Stochastic Search Variable Selection for Identifying Multiple Quantitative Trait Loci

Nengjun Yi,*†,1 Varghese George*† and David B. Allison*†,‡

*Department of Biostatistics, †Section on Statistical Genetics, ‡Clinical Nutrition Research Center, University of Alabama, Birmingham, Alabama 35294-0022

Manuscript received December 5, 2002
Accepted for publication March 13, 2003

ABSTRACT

In this article, we utilize stochastic search variable selection methodology to develop a Bayesian method for identifying multiple quantitative trait loci (QTL) for complex traits in experimental designs. The proposed procedure entails embedding multiple regression in a hierarchical normal mixture model, where latent indicators for all markers are used to identify the multiple markers. The markers with significant effects can be identified as those with higher posterior probability included in the model. A simple and easy-to-use Gibbs sampler is employed to generate samples from the joint posterior distribution of all unknowns including the latent indicators, genetic effects for all markers, and other model parameters. The proposed method was evaluated using simulated data and illustrated using a real data set. The results demonstrate that the proposed method works well under typical situations of most QTL studies in terms of number of markers and marker density.

Most complex traits important to evolution, animal and plant breeding, and medical genetics are influenced by the segregation of multiple genes [quantitative trait loci (QTL)] and environmental factors. There is strong interest in inferring the number, genomic locations, and genetic effects of QTL. Recently, the most widely used methods are interval mapping (Lander and Botstein 1989; Haley and Knott 1992). These methods are developed on the basis of a single-QTL model and detect QTL effects at different genomic locations separately. Although these methods have been successfully applied to detect QTL for a number of traits in a number of organisms, they may result in biased estimates for QTL locations and effects when the traits are actually controlled by multiple, especially linked, QTL (e.g., Haley and Knott 1992).

For complex traits governed by multiple QTL, it is necessary to take the whole genome into account for estimating the number, locations, and genetic effects of QTL. It has been recently shown both theoretically and empirically that multiple-QTL methods can improve power in detecting QTL and eliminate biases in estimates of QTL locations and genetic effects that can be introduced by using a single-QTL model (e.g., Haley and Knott 1992). Composite interval mapping creates a relatively simple and systematic procedure to map multiple QTL (Jansen and Stam 1994; Zeng 1994). This method detects and estimates each individual QTL by conditioning the test on other selected markers to absorb effects of other QTL. In the last decades, several statistical methods have been developed to detect multiple QTL and estimate their locations and effects simultaneously, including the multiple-interval mapping approach (Kao et al. 1999; Zeng et al. 2000), variable selection methods (Ball 2001; Piepho and Gauch 2001; Broman and Speed 2002), and Bayesian methodology with the reversible-jump Markov chain Monte Carlo algorithm (Satagopan and Yandell 1996; Satagopan et al. 1996; Heath 1997; Sillanpää and Arjas 1998; Stephens and Fisch 1998; Xu and Yi 2000; Yi and Xu 2000, 2001; Hoeschele 2001). These methods treat mapping multiple QTL as a problem of model determination and variable selection (Sillanpää and Corander 2002).

In this study, we propose an alternative Bayesian method for identifying multiple quantitative trait loci in experimental designs. Our method is based on a variable selection method, called stochastic search variable selection (SSVS), developed by George and McCulloch (1993). SSVS was originally introduced for linear regression models and has been adopted for more complex models such as generalized linear models (George and McCulloch 1997), log-linear models (Ntzoufras et al. 1997), and multivariate regression models (Brown et al. 1998). The difference between SSVS and other variable selection approaches is that the dimensionality is kept constant across all possible models by limiting the posterior distribution of nonsignificant terms in a small neighborhood of zero instead of removing them from the model as is usually done. Due to this unique property, SSVS is able to (1) be easily implemented via the Gibbs sampler,
(2) evaluate each variable effect on the dependent response, and (3) provide the posterior probability that each variable should be included in the model.

In most QTL studies, a large number of markers are available across the genome, and these markers are usually closely related. It has been hypothesized that the genetic variation of most quantitative traits is actually controlled by a few loci with large effects and a large number of loci with small effects (e.g., Lynch and Walsh 1998). Therefore, only a small number of markers are expected to have large effects on the trait because of being linked to large-effect loci, and most of the markers have nonsignificant effects. Our method considers all markers simultaneously and is able to evaluate not only marker effects of the entire genome, but also the posterior probability of each marker having significant effects.

METHODS

Linear model: We describe the method primarily for a mapping population with only two segregating genotypes, e.g., a backcross, double-haploid lines (DHLs), or recombinant inbred lines (RILs). Assume that we observe \( K \) markers along the genome. Among the \( K \) markers, some may be tightly linked to genes with large effects and therefore have large effects, and others may have only weak effects. Our aim here is to identify which markers are tightly linked to genes with large effects and to estimate the magnitude of their effects. For a continuously distributed trait, the observed phenotypic value of individual \( i \), \( y_i \), can be described by the linear model,

\[
y_i = \mu + \sum_{j=1}^{K} x_{ij} \alpha_j + e_i, \tag{1}
\]

where \( \mu \) is the population mean, \( x_{ij} \) denotes the genotype of marker \( j \) for individual \( i \) and is defined by 0.5 or \(-0.5\) for the two genotypes in the mapping population, \( \alpha_j \) is the effect size associated with marker \( j \), and \( e_i \) is the residual error assumed to follow \( N(0, \sigma_e^2) \).

In practice, some marker data may be missing. Two methods deal with missing marker data. The first method is to replace the missing genotype \( x_{ij} \) by its conditional expectation \( E(x_{ij} | M) = 0.5 p(x_{ij} = 0.5 | M) - 0.5 p(x_{ij} = -0.5 | M) \), where \( M \) is observed marker data for individual \( i \), and \( p(x_{ij} = 0.5 | M) \) and \( p(x_{ij} = -0.5 | M) \) are the conditional probabilities that marker \( j \) for individual \( i \) takes the two genotypes, respectively, and can be calculated using the multipoint method (Jiang and Zeng 1997). The second method is to impute the missing marker genotypes by sampling from the corresponding fully conditional probability distribution. If \( x_{ij} \) is missing, the fully conditional distribution can be derived as

\[
p(x_{ij} = 0.5 | y_i, x_{i,-j}, \mu, \alpha, \sigma_e^2) = \frac{p(y_i | x_{i,-j}, x_{ij} = 0.5, \mu, \alpha, \sigma_e^2) p(x_{ij} = 0.5 | x_{i,-j}, x_{(j+1)})}{\sum_{x_{ij}} p(y_i | x_{i,-j}, x_{ij}, \mu, \alpha, \sigma_e^2) p(x_{ij} | x_{i,-j}, x_{(j+1)})}, \tag{2}
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_K) \) and \( x_{(-j)} = (x_{i,1}, \ldots, x_{i,j-1}, x_{i,(j+1)}, \ldots, x_{i,K}) \), the conditional probability \( p(x_{ij} = 0.5 | x_{i,-j}, x_{(j+1)}) \) depends on the recombinant rates between marker \( j \) and its flanking markers \( (j - 1) \) and \( (j + 1) \), and

\[
p(y_i | x_{i,-j}, x_{ij} = 0.5, \mu, \alpha, \sigma_e^2) \text{ is a normal density function with mean } \mu + \sum_{k=1}^{K} \alpha_k x_{ik} \text{ and variance } \sigma_e^2. \]

Obviously, the first method ignores the probability distributions of the missing genotypes and provides approximate estimates of the missing genotypes. In contrast, the second method can take the probability distributions into account. In this study, we use the second method to describe our Bayesian approach.

Stochastic search marker selection: In variable selection problems, statistical models can be naturally represented by a set of binary indicator variables \( \gamma = (\gamma_1, \ldots, \gamma_K) \), where \( \gamma_j = 1 \) or 0 represents the presence or absence of covariate \( j \) in the model, respectively. The difference between SSVS and other variable selection approaches is that the dimension of parameter space remains unchanged so that the Gibbs sampler can be easily used to explore both the model space and the parameter space (George and McCulloch 1993).

The SSVS constructs the prior distribution for \( (\gamma, \alpha) \) in two stages. The prior distribution of the model indicator variables \( p(\gamma) \) is chosen to reflect prior belief in whether particular markers are linked to QTL. A simple choice might have the \( \gamma_j \)'s independent, so that

\[
p(\gamma) = \prod_{j=1}^{K} p(\gamma_j). \tag{3}
\]

When no information is available, a uniform prior is chosen for each \( \gamma_j \), i.e., \( p(\gamma_j = 0) = p(\gamma_j = 1) = 0.5 \).

The marker effects \( \alpha_j \) (\( j = 1, \ldots, K \)) are given normal prior distributions conditional on the corresponding indicators \( \gamma_j \):

\[
\alpha_j | \gamma_j \sim (1 - \gamma_j) N(0, \tau_j^2) + \gamma_j N(0, c_j^2 \tau_j^2), \quad j = 1, \ldots, K. \tag{4}
\]

The prior parameters \( \tau_j^2 \) and \( c_j^2 \) are chosen so that \( \tau_j^2 \) is “small” and \( c_j^2 \tau_j^2 \) is “large.” Hence, if \( \gamma_j = 0 \), the magnitude of the effect \( \alpha_j \) is small and then the prior distribution for \( \alpha_j \) forces this parameter to be close to zero. If \( \gamma_j = 1 \), the magnitude of the effect \( \alpha_j \) is large and then a nonzero estimate of \( \alpha_j \) should be included in the model and its posterior distribution will largely be determined by the data. On the basis of the above prior specification, a multivariate normal distribution can be used as the joint prior distribution for \( \alpha \) conditional on \( \gamma \), given by

\[
\alpha | \gamma \sim N_K(0, D_{\gamma} \mathbf{R}_{\gamma}), \tag{5}
\]

where \( \mathbf{R} \) is the prior correlation matrix that is usually assigned to be \( \mathbf{R} = \mathbf{I} \) or \( \mathbf{R} \propto (\mathbf{xx}^\top)^{-1} \), and \( \mathbf{D}_{\gamma} = \text{diag}[a_1 \tau_1, \ldots, a_K \tau_K] \) with \( a_i = 1 \) if \( \gamma_i = 0 \) and \( a_i = c_i \) if \( \gamma_i = 1 \). The
prior distribution for \( \mu \) is assumed to be Normal \( N(\eta, \tau^2) \) with prespecified prior mean \( \eta \) and prior variance \( \tau^2 \).

The prior for \( \sigma^2 \) is chosen to be of a scaled inverse -\( \chi^2 \) distribution, \( \text{Inv}(-\chi^2(v_0, \sigma^2)) \), with known hyperparameters \( v_0 \) and \( \sigma^2 \).

On the basis of the prior specifications described above, we can use the Gibbs sampler to generate samples from the posterior distribution \( p(\mu, \alpha, \sigma^2; y|\mathbf{M}) \). Starting with an initial value \( (\mu(0), \alpha(0), \sigma^2(0), \gamma(0)) \), the Gibbs sampler proceeds as follows:

1. Sample the missing marker genotypes from the full conditional posterior distributions described in Equation 2.
2. Sample \( \mu \) from the full conditional posterior distribution:

\[
p(\mu|y, \mathbf{x}, \alpha, \sigma^2) = N\left( \frac{v_0\mu + \sum_{i=1}^{n} x_i y_i}{v_0 + n}, \frac{1}{\tau^2 + n/\sigma^2} \right).
\]

3. The full conditional posterior distribution of \( \alpha \) is multivariate normal, \( N_n((\mathbf{x}x^T + \sigma^2_\alpha \mathbf{D} T \mathbf{R} \mathbf{D}_T)^{-1})^{-1}\mathbf{x}^T(y - \mu), \sigma^2_\alpha(\mathbf{x}x^T + \sigma^2_\alpha \mathbf{D} T \mathbf{R} \mathbf{D}_T)^{-1} \).

Sampling from this distribution requires recomputing \( (\mathbf{x}x^T + \sigma^2_\alpha \mathbf{D} T \mathbf{R} \mathbf{D}_T)^{-1} \) on the basis of new values of \( \sigma^2_\alpha \) and \( \gamma \) and thus may be costly. To avoid computing \( (\mathbf{x}x^T + \sigma^2_\alpha \mathbf{D} T \mathbf{R} \mathbf{D}_T)^{-1} \) and \( \sigma^2_\alpha(\mathbf{x}x^T + \sigma^2_\alpha \mathbf{D} T \mathbf{R} \mathbf{D}_T)^{-1} \), we sample \( \alpha_j \) (\( j = 1, \ldots, K \)) from the full conditional posterior distribution \( p(\alpha_j|y, \mathbf{x}, \mu, \alpha_{(-j)}, \sigma^2_\gamma) \), which is normal distribution (Wang et al. 1994), where \( \alpha_{(-j)} \) denotes all terms of \( \alpha \) except \( \alpha_j \).

4. Sample \( \sigma^2_\alpha \) from the full conditional posterior distribution:

\[
p(\sigma^2_\alpha|y, \mathbf{x}, \mu, \alpha) = \text{Inv}(-\chi^2(v_0 + n, \frac{v_0\sigma^2_\alpha + \sum_{i=1}^{n} x_i^2(y_i - \mu - \sum_{j=1}^{K} x_i \alpha_j)^2}{v_0 + n})).
\]

5. Sample \( \gamma \) from \( p(\gamma|y, \mathbf{x}, \gamma_{(-j)}, \mu, \alpha, \sigma^2_\alpha) = p(\gamma|\gamma_{(-j)}, \mu, \alpha, \sigma^2_\gamma) \), which is Bernoulli with probability

\[
p(\gamma|\gamma_{(-j)}, \mu, \alpha, \sigma^2_\gamma) = \frac{p(\alpha|\gamma_{(-j)}, \gamma = 1)p(\gamma = 1)}{p(\alpha|\gamma_{(-j)}, \gamma = 1)p(\gamma = 1) + p(\alpha|\gamma_{(-j)}, \gamma = 0)p(\gamma = 1)},
\]

where \( \gamma_{(-j)} \) denotes all terms of \( \gamma \) except \( \gamma_j \).

The above steps are repeated until a certain criterion for convergence is reached. The posterior sample \( ((\mu(t), \alpha(t), \sigma^2_\alpha(t), \gamma(t)) : t = 1, 2, \ldots) \) converges in distribution to the joint posterior distribution, \( p(\mu, \alpha, \sigma^2_\alpha, \gamma|\mathbf{y}, \mathbf{M}) \).

The embedded subsequence \( \{\gamma(t) : t = 1, 2, \ldots\} \) thus converges to \( p(\gamma|\mathbf{y}, \mathbf{M}) \). Generally, the markers with large effects will appear most frequently and quickly, making them easier to identify. Therefore, markers with high posterior probability included in the model will most probably be linked to large-effect QTL.

### TABLE 1: Locations and effects of simulated QTL

<table>
<thead>
<tr>
<th>Chromosome 1</th>
<th>Chromosome 2</th>
<th>Chromosome 3</th>
<th>Chromosome 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location (cM)</td>
<td>Effect</td>
<td>Location (cM)</td>
<td>Effect</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>24</td>
<td>1.0</td>
<td>16</td>
<td>-1.0</td>
</tr>
<tr>
<td>30.5</td>
<td>0.1</td>
<td>30.5</td>
<td>0.1</td>
</tr>
<tr>
<td>44</td>
<td>1.0</td>
<td>37</td>
<td>1.0</td>
</tr>
<tr>
<td>60.5</td>
<td>0.1</td>
<td>60.5</td>
<td>0.1</td>
</tr>
<tr>
<td>90.5</td>
<td>0.1</td>
<td>90.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

### SIMULATION STUDIES AND REAL DATA ANALYSIS

**Simulation studies:** The applicability of the proposed method was demonstrated by analyzing simulated data. The experimental sample was from a backcross and contained 300 segregating individuals. Four chromosomes with length 100 cM each were simulated. Twenty-one codominant markers were evenly placed on each chromosome with marker intervals of 5 cM each. We simulated 8 large-effect QTL and 16 small-effect QTL controlling the expression of a quantitative trait. The locations of the simulated QTL and their genetic effects are shown in Table 1. The overall mean and the residual variance were set to be \( \mu = 1 \) and \( \sigma^2 = 1 \), respectively. The genetic variance of QTL \( j \) is calculated by \( v_i = a^2_i/4 \), where \( a_i \) is the true genetic effect. Ignoring the covariance due to linkage, the total genetic variances for the 8 large-effect QTL and 16 small-effect QTL are 2 and 0.04, respectively. Therefore, the phenotypic variances explained by each large-effect QTL and each small-effect QTL are 8 and 0.08\%, respectively. We randomly generated missing markers of 10\%. The design was replicated five times and analyzed using the proposed method. The results averaged over the five replicates were reported.

For each analysis, the initial values \( (\mu^{(0)}, \alpha^{(0)}, \sigma^2_{\alpha^{(0)}}, \gamma^{(0)}) \) were randomly generated from their priors. We used the uniform distribution as prior for \( \gamma \) as described earlier. Following the principles developed in George and McCulloch (1993, 1997) for choosing \( \tau_j \) and \( \epsilon_j \), three different prior variances, i.e., \( (\tau_j^2, \epsilon_j^2) = (0.001, 10), (0.01, 10), (0.01, 100) \), were used for the conditional prior distribution of the genetic effect \( \alpha_j \).
The prior correlation matrix was assigned to be the identity matrix, i.e., $\mathbf{R} = \mathbf{I}$. The prior distribution for $\mu$ was $N(0, 2)$. The hyperparameter $v_0$ for $\sigma^2_e$ was set to zero, which yields the noninformative prior distribution $p(\sigma^2_e) \propto \sigma^{-2}$ (Gelman et al. 1995).

The Gibbs sampler was run for 50,000 cycles after discarding the first 2000 cycles for the burn-in period. It took $\sim 1$ hr to generate each sample with a C++ program on a Pentium 4 PC. The chain was thinned (saved one iteration in every 5 cycles) to reduce serial correlation in the stored samples so that the total number of samples kept in the post-Bayesian analysis was 10,000 (Gelman et al. 1995). The stored sample was used to infer the parameters of interest.

The estimated posterior probabilities for the marker indicators $\gamma_j (j = 1, \ldots, K)$ are given in Figure 1 (left). The posterior probability $p(\gamma_j | y, \mathbf{M})$ was obtained by counting the number of samples in which the marker indicator $\gamma_j$ is 1, divided by the total number of samples. As shown in Figure 1, for almost all markers, the posterior probabilities for the first prior setting were larger than those for the other two settings, and the
posterior probabilities for the second prior setting were larger than those for the third. For the three sets of prior variances, however, the profiles of the posterior probability distributions were similar. These profiles are very peaked, suggesting that the markers corresponding to the peaks have much larger effects than the rest. From these profiles, we also found that for most situations two markers flanking a large-effect QTL have very different posterior probabilities, one being large and another being close to zero. Therefore, our method may be powerful in distinguishing closely linked markers. There are a total of eight main peaks along the four simulated chromosomes on the profiles, and two are on one chromosome. It can be observed that the markers corresponding to the peaks are those that are the closest to the simulated eight large-effect QTL. Therefore, our Bayesian method was shown to be powerful for identifying multiple QTL. None of the simulated small-effect QTL were identified in our analyses. Actually, these QTL had effects close to zero and thus were picked up only occasionally.

The profiles of marker effects are displayed in Figure 1 (right). Although empirical posterior distribution for each marker effect can be depicted, for simplicity we report only the posterior mean over the samples. The marker effects were estimated to be essentially identical for the three sets of prior variances \((\tau_i^2, c_i^2)\). Therefore, these prior variances had an ignorable influence on the posterior inference about the marker effects. As in the case of the posterior probability distribution of marker indicators, the profiles of marker effects also have eight obvious peaks, each corresponding to a marker that is the closest to a large-effect QTL. For markers far from the large-effect QTL, their effects were estimated to be close to zero. From Figure 1, it can be seen that the accuracy of the estimate for a marker effect depends...
on the estimated posterior probability of the marker indicator. When the posterior probability was estimated to be close to one, the estimate of the marker effect was close to the true value. Otherwise, the marker effects were slightly underestimated. These results are expected because the marker effects with the corresponding indicators being zero are forced to be close to zero by the priors. However, we observed that the conditional estimates of marker effects were close to the true value if we used only the posterior samples with the corresponding indicators equal to one.

The empirical posterior distributions for the overall mean and the residual variance are depicted in Figure 2 (a–f). The estimated means for these two parameters were very close to the simulated values and the standard deviations were small, showing that the overall mean and the residual variance were estimated with precision.

For comparison, we also performed the single-marker analyses with the simple regression method for each marker and the usual multiple regression analysis with all markers as predictors. For the single-marker analyses, the profiles of the t-test statistics and the marker effects on chromosome 1 are shown in Figure 3. Apparently, these two profiles have only one peak covering a wide range. This shows that the single-marker analyses fail to separate the two linked QTL. It is also obvious that the marker effects were seriously overestimated. Since the marker density is quite high, results of single-marker analyses should be close to that of the interval mapping. Therefore, the proposed method was shown to be more powerful than the widely used interval-mapping method for detecting multiple QTL. Figure 4 shows the plot of the marker effects against the genome location (centimorgans) of the markers from multiple regression analysis. Obviously, the effects of most markers far from the large-effect QTL have not shrunk in the usual multiple regression analysis, indicating that multiple regression failed to detect clear signals of QTL. A common feature of the proposed Bayesian method and the usual multiple regression method is that all markers are included in the model. The clear advantage of the proposed method is that it uses two different prior distributions for the markers, which force the posterior means of insignificant markers to be close to zero and the posterior distributions of significant markers to be determined by the data.

**Real data analysis:** Data from the North American Barley Genome Mapping Project (Tinker et al. 1996) were analyzed using the proposed Bayesian method. Seven traits were investigated in the project: heading, yield, maturity, height, lodging, kernel weight, and test weight. We present only the results of “heading” here. The DH (double-haploid) population contained 145 lines (n = 145), each grown in a range of environments. A total of 127 mapped markers (K = 127) covering a 1500-cM genome along seven linkage groups were used in the analysis. The average phenotypic values across the environments were calculated for each line and these average values were treated as the original phenotypic values (y) for the analysis. These phenotypic values were further standardized. The standardized records were used in the analysis.

The prior distributions for (μ, α, σ^2, γ), the prior variances (τ^2, c_τ^2), the length of the Gibbs sampler, and the thinning scheme of the posterior sample were set to be the same as those in the analysis of our simulated data described above. The initial values (μ^(0), α^(0), σ^2^(0), γ^(0)) were randomly generated from their priors.

For the three different prior specifications, the plots of the posterior probabilities of the marker indicators are shown in Figure 5, and the marker effects are depicted in Figure 6. As observed in our simulation studies, for almost all markers, the posterior probabilities for the first prior setting were larger than those for the other two settings, and the posterior probabilities for the second prior setting were larger than those for the third. However, the profiles of the posterior probability distributions were proximate. As shown in Figure 6, the marker effects were estimated to be essentially identical for the three sets of prior variances (τ^2, c_τ^2). For the first prior setting, it was found that five markers with posterior probabilities from 0.75 to 0.95 and marker effects of ~ ±0.4, are located at chromosomes 1, 3, 4, and 6, respectively. We also found five markers on chromosomes 3, 4, 5, and 6, respectively, with posterior probabilities from 0.4 to 0.6. Using the interval mapping of Lander and Botstein (1989) and the composite interval mapping of Zeng (1994), however, Tinker et al.
(1996) declared only three QTL on chromosomes 1, 4, and 7, respectively, as significant (data not shown here). Two markers on chromosome 1 were found to have the posterior probability of $\sim 0.76$ and the effect of $\sim -0.4$. However, Tinker et al. (1996) declared only one QTL on chromosome 1 as significant.

For the first prior setting, the posterior means of the overall mean and the residual variance were estimated to be $\hat{\mu} = 0.0882$ and $\hat{\sigma}_e^2 = 0.3102$, respectively. The proportion of phenotypic variance explained by the markers is calculated as $\hat{h}^2 = (\hat{\sigma}_j^2 - \hat{\sigma}_e^2)/\hat{\sigma}_j^2$, where $\hat{\sigma}_j^2 = 1$ is the phenotypic variance for the standardized phenotype. Thus, the proportion of phenotypic variance explained by the markers was estimated to be $\sim 69\%$. From the estimates of the marker effects $\hat{\alpha}_j (j = 1, \ldots, K)$, we calculated the proportion of phenotypic variance explained by marker $j$ as $\hat{h}_j^2 = \hat{\alpha}_j^2 / 4$ and found that the proportion of phenotypic variance explained by each of the five strongest markers was $\sim 4\%$.

DISCUSSION

Mapping multiple QTL can be viewed essentially as a problem of model selection (e.g., Broman and Speed 2002; Sillanpää and Corander 2002). A variety of statistical selection procedures including both non-Bayesian and Bayesian methods have been developed for conventional statistical models. Some of these procedures have been modified to map multiple QTL. In this study, we developed a Markov chain Monte Carlo (MCMC) algorithm on the basis of the SSVS approach of George and McCulloch (1993) for identifying multiple markers. The proposed method was shown to be extremely efficient under typical situations of most QTL studies in terms of the number of markers and the marker density. Compared with the existing Bayesian methods, such as the reversible-jump MCMC, the SSVS approach has advantages on simplicity of computation and diagnosis of convergence (George and McCulloch 1997). The SSVS procedure can even be implemented using the publicly available software BUGS (Congdon 2002) and thus can be widely used in QTL studies.

An essential element of the performance of our Gibbs sampler is its ability to move between two different values of the indicator variable $\gamma_j$. In the analyses of the simulated data and real data, the value of $\gamma_j$ changed frequently, suggesting that the proposed algorithm mixes well and the chain converges quickly. However, as described in George and McCulloch (1993, 1997), convergence can be very slow and thus computational problems can arise when $c_2^j$ is set too large. This setup can lead to very small transition probabilities for $\gamma_j$ to go from 0 to 1 or from 1 to 0. After extensive testing, George and McCulloch (1997) indicated that these problems can be avoided whenever $c_2^j \leq 10,000$. Also, mixing behavior and convergence of the chain is expected to be affected by marker density. In the cases of high-density maps with hundreds of markers, one might consider a second iteration of SSVS with a reduced set of markers based on the first run (George and McCulloch 1993). This two-stage strategy may improve accuracy of estimating the marker effects and the posterior probabilities.
Recently, Xu (2003) proposed a Bayesian method under the random regression model to simultaneously estimate genetic effects associated with markers of the entire genome in inbred line crosses. In his Bayesian framework, each genetic effect was assigned a normal prior distribution with mean zero and a unique variance. The effect-specific prior variance was further assigned a vague prior so that the variance was 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. For a backcross population with $K$ markers, Xu’s method needs to estimate $K$ different marker effects $\alpha_j$ ($j = 1, \ldots, K$) and $K$ different variances $\sigma^2_j$ ($j = 1, \ldots, K$), where $\alpha_j \sim N(0,\sigma^2_j)$. Although this approach can evaluate each marker effect, it does not provide a probability statement about statistical significance for marker effects. In our approach, indicator variables are introduced but all effects included in the model have the same prior variance and all effects excluded from the model have another common prior variance. Therefore, our method not only estimates each marker effect, but also provides the posterior probability that each marker has a significant effect on the trait. The introduction of indicator variables may allow a large number of markers to be included in the model. It is worth noting that both the random-model approach and our method include all markers in the analysis and thus may have the ability to control the genetic variances of a large number of small-effect QTL. Whether this property can improve power...
in detecting multiple QTL and estimating the genetic effects deserves further investigation.

We have applied the SSVS approach to identify multiple QTL by analyzing all markers of the whole genome. When the markers are densely and regularly spaced, the marker analysis would provide reasonable estimates of marker effects and marker posterior probabilities even when QTL are located in the marker intervals. If the marker density is low and irregularly spaced, however, the marker analysis will be biased. In these situations, however, we can extend the proposed method to allow for finer structure mapping by two ways. The first approach is to use the multiple imputation method to generate the missing genotypes at grids of points between markers (Sen and Churchill 2001). The imputed genotypes are then incorporated into our Bayesian procedure. The second approach is to substitute markers by positions in the marker intervals if we assume that at most one QTL is on any marker interval. This approach requires searching the optimal positions within the marker intervals. The algorithms for updating QTL positions have been developed (e.g., Xu and Yi 2000; Yi and Xu 2000, 2001) and can be easily incorporated into our procedure.

The SSVS approach has been extended to the multivariate regression model (Brown et al. 1998). In QTL-mapping studies, the joint analysis of multiple traits can provide formal procedures to test a number of biologically interesting hypotheses concerning the nature of genetic correlations between different traits. Under cer-
tain situations, the joint analysis can improve statistical power in detecting QTL and estimating the genetic parameters. We can extend the proposed method by applying the SSVS approach to jointly identify QTL for correlated multiple traits. In this study, we considered mapping multiple QTL under the nonepistatic model. A growing number of experiments provide strong evidence of the presence of interactions between genes for many complex traits. Under the epistatic model, the number of genetic effects increases exponentially as the number of markers increases. The multiple-stage SSVS approach can be employed to identify interacting QTL.

We are grateful to two anonymous reviewers for their helpful comments. This work was supported by National Institutes of Health grants R01ES009912, P41RR06609, R01DK054298, and P30DK56336 to D.B.A.

**LITERATURE CITED**


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.


Communicating editor: J. B. Walsh