example, in measurements of tumor mass, some individuals may exhibit no tumors; in measurements of time to death after a bacterial infection, some individuals may recover from the infection and fail to die. If an appreciable portion of individuals share a common phenotype value (generally either the minimum or the maximum observed phenotype), the standard approach to QTL mapping can behave poorly. We describe several alternative approaches for QTL mapping in the case of such a spike in the phenotype distribution, including the use of a two-part parametric model and a nonparametric approach based on the Kruskal-Wallis test. The performance of the proposed procedures is assessed via computer simulation. The procedures are further illustrated with data from an intercross experiment to identify QTL contributing to variation in survival of mice following infection with Listeria monocytogenes.

METHODS
Consider \( n \) \( F_2 \) progeny from an intercross between two inbred strains. Let \( y_i \) denote the quantitative phenotype for individual \( i \). We assume, without loss of generality, that the spike in the phenotype distribution is at 0. Let \( z_i = 0 \) if \( y_i = 0 \) and \( z_i = 1 \) if \( y_i > 0 \). Consider data on
a set of genetic markers, with a known genetic map. Let \( m \) denote the multipoint marker data for individual \( i \).

Conditional and binary trait analyses: A simple approach for QTL mapping in this situation is to first analyze the quantitative phenotype, \( y_i \), using only the individuals for which \( y_i > 0 \), by standard interval mapping using a normal model (Landers and Botstein 1989), and then separately analyze the binary trait \( z_i \).

The analysis of the binary trait deserves further explanation. Xu and Atchley (1996) described maximum-likelihood estimation for a binary trait in the context of composite interval mapping (Zeng 1993, 1994). Visscher et al. (1996) and McIntyre et al. (2001) described approximate methods for analysis of binary traits. We prefer the approach of Xu and Atchley (1996). We briefly describe the special case of no marker covariates.

We consider some fixed position in the genome as the location of a putative QTL and let \( g_i = 1, 2, \) or \( 3 \), according to whether individual \( i \) has genotype AA, AB, or BB, respectively, at the QTL. Let us assume that the binary phenotypes, \( z_i \), are independent, and let \( \pi_i = \Pr(z_i = 1 | g_i = j) \). Given the marker data, \( m_i \), but not knowing the QTL genotypes \( g_i \), the \( z_i \) follow mixtures of Bernoulli distributions (analogous to the mixtures of normals that arise in standard interval mapping).

We assume that we may calculate \( p_i = \Pr(g_i = j | m_i) \), the QTL genotype probabilities, given the observed multipoint marker data. Under no crossover interference and no genotyping errors, the distribution depends only on the nearest flanking typed markers, but one may also use the approach of Lincoln and Landers (1992) to take account of the presence of genotyping errors.

The likelihood for the parameters \( \pi = (\pi_j) \), given the observed data \( (m_i, z_i) \), is then

\[
L(\pi) = \prod_i \sum_j p_i(\pi_j)^{z_i}(1 - \pi_j)^{(1-z_i)}.
\]

We obtain maximum-likelihood estimates (MLEs), \( \hat{\pi}_j \), using a form of the expectation-maximization (EM) algorithm (Dempster et al. 1977). At iteration \( s + 1 \), we have estimates of the parameters, \( \hat{\pi}^{(i)} \). In the E-step, we calculate weights for each individual and for each genotype:

\[
w_i^{(s+1)} = \Pr(g_i = j | z_i, m_i, \hat{\pi}^{(i)}) = \frac{p_i(\hat{\pi}_j^{(i)})^{z_i}(1 - \hat{\pi}_j^{(i)})^{(1-z_i)}}{\sum_j p_i(\hat{\pi}_j^{(i)})^{z_i}(1 - \hat{\pi}_j^{(i)})^{(1-z_i)}}.
\]

In the M-step, we reestimate the probabilities \( \pi_j \) as weighted proportions using the weights, \( w_i^{(s+1)} \):

\[
\hat{\pi}_j^{(s+1)} = \frac{\sum_i z_i w_i^{(s+1)}}{\sum_i w_i^{(s+1)}}.
\]

We begin the algorithm by taking \( w_i^{(0)} = p_i \) and iterate until the estimates converge, giving the MLE, \( \hat{\pi} \).

We next calculate a LOD score for the test of \( H_0: \pi_j = \pi \). First note that the MLE, under \( H_0 \), of the common probability \( \pi \) is the overall proportion, \( \hat{\pi}_0 = \sum z_i / n \). Letting \( \pi_0 = (\pi_0, \pi_0, \pi_0) \), the LOD score is

\[
\text{LOD} = \log_{10} \left( \frac{L(\hat{\pi})}{L(\hat{\pi}_0)} \right).
\]

As with standard interval mapping, the likelihood under \( H_0 \) is calculated once, while the EM algorithm is performed at each position in the genome (in practice, at 1-cM steps), producing a LOD curve for each chromosome.

Two-part model: The two separate analyses described above suggest the following two-part, single-QTL model. We again consider \( n \) \( F_2 \) progeny and some fixed position in the genome as the location of a putative QTL. Let \( y_i, z_i, g_i \), and \( m_i \) be defined as above, and again let \( p_0 = \Pr(g_i = j | m_i) \).

We assume that the \( (m_i, y_i, z_i) \) are mutually independent, that \( \Pr(z_i = 1 | g_i = j) = \pi_j \) and that \( \gamma_i(g_i = j, z_i = 1) \sim \text{normal}(\mu_j, \sigma^2) \). In other words, the probability that an individual with QTL genotype \( j \) has the null phenotype is \( 1 - \pi_j \); if this individual’s phenotype is nonnull, it follows a normal distribution with mean \( \mu_j \) depending on the QTL genotype, and with SD \( \sigma \), independent of genotype.

This model contains seven parameters, \( \theta = (\pi_1, \pi_2, \pi_3, \mu_1, \mu_2, \mu_3, \sigma) \). The likelihood function is

\[
L(\theta) = \prod_i \sum_j p_i(1 - \pi_j)^{1-y_i(\pi_j, \mu_j, \sigma)} f(y_i | \mu_j, \sigma),
\]

where \( f(y, \mu, \sigma) \) is the density function for a normal distribution with mean \( \mu \) and SD \( \sigma \).

We may again obtain MLEs with a form of the EM algorithm. Assume at iteration \( s + 1 \) we have estimates \( \hat{\theta}^{(i)} \). In the E-step, we calculate weights for each individual and each genotype:

\[
w_i^{(s+1)} = \Pr(g_i = j | y_i, z_i, m_i, \hat{\theta}^{(i)}) = \begin{cases} p_i(1 - \hat{\pi}_j^{(i)}) & \text{if } z_i = 0, \\ \frac{p_i \hat{\pi}_j^{(i)} f(y_i | \hat{\mu}_j^{(i)}, \sigma^{(i)})}{\sum_k p_k \hat{\pi}_k^{(i)} f(y_i | \hat{\mu}_k^{(i)}, \sigma^{(i)})} & \text{if } z_i = 1. \end{cases}
\]

In the M-step, we obtain revised estimates of the parameters according to the following equations:

\[
\hat{\pi}_j^{(s+1)} = \frac{\sum_i w_i^{(s+1)} z_i}{\sum_i w_i^{(s+1)}}
\]

\[
\hat{\mu}_j^{(s+1)} = \frac{\sum_i w_i^{(s+1)} y_i z_i}{\sum_i w_i^{(s+1)} z_i}
\]

\[
\hat{\sigma}^{(s+1)} = \sqrt{\frac{\sum_i (y_i - \hat{\mu}_j^{(s+1)})^2 w_i^{(s+1)} z_i}{\sum_i z_i}}.
\]

We again start the algorithm by taking \( w_i^{(0)} = p_i \) and iterate until the estimates converge, producing the MLEs, \( \hat{\theta} \).

We may calculate a LOD score for the test of \( H_0: \pi_j = \pi \).
\( \pi, \mu_i \equiv \mu \). We first note that, under \( H_0 \), the MLEs of the three parameters, \( \pi, \mu, \) and \( \sigma \), are

\[
\hat{\pi}_0 = \frac{\sum z_i}{n} \\
\hat{\mu}_0 = \frac{\sum z_i \hat{\mu}_i}{\sum z_i} \\
\hat{\sigma}_0 = \sqrt{\frac{\sum (y_i - \hat{\mu}_0)^2 z_i}{\sum z_i}}.
\]

In other words, \( \hat{\pi}_0 \) is the proportion of individuals with a positive phenotype, and \( \hat{\mu}_0 \) and \( \hat{\sigma}_0 \) are the sample mean and SD, among individuals with positive phenotypes. Letting \( \hat{\theta}_0 = (\hat{\pi}_0, \hat{\mu}_0, \hat{\sigma}_0, \hat{\sigma}_{i0}) \), the LOD score is

\[ \text{LOD} = \log_{10}(L(\hat{\theta})/L(\hat{\theta}_0)). \]

Note that in the case of complete QTL genotype information (i.e., when the putative QTL is at a marker that has been fully typed), the \( p_j \) are all either 1 or 0, and the two parts of the model separate fully. As a result, the MLEs under the two-part model are exactly those obtained by the two separate analyses (the analysis of the binary trait and the conditional analysis of the quantitative trait, for those individuals with nonzero phenotype). Further, the LOD score for the two-part model is simply the sum of the LOD scores from the two separate analyses.

**Nonparametric analysis:** Kruglyak and Lander (1995) described an extension of the Wilcoxon rank-sum test for nonparametric interval mapping in a backcross. The rank-sum test is a nonparametric version of the two-sample \( \chi^2 \)-test. In the case of an intercross, they suggested tests for the additive or dominant effects at a putative QTL. An alternative approach is to extend the Kruskal-Wallis test statistic, a nonparametric version of a one-way analysis of variance, for the comparison of two or more samples (e.g., see Lehmann 1975). We describe such an extension below.

Rank the phenotypes, \( y_i \), from 1, \ldots, \( n \), and let \( R_i \) denote the rank for individual \( i \). In the case of ties, use the average rank within each group of ties. We again consider some fixed position in the genome as the location of a putative QTL, and let \( p_j = \Pr(g = j|m) \), the QTL genotype probabilities for individual \( i \), given the available multipoint marker data. Whereas, in the Kruskal-Wallis test statistic, one considers the sum of the ranks within each group, here the exact assignment of individuals to QTL genotype groups is not known; rather, individual \( i \) has prior probability \( p_j \) of belonging to group \( j \). We follow the approach of Kruglyak and Lander (1995) and consider the expected rank sum, \( S_j = \sum \hat{p}_j R_i \). We then consider the statistic

\[ H = \sum_j \frac{(n - \sum \hat{p}_j) [(S_j - E_{0j})^2 / V_{0j}] \left[ \sum \hat{p}_j R_i - n + 1 \right]^2}{n(n + 1)} \]

where \( E_{0j} \) and \( V_{0j} \) are the mean and variance of \( S_j \) under the null hypothesis of no linkage, considering the \( p_j \) as fixed. After some algebra, we obtain the formula

\[ H = \frac{12}{n(n + 1)} \sum_j (n - \sum \hat{p}_j) \left( \sum \hat{p}_j \right)^2 \left[ \sum \hat{p}_j R_i - n + 1 \right]^2 \left[ \sum \hat{p}_j R_i - n - 1 \right]^2 \]

In the case that the putative QTL is at a fully typed genetic marker, the \( p_j \) will all be 0 or 1, and the above statistic reduces to the Kruskal-Wallis test statistic.

Kruglyak and Lander (1995) had randomized any tied phenotypes, a reasonable approach in the case of very few ties. In our application, however, a large proportion of the individuals share a common phenotype. Thus, rather than randomizing ties, we assign the average rank to each individual within a set of tied phenotypes. A standard correction for the case of ties is to use the statistic \( H' = H/D \), where \( D = 1 - \sum (t^3 - t)/n(n - n) \), with \( t_k \) being the number of values in the \( k \)th group of ties. Note that if there are no ties, \( D = 1 \) and so \( H' = H \). [Of course, if one uses a permutation test (Churchill and Doerge 1994) to obtain the genome-wide significance threshold, as we recommend, the correction factor is unnecessary.] As the nonparametric statistic \( H' \) follows, approximately, a \( \chi^2 \) distribution under the null hypothesis of no linkage, we convert the statistic to the LOD scale by taking \( \text{LOD} = H'/2 \ln 10 \).

**EXAMPLE**

To illustrate our methods, we consider the data of Boyartchuk et al. (2001), on the time to death following infection with *L. monocytogenes* in 116 intercross mice. Approximately 30% of the mice recovered from the infection and survived to the end of the experiment (264 hr).

![Histogram of survival time, following infection with *Listeria monocytogenes*, of 116 intercross mice. Approximately 30% of the mice recovered from the infection and survived to the end of the experiment (264 hr).](Image 341x637 to 533x743)
our two-part model (“two-part”); and the nonparametric interval-mapping method based on the Kruskal-Wallis test statistic (“NP”).

Genome-wide LOD thresholds were obtained by permutation tests (Churchill and Doerge 1994), using 11,000 permutation replicates. The estimated 95% genome-wide LOD thresholds for the four methods, binary, QT, two-part, and NP, were 3.54, 3.96, 4.91, and 3.27, respectively. The estimated standard errors (SEs) for these thresholds were \( \approx 0.02 \).

Because of the large differences in the LOD thresholds for the four methods, we converted the LOD curves to a common scale, the estimated experiment-wise \( P \) values derived from the permutation tests. The results indicated evidence for QTL on chromosomes 1, 5, 13, and 15. In Figure 2, the statistic \(-\log_{10} P\) for each method is displayed for these selected chromosomes.

The locus on chromosome 1 appears to have an effect only on the average time to death among the nonsurvivors. The locus on chromosome 5 appears to have an effect only on the chance of survival. The loci on chromosomes 13 and 15 have an effect on both the chance of survival and the average time to death among nonsurvivors. Note that the locus on chromosome 15 achieved the 5% genome-wide significance level only with the nonparametric interval-mapping method.

SIMULATIONS

To better understand the relative performance of these approaches for QTL mapping in the case of a spike in the phenotype distribution, we performed a small simulation study. We first estimated the 95% genome-wide LOD threshold for each method, in the case of 250 intercross individuals with 25% having the null phenotype and an autosomal genome modeled after the genetic map for the mouse described in Rowe et al. (1994), consisting of 19 autosomes with total length 1300 cM. Genetic markers were equally spaced on each chromosome, with a marker spacing of 10–12 cM. (The intermarker spacing was slightly different for each chromosome, so that the chromosomal lengths could match those in the genetic maps of Rowe et al. 1994.) A random 10% of the marker genotype data was missing. We simulated a phenotype that was independent of the marker data. Each individual had probability 25% of having a null phenotype; otherwise their phenotype was drawn from a normal distribution with mean 10 and SD 1.

For each of 10,000 replicates, we simulated such data under the null hypothesis of no QTL, applied each of the four methods, and recorded the maximum LOD score, genome-wide, for each method. The 95th percentiles of the maximum LOD score, for the four methods, binary, QT, two-part, and NP, were 3.55, 3.53, 4.64, and 3.41, respectively. Note that the binary, QT, and NP methods have similar LOD thresholds. The LOD threshold for the two-part model is much higher, due to the fact that the corresponding statistical test concerns four free parameters, rather than two.

We also considered a fifth approach, in which one takes the maximum of the LOD scores from the binary and conditional quantitative trait analyses. For this approach, we used a Bonferroni correction and declared significant linkage if the LOD scores for either the binary trait analysis or the conditional quantitative trait analysis exceeded the corresponding 97.5% genome-wide LOD thresholds, which were estimated to be 3.88 and 3.86, respectively.

To investigate the power and precision of each of these methods, we simulated data under the two-part model described above, with a single QTL located between two markers near the center of chromosome 1 (of length 103 cM). The QTL was taken to have multiplicative effect \( \phi_\pi \) on the probabilities \( \pi \), and additive effect \( \Delta_\mu \) on the conditional means \( \mu \). The probabilities, \( \pi \), were chosen so that \( \pi_1 = \phi_\pi \pi_1 \) and \( \pi_2 = \phi_\pi \pi_2 \), and so
that the overall proportion of individuals with positive phenotypes was \( \pi_1/4 + \pi_2/2 + \pi_3/4 = 75\% \). The means were chosen so that \( \mu_1 = \mu_2 - \Delta_\mu \) and \( \mu_3 = \mu_2 + \Delta_\mu \), with \( \mu_2 = 10 \). The residual SD was \( \sigma = 1 \). We considered the values \( \phi_\pi = 1, 1.5, \) and \( 2 \) and \( \Delta_\mu = 0, 0.4, \) and \( 0.6 \). (Note that \( \phi_\pi = 1 \) and \( \Delta_\mu = 0 \) correspond to no QTL effect.)

We performed 4000 simulations of 250 intercross individuals, for all pairs of effects \( (\phi_\pi, \Delta_\mu) \), except for the case \( \phi_\pi = 1, \Delta_\mu = 0 \). The latter corresponds to the null hypothesis of no QTL; simulations for this case were used to estimate the LOD thresholds (see above). In each case, we applied the four methods to the simulated data on chromosome 1 (containing the QTL), calculated the maximum LOD score on that chromosome, and finally calculated the power of each test, as the proportion of the simulation replicates for which the maximum LOD score exceeded the corresponding 95% genome-wide LOD threshold. The power of the fifth procedure, taking the maximum of the binary and conditional quantitative trait LOD scores, was estimated as the proportion of the 4000 replicates in which either the binary or the conditional quantitative trait LOD score exceeded its corresponding 97.5% genome-wide LOD threshold.

The estimated power of the procedures appears in Figure 3. In Figure 3, A and D, the QTL had effect only on the probabilities, \( \pi_\pi \). In these cases, the conditional analysis of the quantitative trait had no power, and the analysis of the binary trait had the greatest power. The two-part model was somewhat inferior to the binary trait analysis, but had greater power than the nonparametric method. Use of the maximum of the binary and conditional quantitative trait LOD scores (with correction for the use of two tests) had somewhat greater power than the two-part model.

In Figure 3, G and H, the QTL had effect only on the conditional means, \( \mu_\pi \). In these cases, analysis of the binary trait had no power, and the conditional analysis of the quantitative trait had the greatest power. The results for the other methods were similar to the results in Figure 3, A and D: the two-part model was superior to the nonparametric method, but inferior to either the conditional quantitative trait analysis on its own or the
maximum of the binary and conditional quantitative trait analyses.

In Figure 3, B, C, E, and F, the QTL had effect on both the probabilities, $\pi$, and the conditional means, $\mu$. In these cases, the nonparametric method was best, although the use of the two-part model was competitive; both of these approaches showed considerable gains over either of the two separate analyses and over the maximum of the two separate analyses.

Figure 4 contains the results on the precision of QTL localization for the four basic methods. For each method and for each setting of the parameter values ($\phi, \Delta\mu$), the root-mean-square (RMS) of the error in the estimated QTL location, among simulation replicates in which there was significant evidence for the presence of a QTL (i.e., in which the maximum LOD score exceeded the corresponding 95% genome-wide LOD threshold), was calculated. Results for the conditional quantitative trait analysis (QT) for Figure 4, A and D, and for the binary trait analysis for Figure 4, G and H, are not shown, since these methods have no power to detect a QTL with the corresponding parameter settings. The results in Figure 4 mirror those in Figure 3. The methods with the highest power have the greatest precision of QTL localization (i.e., the smallest RMS error), while those with the lowest power have the lowest precision.

In summary, if a QTL has an effect only on the probabilities, $\pi$, or the conditional means, $\mu$, greatest power to detect the QTL is obtained with the separate analysis of that aspect of the data. If a QTL has an effect on both the probabilities, $\pi$, and the conditional means, $\mu$, the nonparametric method performed best. In all cases, analysis under the two-part model (with which the data were simulated) was second place, in terms of power. Note that further simulations, with 100 rather than 250 intercross individuals and with the proportion of individuals with the null phenotype taken to be 15 or 35% rather than 25%, gave qualitatively similar results (data not shown).

**DISCUSSION**

We have considered the problem of QTL mapping in the case of a spike in the phenotype distribution, a common departure from the usual normality assumption in standard interval mapping. Standard interval mapping works reasonably well when the spike is not too far from the rest of the phenotype distribution and contains only a small proportion of the individuals.
When the spike is well separated and contains an appreciable proportion of the data, maximum-likelihood estimation under a normal mixture model has a tendency to produce spurious LOD score peaks in regions of low genotype information (e.g., widely spaced markers).

We developed a parametric, two-part model for QTL mapping in this situation and have described an extension of the Kruskal-Wallis test statistic for nonparametric interval mapping in the case of an intercross. These approaches serve to combine the analysis of the binary trait with the conditional analysis of the quantitative trait among individuals with positive phenotype.

The interpretation of the results of analysis with the two-part model may deserve further explanation. A QTL identified through the two-part model may influence the probability of having a nonnull phenotype or the average phenotype among individuals with positive phenotype values or both. Inspection of the estimated QTL effects (the \( \hat{\pi}_j \) and \( \hat{\mu}_j \)) or of the results of the separate binary and conditional quantitative trait analyses should assist in discriminating between these cases.

In our simulation results, most interesting was the comparison among the two-part model, the nonparametric method, and the maximum of the binary and conditional quantitative trait analyses. In the case that QTL have an effect on both the parameters \( \pi_j \) (the probability that an individual with QTL genotype \( j \) will have a positive phenotype) and \( \mu_j \) (the conditional mean phenotype, among individuals with positive phenotype and QTL genotype \( j \)), the nonparametric approach was seen to have greater power than analysis under the two-part model; this is largely due to the fact that the genome-wide LOD threshold is considerably larger for the latter method. In the case that QTL have an effect on only the \( \pi_j \) or only the \( \mu_j \), the maximum of the separate analyses will have greatest power, and the nonparametric method will have the least power. Thus, analysis under the two-part model is always second best. On the other hand, the overall average power, across the eight parameter settings considered herein, was greatest for the two-part model. Further, the parametric, two-part model may be more useful in consideration of multiple-QTL models.

Thus, while nonparametric interval mapping is a valuable general method, analysis under the two-part model may be preferred for the situation considered here. The extensions of the two-part model for use with multiple QTL (for example, by combining a logistic model for the probabilities with a linear model for the conditional means) deserve exploration.

The methods described in this article have been implemented in the QTL mapping software, R/qtl (http://www.biostat.jhsph.edu/~kbroman/qtl), an add-on package for the general statistical software, R (IHAKA and GENTLEMAN 1996).

The author thanks Victor Boyartchuk and William Dietrich for providing the Listeria data. This work was supported in part by a Faculty Innovation Fund grant from the Johns Hopkins Bloomberg School of Public Health.

LITERATURE CITED


