

Estimating Polygenic Effects Using Markers of the Entire Genome

Shizhong Xu¹

Department of Botany and Plant Sciences, University of California, Riverside, California 92521

Manuscript received June 25, 2002

Accepted for publication November 13, 2002

ABSTRACT

Molecular markers have been used to map quantitative trait loci. However, they are rarely used to evaluate effects of chromosome segments of the entire genome. The original interval-mapping approach and various modified versions of it may have limited use in evaluating the genetic effects of the entire genome because they require evaluation of multiple models and model selection. Here we present a Bayesian regression method to simultaneously estimate genetic effects associated with markers of the entire genome. With the Bayesian method, we were able to handle situations in which the number of effects is even larger than the number of observations. The key to the success is that we allow each marker effect to have its own variance parameter, which in turn has its own prior distribution so that the variance can be estimated from the data. Under this hierarchical model, we were able to handle a large number of markers and most of the markers may have negligible effects. As a result, it is possible to evaluate the distribution of the marker effects. Using data from the North American Barley Genome Mapping Project in double-haploid barley, we found that the distribution of gene effects follows closely an L-shaped Gamma distribution, which is in contrast to the bell-shaped Gamma distribution when the gene effects were estimated from interval mapping. In addition, we show that the Bayesian method serves as an alternative or even better QTL mapping method because it produces clearer signals for QTL. Similar results were found from simulated data sets of F_2 and backcross (BC) families.

THE genetic variation of a quantitative trait is controlled by the segregation of multiple genes. In classical quantitative genetics, the overall genetic variance is described by the infinitesimal model, which assumes that the number of loci is infinitely large, each with an infinitely small effect. The genetic variances of individual loci are so small that they cannot be investigated separately, but collectively via phenotypic resemblance between relatives (LYNCH and WALSH 1998). It has been hypothesized that the genetic variance of most quantitative traits is actually controlled by a few loci with large effects and a large number of loci with small effects. Under this hypothesis, the distribution of the gene effect across loci may be described by a negative exponential distribution (OTTO and JONES 2000). The effects of the major genes can be studied via segregation analysis. The numerous genes with small effects, however, still cannot be investigated individually. As a result, evaluation of the hypothesis of negative exponential distribution of gene effect appears to be impossible.

With the advent of new molecular technology, saturated markers are being generated along the genome. Investigators are now able to investigate not only the effects of the major genes but also their locations in the genome. This is called quantitative trait loci (QTL) mapping (LANDER and BOTSTEIN 1989). However, QTL mapping involves multiple tests for individual loci. Only

significant loci are reported. As GORING *et al.* (2001 and references therein) stated that the reported QTL are almost always biased upward, they are not of much use for evaluating the distribution of the gene effect across loci. OTTO and JONES (2000) recently incorporated statistical test information into the study of QTL distribution, using a truncated negative exponential distribution. Their method actually depends on results of interval mapping of QTL.

Interval mapping requires multiple tests under multiple models. The test statistic becomes a function of the genome location and forms a test statistic profile after the entire genome has been searched. Permutation tests (CHURCHILL and DOERGE 1994) or other means of multiple test adjustment (PIEPHO 2001) are required to control the genome-wide type I error rate at a desired level. Upon completion of the genome search, the QTL effects need to be compiled and the total variance explained by the detected QTL needs to be calculated. However, QTL effects are estimated from different models. As a result, some inconsistency may often occur, such as the total variance explained by the QTL is too high. In addition, multiple estimates of the residual variance are generated and choosing the proper one for calculating the total phenotypic variance has become a problem. Models that include multiple QTL have been developed (SILLANPAA and ARJAS 1998; KAO *et al.* 1999). With these models, the problems of multiple tests and variance evaluation have been eliminated, but a new problem has been introduced with regard to model

¹Author e-mail: xu@genetics.ucr.edu

selection. A few issues need close attention in model selection. The criteria of deleting and inserting a QTL may be arbitrary. The sampling space of possible models (with different combinations of presence and absence of each putative QTL) may be too large to be fully explored. In addition, model selection will also cause biased estimates of gene effects if a single model is selected as the final model, although the biases can be reduced using the Bayesian method where several models are considered (SATAGOPAN *et al.* 1996). These problems have not been fully resolved. Any problems occurring in interval mapping will devalue the significance of the work by OTTO AND JONES (2000).

In this study, we propose a method for simultaneously evaluating marker effects of the entire genome. By marker effect, we mean the QTL effects associated with markers. If the marker density is relatively high, most of the QTL effects will be picked up by the markers and the results may be used to evaluate the distribution of gene effect across the genome. Hereon, we use the words QTL effect and marker effect interchangeably. Two problems are associated with simultaneous evaluation. One is how to handle the large number of markers in a single model. The other is how to deal with the markers with close-to-zero effects. We handle these problems by using a Bayesian method under the random regression coefficient model. In the Bayesian framework, each gene effect is assigned a normal prior with mean zero and a unique variance. The effect-specific prior variance is further assigned a vague prior so that the variance can be estimated from the data. This approach is analogous to the Bayesian method of MEUWISSEN *et al.* (2001) for BLUP prediction of gene effects in outbred populations.

METHODS

Linear model: Let y_i for $i = 1, \dots, n$ be the phenotypic value of the i th individual in a mapping population with only two segregating genotypes, *e.g.*, a backcross (BC) or a double-haploid (DH) population. The linear model for y_i is

$$y_i = b_0 + \sum_j^p x_{ij} b_j + e_i, \quad (1)$$

where b_0 is the population mean, p is the total number of markers in the entire genome, x_{ij} is a dummy variable indicating the genotype of the j th marker for individual i , b_j is the QTL effect associated with marker j , and e_i is the residual error with a $N(0, \sigma_0^2)$ distribution. For a DH population, an individual can take only one of the two genotypes, A_1A_1 and A_2A_2 , at any locus. The dummy variable is defined as $x_{ij} = 1$ for A_1A_1 and $x_{ij} = -1$ for A_2A_2 . Define the genetic effects associated with A_1A_1 and A_2A_2 by G_{11} and G_{22} , respectively, and then the regression coefficient is $b_j = G_{11} - G_{22}$. This model is the multiple regression model of ZENG (1993). The partial regression

coefficient b_j is the effect of marker j associated with the trait. It will absorb partly the effects of all QTL located between markers $j - 1$ and $j + 1$, as shown by ZENG (1993).

For an F_2 population, a dominance effect is associated to each marker locus. The linear model becomes

$$y_i = b_0 + \sum_j^p x_{ij} b_j + \sum_j^p w_{ij} d_j + e_i, \quad (2)$$

where x_{ij} and w_{ij} are defined as $x_{ij} = \sqrt{2}$ and $w_{ij} = -1$ for genotype A_1A_1 , $x_{ij} = 0$ and $w_{ij} = 1$ for A_1A_2 , and $x_{ij} = -\sqrt{2}$ and $w_{ij} = -1$ for A_2A_2 . Let G_{11} , G_{12} , and G_{22} be the genotypic values for the three genotypes. The regression coefficients are defined as $b_j = G_{11} - G_{22}$ for the additive effect and $d_j = 2G_{12} - G_{11} - G_{22}$ for the dominance effect. Note that x and w coded this way are independent and each has a zero expectation and a unity variance.

With a high marker density, most of the marker intervals will contain no QTL. Therefore, most of the regression coefficient will have a theoretical value of zero. In addition, the dummy variables will be highly correlated across loci, leading to a high degree of multicollinearity. When the number of markers exceeds the number of individuals, the ordinary least-squares approach will have no unique solution. Therefore, we must utilize a method that can handle the problem of multicollinearity. We show that the Bayesian regression method is the ideal solution for this problem.

Bayesian estimation: The Bayesian estimation is described only in the context of DH populations because, with a minor modification, the method can be applied to F_2 populations as well. Our Bayesian model differs from the usual regression model in that each b_j is assumed to be sampled from a normal distribution with mean zero and variance σ_j^2 .

In the Bayesian framework, we treat everything as a random variable, including the parameters. Each random variable has a distribution. We classify variables into observables and unobservables. The observables include $\mathbf{y} = \{y_i\}$ for $i = 1, \dots, n$ and marker information. The unobservables include $\mathbf{b} = \{b_j\}$ and $\mathbf{v} = \{\sigma_j^2\}$ for $j = 0, \dots, p$. The distribution of the observables is a function of the unobservables and is called the likelihood function. The distribution of the unobservables is called the prior distribution. The purpose of Bayesian analysis is to infer the conditional distribution of the parameters given the observed data, called the posterior distribution. Bayesian analysis implemented via the Markov chain Monte Carlo (MCMC) does not need an explicit form of the posterior distribution; rather, it draws a sample of the unobservables from the joint posterior distribution. From the joint posterior sample, one can easily obtain the desired Bayesian estimates, such as the posterior means and posterior variances.

In this study, we choose the following prior distributions, $p(b_0) \propto 1$, $p(\sigma_0^2) \propto 1/\sigma_0^2$, $p(b_j) = N(0, \sigma_j^2)$, and

$p(\sigma_j^2) \propto 1/\sigma_j^2$ for $j = 1, \dots, p$. The joint prior of the unobservables $p(\mathbf{b}, \mathbf{v})$ takes the product of the priors of individual parameters. The likelihood is

$$p(\mathbf{y}|\mathbf{b}, \mathbf{v}) = \prod_{i=1}^n p(y_i|\mathbf{b}, \sigma_0^2) \propto (\sigma_0^2)^{-n/2} \times \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (y_i - b_0 - \sum_{j=1}^p x_{ij}b_j)^2 \right\}. \quad (3)$$

The joint posterior distribution has a form of

$$p(\mathbf{b}, \mathbf{v}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{b}, \mathbf{v})p(\mathbf{b}, \mathbf{v}). \quad (4)$$

Genotypes of missing markers were generated randomly in each iteration on the basis of the probability inferred jointly from the nearest nonmissing flanking markers and the phenotype. The probability from the markers is treated as the prior probability. After incorporation of the marker (QTL) effects through the phenotype, the probability becomes the posterior probability, which is used to generate the missing marker genotype. In HD, BC, and F₂ populations, a codominant marker is either fully informative or noninformative. Therefore, using the nearest nonmissing markers is equivalent to using the multipoint method. For dominant markers, a marker in an F₂ population can be partially informative and the multipoint method (JIANG and ZENG 1997) should be used.

In the MCMC-implemented Bayesian analysis, we sample the unobservables from the above joint posterior distribution. The sampling is performed in the following sequences.

Step 1. Initialization: We first initialize all unobservables, denoted by

$$\mathbf{Q}^{(0)} = [b_0^{(0)}, \dots, b_p^{(0)}, \sigma_0^{2(0)}, \dots, \sigma_p^{2(0)}].$$

The location parameters \mathbf{b} are initialized with zero value and the scale parameters \mathbf{v} are initialized with a positive number. The unobservables also include missing marker genotypes, which are initialized by random sampling on the basis of the posterior probability, conditional on the initial values of the model effects and variances.

Step 2. Updating b_0 : The conditional posterior distribution of b_0 is Normal with mean $\bar{b}_0 = (1/n) \sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}b_j^{(0)})$ and variance $s_0^2 = (1/n) \sigma_0^{2(0)}$, from which a new b_0 is sampled. The sampled b_0 is denoted by $b_0^{(1)}$, which will replace $b_0^{(0)}$ in all subsequent sampling processes.

Step 3. Update b_j : The conditional posterior distribution for b_j is Normal with mean

$$\bar{b}_j = \left(\sum_{i=1}^n x_{ij}^2 + \sigma_0^{2(0)}/\sigma_j^{2(0)} \right)^{-1} \sum_{i=1}^n x_{ij} \left(y_i - b_0^{(0)} - \sum_{k \neq j} x_{ik} b_k^{(0)} \right) \quad (5)$$

and variance

$$s_j^2 = \left(\sum_{i=1}^n x_{ij}^2 + \sigma_0^{2(0)}/\sigma_j^{2(0)} \right)^{-1} \sigma_0^{2(0)}, \quad (6)$$

which are used to sample b_j . The newly sampled b_j is denoted by $b_j^{(1)}$ and will replace $b_j^{(0)}$ in all subsequent sampling processes.

Step 4. Update σ_0^2 : The residual variance is sampled from a scaled inverted chi-square distribution; that is, $\sigma_0^{2(1)} = (1/\chi_n^2) \sum_{i=1}^n (y_i - b_0^{(0)} - \sum_{j=1}^p x_{ij}b_j^{(0)})^2$, where χ_n^2 is a random number sampled from a chi-square distribution with n d.f. The variances are immediately updated: $\sigma_0^{2(0)} = \sigma_0^{2(1)}$.

Step 5. Update σ_j^2 for $j = 1, \dots, p$: We sample σ_j^2 from a scaled inverted chi-square distribution; that is, $\sigma_j^{2(1)} = b_j^{2(0)}/\chi_1^2$, where χ_1^2 is a random number sampled from a chi-square distribution with 1 d.f.

Step 6. Update missing marker genotypes.

Step 7. Repeat 2–6: At this moment, we have completed one sweep of the MCMC and are ready to continue our sampling for the next round. When the chain converges to the stationary distribution, the sampled parameters actually follow the joint posterior distribution. When the sample of a single parameter is viewed, this univariate sample is actually the marginal posterior sample for this parameter.

RESULTS

DH mapping in barley: Data from the North American Barley Genome Mapping Project (TINKER *et al.* 1996) were used for analysis. Seven traits were investigated in the project: yield, heading, maturity, height, lodging, kernel weight, and test weight. The DH population contained 145 lines ($n = 145$); each was grown in a range of environments. A total of 127 mapped markers ($p = 127$) covering ~ 1500 cM of the genome along seven linkage groups were used in the analysis. All seven traits were reanalyzed in this study, but only the result of kernel weight was represented here. Note that the data sets were updated after they were first published in 1996, but the difference between the updated and the original data was minor so that the results are still comparable between the current analysis and the analysis by the original authors.

The average phenotypic value across the environments was calculated for each line and these average values were treated as the original phenotypic records (y_i) for analysis. Although results of interval mapping showed significant QTL-by-environmental interaction, most QTL showed effects in the same direction across environments. Therefore, we report only the analysis of average values across environments. In the QTL mapping program, the phenotypic values were further standardized using $y_i = (y_i - \bar{y})/s_y$, where \bar{y} is the mean and s_y is the standard deviation of y . The standardized record was subject to Bayesian analysis. With the standardization, users can always choose the default initial values provided by the program.

The default initial values were set at $b_j = 0$ and $\sigma_j^2 = 0.5$ for $j = 0, \dots, p$. The length of the Markov chain

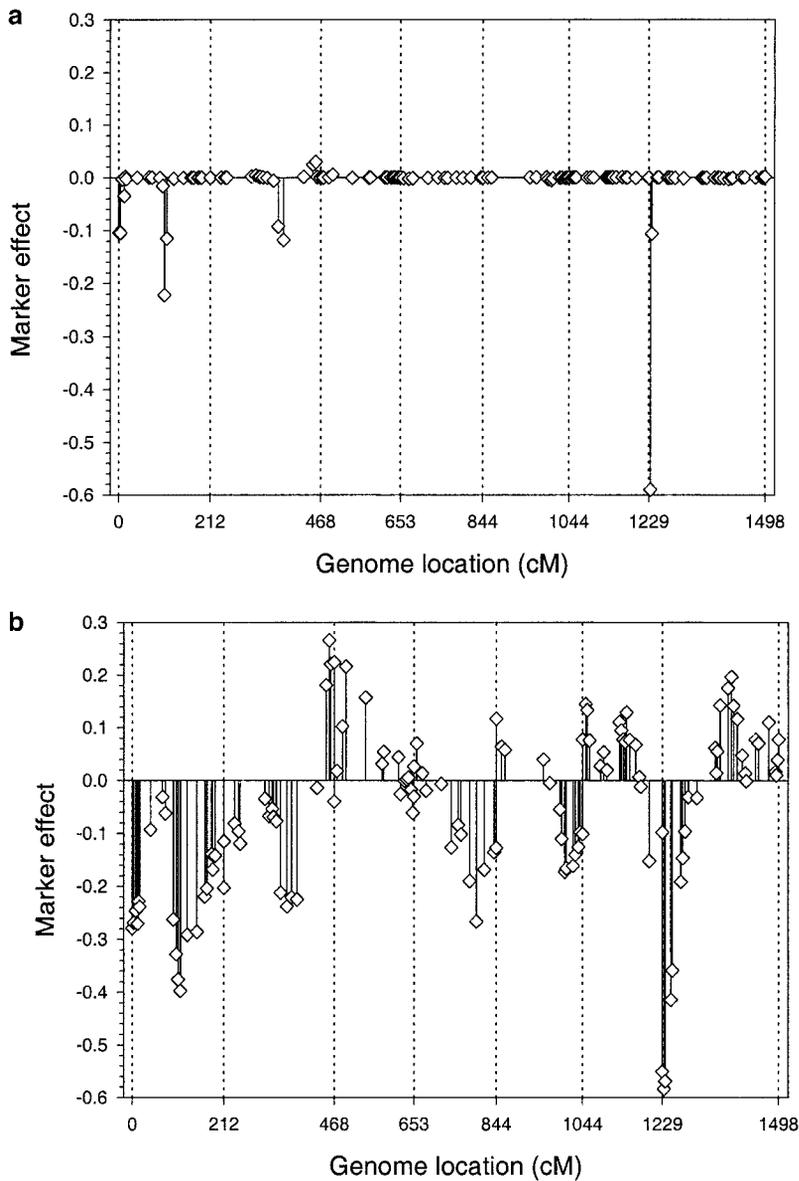


FIGURE 1.—Marker effects of kernel weight in barley plotted against marker locations along the genome. (a) Multiple-marker Bayesian analysis; (b) individual-marker regression analysis. The dotted vertical reference lines separate the seven linkage groups.

contained 51,000 sweeps. The sampled parameter values from the first 1000 sweeps of the chain (burn-in period) were discarded from the analysis. From there on, the observations were saved for every 50 sweeps to reduce the series correlation. Therefore, the posterior sample contains 1000 observations for post-Bayesian analysis. The posterior means of the marker effects (as the Bayesian estimates) are reported.

For comparison, we also performed the single-marker analyses with the simple regression method for each marker. Since the marker density is quite high, results of single-marker analyses should be close to those of interval mapping. Figure 1 shows the plot of the marker effects against the genome location (cM) of the markers for kernel weight. Note that the seven linkage groups have been ligated into a single genome. The genome location of each marker takes the cumulative position measured from the left to the right. For example, the

first marker of the second chromosome occupies the same position (212 cM) as the last marker of the first chromosome. Figure 1a (multiple-marker Bayesian) clearly shows four candidate regions with evidence of QTL and these regions coincide with the peaks shown in Figure 1b (individual marker regression). The two regions with larger effects have been declared by TINKER *et al.* (1996) as significant. One striking result found in the multiple-marker Bayesian analysis is that the markers with large effects in the single-marker analysis maintain their large effects in the Bayesian analysis while the markers with small effects in regression have been shrunk in the Bayesian analysis. The signals of QTL are so much clearer in the Bayesian analysis than in the regression analysis. We also analyzed the remaining six traits and they all show similar patterns; *i.e.*, the major peaks in Bayesian analysis coincide with those of the single-marker analysis and the interval mapping re-

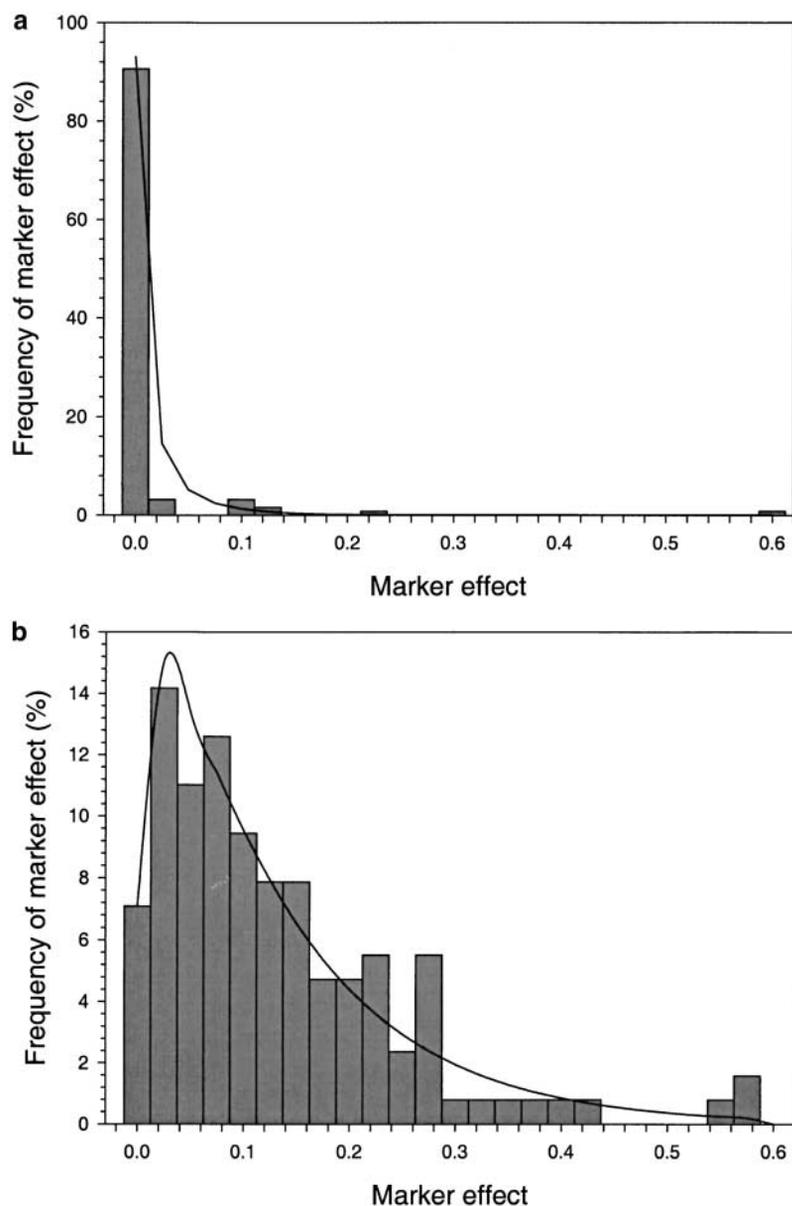


FIGURE 2.—Frequency distributions of marker effects of kernel weight in barley. (a) Multiple-marker Bayesian analysis; (b) individual-marker regression analysis.

ported by TINKER *et al.* (1996; data not shown). Therefore, the proposed Bayesian analysis including all markers in the genome can serve as an alternative (and even better) QTL-mapping method to the interval-mapping method.

The second striking feature of the Bayesian analysis is that most of the markers have an estimated effect close to zero, which follows the prediction of the oligogenic model. The single-marker regression analysis, however, produces spurious effects for many markers. Although it provides a good tool for QTL detection, it is simply not useful for evaluation of the genetic effects of the genome. Figure 2 shows the distribution of the absolute value of estimated gene effect along the genome for kernel weight. The estimates of multiple-marker analysis fit the Gamma distribution with the scale and shape parameters of $\alpha = 0.0579$ and $\beta = 0.2233$, respectively, while the estimates of the individual marker analysis fit

the Gamma distribution with $\alpha = 0.1145$ and $\beta = 1.1396$. The shapes of the two distributions are quite different. The multiple-marker analysis generated an L-shaped distribution because $\beta < 1$ and the individual-marker analysis generated a bell-shaped distribution because $\beta > 1$. The distributions of QTL effects estimated for the remaining six traits follow the same trends (see Table 1). Therefore, the Bayesian analysis is a viable tool for evaluating the polygenic effects of the entire genome.

The proportion of phenotypic variance explained by the markers is expressed as $\hat{h}^2 = 1 - \hat{\sigma}_0^2$ where $\hat{\sigma}_0^2$ is the Bayesian estimate of the residual variance. This formula is a special form of $\hat{h}^2 = (\hat{\sigma}_y^2 - \hat{\sigma}_0^2) / \hat{\sigma}_y^2$, where $\hat{\sigma}_y^2 = 1$ is the phenotypic variance (after the standardization). With the single-marker analysis, we cannot find a proper $\hat{\sigma}_0^2$ to use because each marker has its own $\hat{\sigma}_0^2$. If we took $\hat{h}^2 = \sum_{j=1}^p \hat{h}_j^2 = \sum_{j=1}^p (1 - \hat{\sigma}_{0j}^2)$, where $\hat{\sigma}_{0j}^2$ is the residual

TABLE 1

Parameters of Gamma distributions of QTL effects in barley

Trait	Multiple-marker analysis ^a		Individual-marker analysis	
	α	β	α	β
Kernel weight	0.0579	0.2233	0.1145	1.1396
Yield	0.0100	0.4194	0.0862	1.3148
Heading	0.0285	0.3132	0.0955	1.6229
Height	0.0181	0.3154	0.0862	1.3109
Maturity	0.0323	0.2928	0.0985	1.3743
Test weight	0.0184	0.3097	0.0854	1.3403
Lodging	0.0165	0.3034	0.1059	1.0591

^a Note that α is the scale parameter and β is the shape parameter. When $\beta < 1$ the curve is L-shaped, when $\beta = 1$ the curve is exponential, and when $\beta > 1$ the curve is bell shaped.

variance when the j th marker is fitted, we would soon end up with $\hat{h}^2 > 1$, which contradicts with the definition of h^2 . Therefore, we report only \hat{h}^2 from the Bayesian analysis (Table 2). During the sampling process, sometimes the residual variance can be >1 , which explains the -5% values of the posterior distribution of h^2 . Kernel weight has the highest polygenic variance (0.6484), followed by maturity (0.4549) and heading (0.4042). The remaining traits show lower polygenic variances (explained by markers). Overall, the polygenic variances explained by markers were smaller than those reported by TINKER *et al.* (1996) with the traditional (nonmarker) analysis. Test weight has a high h^2 in the traditional analysis ($h^2 = 0.61$) but a low h^2 explained by markers ($h^2 = 0.1665$). It is possible that some of the major genes may be located in the regions with large gaps of markers. For example, the largest gap (~ 70 cM long) occurs between markers aHor2 and NWG943 on chromosome 5. A gap with this size will certainly fail to pick up any QTL in between.

The two major peaks identified with the Bayesian analysis (Figure 1) for kernel weight coincide with the two QTL identified with the interval mapping. The two

TABLE 2

Proportion of phenotypic variance explained by markers

Trait	Mean	Standard deviation	5% ^a	95% ^a
Kernel weight	0.6484	0.0735	0.5189	0.7550
Yield	0.1343	0.1739	-0.1895	0.3956
Heading	0.4042	0.1103	0.2038	0.5726
Height	0.2131	0.1747	-0.0750	0.5049
Maturity	0.4549	0.1040	0.2614	0.6047
Test weight	0.1665	0.1910	-0.1679	0.4374
Lodging	0.2318	0.1268	0.0109	0.4210

^a The percentiles 5% and 95% stand for the 5th and 95th percentile values of the posterior distribution (90% credibility interval).

peaks with small effects, however, failed to be detected with the interval mapping. We performed several additional analyses to verify whether these subpeaks are true or due to stochastic error in MCMC. We found that these subpeaks occurred most of the times but failed to show up in a few replications (data not shown). Therefore, the QTL evidence of the two subpeaks is not strong. However, in Bayesian analysis, we do not claim insignificance of QTL. Instead, we report small posterior estimates for the two peaks. The two major QTL identified remained in the model for all replicates.

To explore the behavior of the QTL-effect profile under the null model, we reshuffled the phenotypic data so that the association between the markers and the phenotype would be artificially destroyed. We then performed Bayesian analysis on the reshuffled data. This is equivalent to the permutation analysis of CHURCHILL and DOERGE (1994). Among analyses of 10 reshuffled data sets, most of them showed a very flat profile (close to zero estimates for all markers). The 10 profiles were drawn in the same graph (Figure 3), which shows very little variation of the estimated marker effects among the reshuffled data. We now feel more confident that the major peaks identified are not likely due to spurious effects.

F₂ mapping with simulated data: To demonstrate that the proposed Bayesian method can handle data with the number of effects larger than the number of individuals, we simulated 301 markers in an F₂ population with 300 individuals. Each marker is associated with an additive and a dominance effect, and thus the model includes 602 effects. For convenience of programming, we arranged the 301 markers in a single large chromosome with 5 cM between consecutive markers. The total length of the hypothetical chromosome is 1500 cM. We simulated four QTL with their locations and sizes listed in Table 3. The true population mean and residual variance are $b_0 = 5.0$ and $\sigma_0^2 = 10$, respectively. The genetic variance due to each QTL is determined by $v_g = a^2 + d^2$, where a and d are the additive and dominance effects, respectively. The total genetic variance for the four QTL is ~ 18.0 (excluding the negligible covariance due to linkage). Therefore, the proportion of the total phenotypic variance explained by the four QTL is $H = 18.0 / (18.0 + 10.0) = 64.26\%$.

First, we made a simplification for the distribution of b_j . We assumed that $b_j \sim N(0, 1/\lambda)$ for $j = 1, \dots, p$, where λ is a constant positive number. This leads to the usual Bayesian regression analysis with a common variance for all b_j . It is also analogous to the ridge regression (HOERL and KENNARD 1970). When λ is chosen as a very small positive number, there is no unique solution for the model with this many effects. We then gradually increased λ until a unique solution is possible for the regression coefficients. We examined the estimated regression coefficient and plotted them against the chromosome location (Figure 4, a and b). Note that the ridge estimates of the effects vary widely around zero

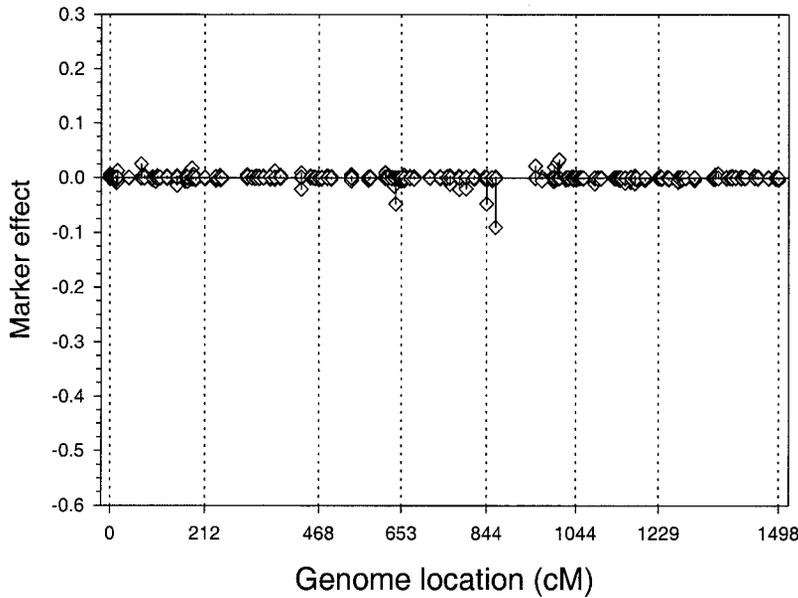


FIGURE 3.—Multiple-marker Bayesian analysis (null model). Marker effects plotted against marker locations along the genome of kernel weight in barley from 10 reshuffled data sets (permutation analysis).

with no clear signals at the QTL positions. Further increase of λ has reduced the variation of the estimates, but still there are no signals of QTL along the genome. Instead of choosing λ subjectively, we attempted to let the data speak for themselves, where we treated λ as a random variable sampled from its conditional posterior distribution, $\lambda = \chi_p^2 / \sum_{j=1}^p b_j^2$, where χ_p^2 is a sampled chi-square variable with p degrees of freedom. The result is almost identical to the situation of constant λ (data not shown). When we adopted the Bayesian approach developed in this study using b_j specific variances, the results are strikingly different (Figure 4, c and d). The signals of QTL become extremely clear at the true positions. The estimated effects of the identified QTL are very close to those of the true position simulated (Table 4). The QTL located at position 250 cM is weak (explaining $\sim 7\%$ of the phenotypic variance, 3.5% by additive effect and 3.5% by dominance). The estimated additive effect is half the size of the true value. The dominance effect is also reduced by half, but the other half is picked up by the next marker 5 cM away from the true position. This is expected because a small QTL should be hard to estimate. The estimated residual variance is 9.75, close to the true value of 10. The estimated proportion of variance explained by the five listed mark-

ers is $\hat{h}^2 = 17.08 / (17.08 + 9.75) = 0.6365$, almost hitting the true value of 0.6426.

BC mapping with simulated data: We also simulated data from a BC family to examine the behavior of the method in some interesting situations. First, we simulated four QTL with exactly the same setup as the F_2 simulation except that the first and last QTL have negative effects (QTL with effects in different directions). We set the genetic effects to 2.828, -1.414, 2.000, and -2.000 for the four QTL, respectively. This made the variance of each QTL identical to the variance listed in Table 3 for the F_2 design. For example, the first QTL variance is $8 = 2.828^2$ and the second is $2 = (-1.414)^2$, etc. The estimated marker effects plotted against the genome location are shown in Figure 5. The method works equally well as the situation where QTL act in the same direction.

We then simulated 11 QTL evenly placed in the single large chromosome of 1500 cM. The QTL are numbered from 1 to 11 with variances in descending order: 20, 10, 5, 2.5, 1.25, 0.625, 0.3125, 0.3125, 0.3125, 0.3125, and 0.3125. The total variance of the 11 QTL is ~ 40 (ignoring the covariance due to linkage). The phenotypic variance is $40 + 10 = 50$. Therefore, the first QTL explains 40% of the phenotypic variance and the second QTL explains 20% of the phenotypic variance and so on. The actual effects of the 11 QTL simply take the square root of the corresponding variances. The result of Bayesian analysis is depicted in Figure 6. The signals of the large QTL are quite clear until the variance is reduced to 0.625, under which the method failed to give any meaningful estimates. The smallest QTL that the method can pick up in this example explains 1.25% of the phenotypic variance.

Finally, we simulated four QTL with effects equal to 2.828, 1.414, 2.000, and 2.000, respectively. QTL nos. 1, 2, and 4 are located at positions 0, 250, and 750 cM,

TABLE 3

Locations and sizes of the four QTL used in the simulation

QTL	Position (cM)	Additive (a)	Dominance (d)	Variance (v_g)
1	0	2.0	2.0	8.0
2	250	1.0	1.0	2.0
3	500	2.0	0.0	4.0
4	750	0.0	2.0	4.0
	Σ			18.0

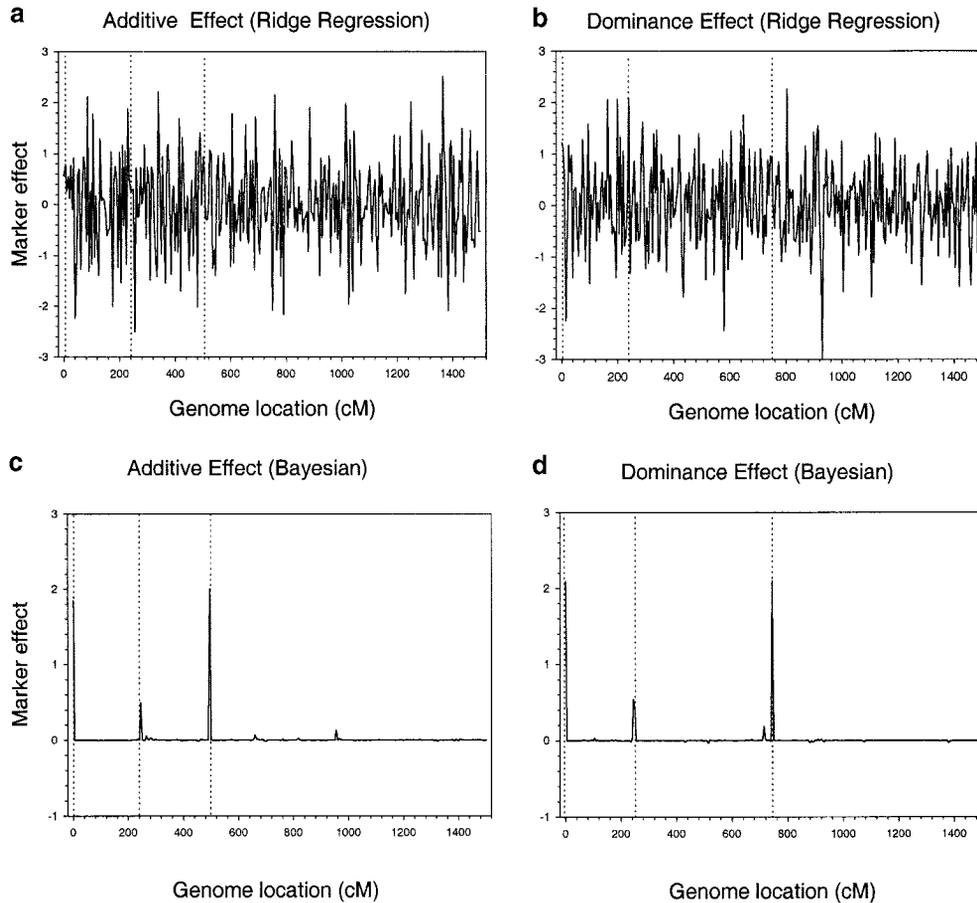


FIGURE 4.—Marker effects estimated from ridge regression and Bayesian analysis plotted against marker locations along a simulated large chromosome of 1500 cM. (a and b) The additive effect and dominance effects from ridge regression; (c and d) the additive and dominance effects from Bayesian analysis. The vertical reference lines indicate the positions of the markers with non-zero simulated QTL effects.

respectively, whereas the position of QTL no. 3 varies from 255 to 290 cM with a 5-cM increment. From this simulation experiment we can examine the ability of the method to separate closely linked QTL (nos. 2 and 3). Figure 7 shows the plots of the marker effects on the genome location when the two linked QTL (nos. 2 and 3) are (a) 5 cM, (b) 10 cM, (c) 15 cM, (d) 20 cM, (e) 25 cM, (f) 30 cM, (g) 35 cM, and (h) 40 cM apart. We can see that the method separates the two closely linked QTL very well when the distance between the two QTL is >5 cM. When the distance is 5 cM the two markers are adjacent with no intermediate markers to separate, and thus they are inseparable.

DISCUSSION

WHITTAKER *et al.* (2000) first applied ridge regression to marker-assisted selection and showed that ridge regression can substantially improve the selection efficiency. In their analysis, markers included in the model were selected on the basis of QTL mapping results and the number of markers was much smaller than the number of individuals. We demonstrated that ridge regression is not a viable choice for QTL mapping if the model includes markers of the entire genome. The reason is that ridge regression treats all effects equally across loci. Our prior knowledge is that most markers have negligible effects. We need a method to discriminate the effects

TABLE 4
Estimated QTL parameters from the simulated data

QTL	Position (cM)	a	a_L	a_U	d	d_L	d_U
1	0	1.8460	1.4901	2.1711	2.0976	1.7767	2.4373
2a	250	0.4954	0.0000	1.0128	0.5384	0.0000	1.2317
2b	255	0	0	0	0.3986	0	1.1795
3	500	2.0020	1.6694	2.2914	0	0	0
4	750	0	0	0	2.0956	1.7322	2.4854

Note that a and d stand for the additive and dominance effects, respectively; a_L and a_U stand for the 5th and 95th percentiles of the posterior sample for the additive effect; d_L and d_U are defined similarly for the dominance effect.

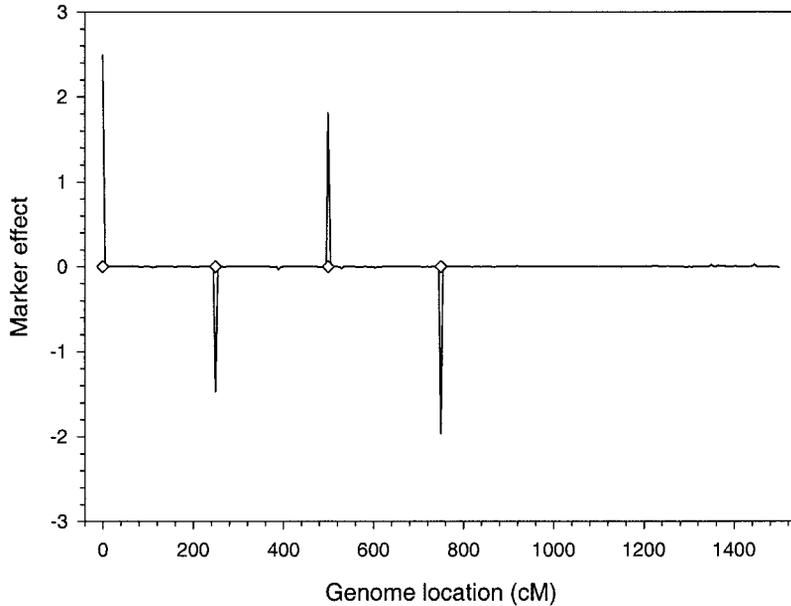


FIGURE 5.—Bayesian estimates of marker effects in the simulated BC family when the QTL have effects in different directions. The open diamonds (◇) indicate the true positions of the four simulated QTL.

across markers. The inverse of the variance actually serves as a coefficient of penalty. If a marker has a noticeable effect, it should not be penalized as severely as a marker with a negligible effect and thus should be given a large variance. An extremely small variance will cause a close to zero estimate of b_j . In fact, most of the markers will be given an extremely small σ_j^2 and thus their effects will be negligibly small. Updating the variance σ_j^2 for the j th marker is important because it depends on the sampled b_j from the previous round; *i.e.*, $\sigma_j^2 = b_j^2 / \chi_1^2$. If $b_j \rightarrow 0$, then $\sigma_j^2 \rightarrow 0$ in general. However, dividing b_j^2 by a chi-square variable allows σ_j^2 to have a chance to recover because χ_1^2 can be very small by chance.

As demonstrated in Figure 2, the marker effects fit a Gamma distribution nicely. However, the shape of the

Gamma distribution from multiple-marker analysis (L-shaped) is quite different from that of the individual marker analysis (bell-shaped). An L-shaped Gamma distribution is probably closer to reality. HAYES and GODDARD (2001) collected data from many QTL mapping experiments in pigs and dairy cattle and investigated the distribution of the QTL effects. They found that the distribution of QTL effects is more toward L-shaped distribution, especially in the cattle, although the distribution in pigs is slightly bell-shaped. MACKAY's (1996) experiments in *Drosophila* showed that many loci have small effects on abdominal and sternopleural bristle number, but few loci cause most of the genetic variation. EDWARDS *et al.* (1987) investigated the associations of markers with 82 traits in corn and discovered the L-shaped distribution of QTL effects. In addition to the

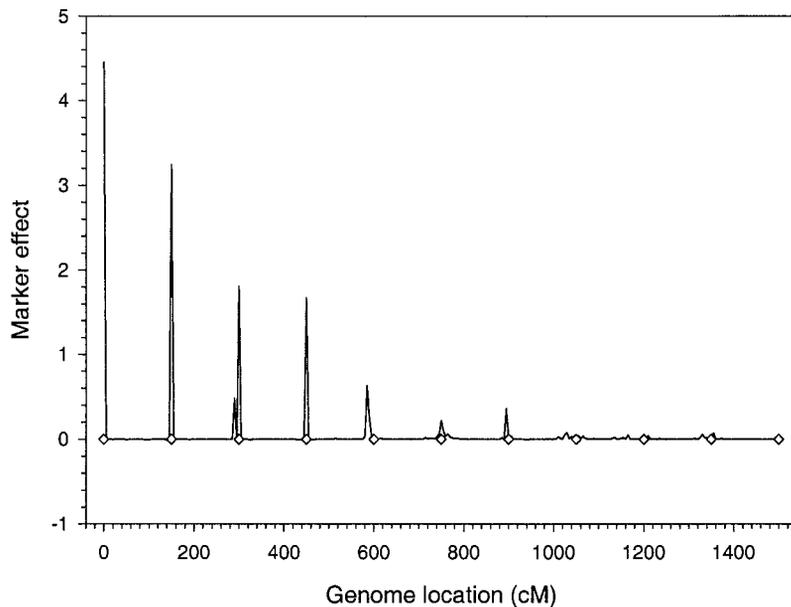


FIGURE 6.—Bayesian estimates of marker effects in the simulated BC family with 11 QTL of various sizes. The open diamonds (◇) indicate the true positions of the simulated QTL.

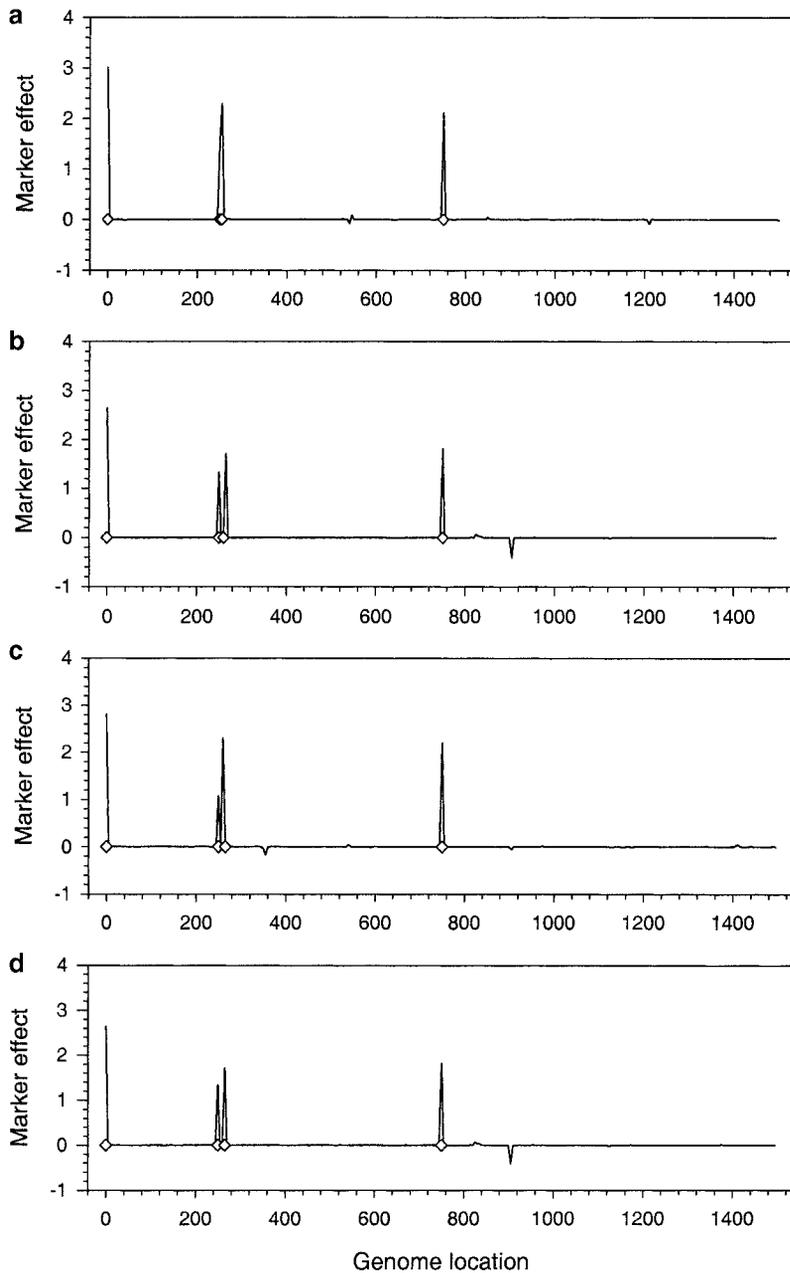


FIGURE 7.—Bayesian estimates of marker effects in the simulated BC family with four QTL, two of which (QTL nos. 2 and 3) are tightly linked from 5 cM (a) to 40 cM (h) with a 5-cM increment. The open diamonds (\diamond) indicate the true positions of the simulated QTL.

empirical evidence of L-shaped distribution, BOST *et al.* (1999, 2001) proved that the L-shaped distribution can be generated by the intrinsic property of metabolic pathways due to the summation property of control coefficients of the enzymes on the variation of the fluxes. Both MACKAY (2001) and BOST *et al.* (2001) accepted the fact that some experimental and statistical artifacts may cause the L-shaped distribution.

One caveat in fitting the Gamma distribution of marker effect needs to be clarified. We normally fit a model by using observations (true gene effects in this case), but here we used estimated gene effects to fit a model and completely ignored the estimation errors of the gene effects. When fitting the Gamma distribution using estimated gene effects, HAYES and GODDARD

(2001) took into account the estimation errors by assuming normal distributions for the estimated QTL effects. However, numerical integration was required to fit the Gamma model, and, therefore, for simplicity we did not consider the estimation errors when fitting the Gamma model. In fact, it is possible to choose a Gamma prior for each marker effect and directly estimate the scale and shape parameters in the Bayesian analysis. The problem is that Gamma is not a conjugate prior in QTL mapping and we would not be able to take advantage of the Gibbs sampling (GEMAN and GEMAN 1984) and would be forced to use the Metropolis-Hastings algorithm (METROPOLIS *et al.* 1953; HASTINGS 1970), which is less efficient than the Gibbs sampling. Although it will be interesting to compare the results of fitting Nor-

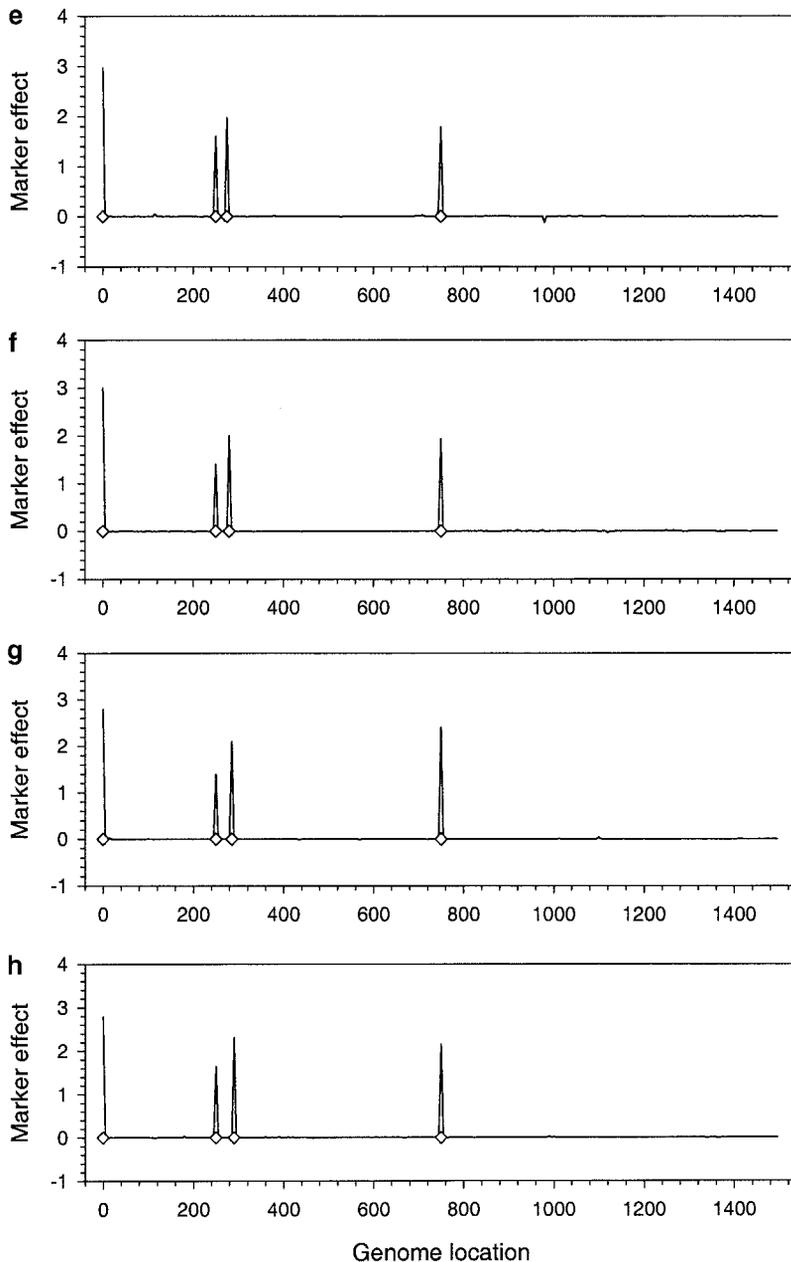


FIGURE 7.—Continued.

mal and Gamma priors for the gene effects, we decided to leave this as an option for future studies.

The original idea of our work was stimulated by the work of MEUWISSEN *et al.* (2001), who proposed the Bayesian method to simultaneously evaluate effects of a large number of markers in a genome. They took a mixed-model approach by treating each QTL allele as a random effect sampled from a normal distribution. Amazingly, they were able to estimate effects of 1000 markers involving 50,000 allelic effects with only 2200 phenotypic records. The authors investigated this problem also from the marker-assisted selection point of view under the framework of association study at the population level. Their work is different from ours in that (i) theirs was an association study at the population level whereas ours is a linkage-mapping study at a family

level; (ii) in their study, each locus had >50 different alleles and these alleles had a common variance whereas in our QTL mapping study, each locus has only two alleles and the allelic difference has a unique variance; and (iii) they analyzed simulated data whereas we analyzed both the simulated data and the data collected from the fields. Note that the QTL alleles in their study were actually defined as marker haplotypes, the combinations of alleles of several consecutive markers. In line-crossing experiments, as investigated in this study, we can deal directly with marker alleles instead of consecutive marker haplotypes. Therefore, the work in this study is more related to QTL mapping than to association study.

Traditional methods of QTL mapping include single-marker analysis and interval mapping. In single-marker

analysis, one marker is analyzed at a time and the model effect is the effect of the marker in question. Interval mapping allows the effect of an arbitrary position between two flanking markers to be estimated. When the marker density is sufficiently high, the single-marker analysis reaches its asymptotic limit—the interval mapping. Recently, a multiple-interval mapping was proposed in which a single linear model may contain all possible QTL (KAO *et al.* 1999). Since there has been no convenient way to handle too many intervals simultaneously, intervals must be selectively included in the model, generating a model selection problem. Let us call the multiple-interval mapping with model selection the selective multiple-interval mapping. The true multiple-interval mapping should include all intervals defined by markers. The multiple-marker analysis proposed in this study will reach its asymptotic limit—the true multiple-interval mapping (including all intervals) when the marker density is sufficiently high. If the marker density is low, interval mapping and multiple-interval mapping may offer some advantage over marker analysis if a QTL is located in the middle of a large interval because they can point to one side rather than to two sides of the marker. If a QTL is located right at a marker, interval mapping offers no advantage over single-marker analysis. No one will try to clone a QTL identified by the interval mapping that is located between two distant markers. One should always try fine mapping using saturated markers in an expanded population to further localize the QTL before considering cloning. From that point of view, multiple-marker analysis and multiple-interval mapping will provide the same amount of information.

The thorniest problem in multiple-QTL analysis comes from model selection, which has recently become the focus of QTL-mapping studies (KAO *et al.* 1999; BALL 2001; GORING *et al.* 2001; SEN and CHURCHILL 2001; YI and XU 2001; BROWMAN and SPEED 2002; SILLANPAA and CORANDER 2002). The all-marker analysis developed here is a model-selection-free approach. It has prevented all the problems of model selection, but the price is the inclusion of all model effects, even effects with zero values. Because of the high dimensionality of the model, it violates the usual rule of parsimony in model fitting. Fortunately, we were able to penalize the small effects and give them negligible weights so that their inclusion should have negligible effects on the analysis. One of the nice properties of Bayesian analysis is its ability to handle a large number of model effects. The all-marker analysis developed in this study has fully taken advantage of this property and results of data analysis have clearly verified this notion.

When two markers are closely linked, the effect of one marker is usually split between the two markers, as demonstrated in our simulation studies. In this case, we may suppress one of the two markers. The ability to

handle closed markers depends on the sample size and the type of population. Large sample sizes will allow separation of more closed markers. Populations carrying historical recombination events, *e.g.*, recombinant inbred lines, can also handle more closed markers.

The next step of the multiple-marker analysis is to estimate epistatic effects between pairwise markers. With the epistatic model, the number of effects increases rapidly as the number of markers increases. It is not clear at this moment whether a model with many hundred times more effects than the number of observations still works with the proposed method. We are confident that additive and dominance effects can be handled easily for data generated in most QTL-mapping projects.

Dr. Chenwu Xu (postdoctorate) helped download the data from the internet and helped perform some preliminary data manipulation to meet the required format of the C++ program. Ms Hui Wang (Ph.D. student) performed part of the simulation experiments. Both are greatly appreciated for their contributions to the project. This work was supported by the National Institutes of Health (grant R01-GM55321) and the U.S. Department of Agriculture National Research Initiative Competitive Grants Program (00-35300-9245).

LITERATURE CITED

- BALL, R. D., 2001 Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics* **159**: 1351–1364.
- BOST, B., C. DILLMANN and D. DE VIENNE, 1999 Fluxes and metabolic pools as model traits for quantitative genetics. I. The L-shaped distribution of gene effects. *Genetics* **153**: 2001–2012.
- BOST, B., D. DE VIENNE, F. HOSPITAL, L. MOREAU and C. DILLMANN, 2001 Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. *Genetics* **157**: 1773–1787.
- BROWMAN, K. W., and T. P. SPEED, 2002 A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Stat. Soc. B* **64**: 1–16.
- CHURCHILL, G. A., and R. W. DOERGE, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963–971.
- EDWARDS, M. D., C. W. STUBER and J. F. WENDEL, 1987 Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* **116**: 113–125.
- GEMAN, S., and D. GEMAN, 1984 Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Machine Intell.* **6**: 721–741.
- GORING, H. H. H., J. D. TERWILLIGER and J. BLANGERO, 2001 Large upward bias in estimation of locus-specific effects from genome-wide scans. *Am. J. Hum. Genet.* **69**: 1357–1369.
- HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.
- HAYES, B., and M. E. GODDARD, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33**: 209–229.
- HOERL, A. E., and R. W. KENNARD, 1970 Ridge regression: application to nonorthogonal problems. *Technometrics* **12**: 68–82.
- JIANG, C., and Z-B. ZENG, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.
- 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.
- LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

- MACKAY, T. F., 1996 The nature of quantitative genetic variation revisited: lesson from *Drosophila* bristles. *Bioessays* **18**: 113–121.
- MACKAY, T. F., 2001 The genetic architecture of quantitative traits. *Annu. Rev. Genet.* **35**: 303–339.
- METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER and E. TELLER, 1953 Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**: 1087–1091.
- MEUWISSEN, T. H. E., B. J. HAYES and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.
- OTTO, S. P., and C. D. JONES, 2000 Detecting the undetected: estimating the total number of loci underlying a quantitative trait. *Genetics* **156**: 2093–2107.
- PIEPHO, H. P., 2001 A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157**: 425–432.
- SATAGOPAN, J. M., B. S. YANDELL, M. A. NEWTON and T. C. OSBORN, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**: 805–816.
- SEN, S., and G. A. CHURCHILL, 2001 A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.
- SILLANPAA, M. J., and E. ARJAS, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.
- SILLANPAA, M. J., and J. CORANDER, 2002 Model choice in gene mapping: what and why. *Trends Genet.* **159**: 371–387.
- TINKER, N. A., D. E. MATHER, B. G. ROSSNAGEL, K. J. KASHA and A. KLEINHOF, 1996 Regions of the genome that affect agronomic performance in two-row barley. *Crop Sci.* **36**: 1053–1062.
- WHITTAKER, J. C., R. THOMPSON and M. C. DENHAM, 2000 Marker-assisted selection using ridge regression. *Genet. Res.* **75**: 249–252.
- YI, N., and S. XU, 2001 Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics* **157**: 1759–1771.
- ZENG, Z-B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA* **90**: 10972–10976.

Communicating editor: J. B. WALSH

