Bayesian Methods for Quantitative Trait Loci Mapping Based on Model Selection: Approximate Analysis Using the Bayesian Information Criterion

Roderick D. Ball

New Zealand Forest Research Institute, Rotorua 3201, New Zealand

Manuscript received April 13, 2000
Accepted for publication August 30, 2001

ABSTRACT

We describe an approximate method for the analysis of quantitative trait loci (QTL) based on model selection from multiple regression models with trait values regressed on marker genotypes, using a modification of the easily calculated Bayesian information criterion to estimate the posterior probability of models with various subsets of markers as variables. The BIC-δ criterion, with the parameter δ increasing the penalty for additional variables in a model, is further modified to incorporate prior information, and missing values are handled by multiple imputation. Marginal probabilities for model sizes are calculated, and the posterior probability of nonzero model size is interpreted as the posterior probability of existence of a QTL linked to one or more markers. The method is demonstrated on analysis of associations between wood density and markers on two linkage groups in *Pinus radiata*. Selection bias, which is the bias that results from using the same data to both select the variables in a model and estimate the coefficients, is shown to be a problem for commonly used non-Bayesian methods for QTL mapping, which do not average over alternative possible models that are consistent with the data.

QUANTITATIVE trait loci (QTL) mapping is the process of finding and estimating associations between a continuous quantitative trait and a set of DNA markers that have been previously placed on a genetic map, with the ultimate goal of determining the genetic architecture of a trait, or finding markers that can be used to select for preferred values of the trait. The map is generally assumed to be known and correct and should cover a significant proportion of the genome. QTL mapping works on the principle that if a locus (called a QTL) on the genome is causing variation in a trait, and data are obtained from a cross (or pedigree) in which the QTL is segregating, then values of the trait will be correlated with markers linked to that locus. The closer the marker, the closer the correlation. For a marker at a given distance from the QTL, the larger the effect of the QTL, the larger the effect of the marker, as can be estimated from differences between subsets of the population with different marker classes. The statistical problem is to estimate the effects and locations of QTL or the effects of using associated markers to select progeny. The problem is challenging statistically because one or more QTL for a trait could be located anywhere on the genome.

Non-Bayesian QTL mapping: A common approach is to carry out a hypothesis test at each marker (a *t*-test or ANOVA) or at each of a series of points on the genome (interval mapping; Lander and Botstein 1989). An alternative to interval mapping, also involving multiple hypothesis tests, is based on regression on flanking markers (Haley and Knott 1992). See Paterson (1995) and Doerge *et al.* (1997) for reviews.

Markers or loci where the test statistic exceeds a threshold are chosen and considered to be "detected." Problems with these methods are that QTL are often detected but where an independent verification population is used the markers are subsequently not verified and/or the estimated effects are much smaller in the verification population (see, *e.g.*, Beavis 1994; Wilcox *et al.* 1997; Melchinger *et al.* 1998). The latter problem is an example of selection bias, which is well known to statisticians in the context of stepwise regression (Miller 1990). Selection bias occurs when the same data are used to both select the variables in a regression model and to estimate the coefficients.

Bayesian statistics: Bayesian statistics aim to compute probability distributions for the underlying parameters in a model. With this information it is, in principle, possible to compute probabilities of any events or quantities of interest such as the probability of a linked QTL in a region or the expected gain from marker-aided selection.

An important aspect of Bayesian analysis is the use of *marginal distributions*. In the method of this article, the probability distributions of estimates and predictions are averaged over the values of parameters in the model and over possible models, rather than with parameters

Address for correspondence: New Zealand Forest Research Institute, P.B. 3020, Rotorua 3201, New Zealand.
E-mail: rod.ball@forestresearch.co.nz

set to their most likely values in the most likely model, as is usually the case in non-Bayesian methods, such as maximum likelihood. In a single model, estimates and confidence intervals from maximum likelihood will be similar to their Bayesian counterparts provided the sample size is large enough. More significant differences arise, however, when testing "precise hypotheses" (Berger and Berry 1988). Existence or nonexistence of a QTL linked to a particular marker is one example. This difference occurs because Bayesian inference considers the probability of the data under each of the two possible models, *e.g.*, H_0:$\theta = 0$ and H_1:$\theta \neq 0$. The non-Bayesian hypothesis test considers the tail probability for a test statistic under H_0, which can be shown to be approximately equivalent to a tail probability of the posterior distribution for the parameter, θ, being tested *under* H_1. This does not allow for the finite nonzero *prior probability* (it must be nonzero or we would not need to test it) that H_0 is true. More generally, *where multiple models are consistent with the data, it is necessary to consider all these models according to their probabilities for valid statistical inference* (*cf.* Raftery 1995, Raftery *et al.* 1997).

Bayesian QTL mapping: A number of articles on Bayesian approaches to QTL mapping have appeared (reviewed by Hoeschele *et al.* 1997).

Several more recent articles have appeared that simultaneously consider multiple models with different numbers of QTL (Satagopan and Yandell 1996; Satagopan *et al.* 1996; Heath 1997; Sillanpää and Arjas 1998, 1999; Stephens and Fisch 1998). These articles use the "reversible jump" methodology of Green (1995) for constructing a sampler that jumps between models of different dimension. A major challenge remains to obtain a rapidly converging sampler for the full Bayesian model (D. A. Stephens, personal communication). The methods are complex to program and as yet there is no publicly available program with demonstrated rapid convergence.

More easily implemented and faster methods are useful, for people without access to the above programs, for checking the results of the more complex programs, or for preliminary or exploratory analysis as data are collected and may be useful for generating approximate starting values for algorithms for the full Bayesian models. Moreover, when data are limited and the actual genetic architecture is unknown, more generic statistical methods based on linear models with main effects and interactions may be more appropriate and more efficient (in terms of ease of application, rate of convergence, and validity) than the detailed modeling of genetic parameters from a particular genetic architecture. The information from the simpler methods is adequate to assess the evidence for the existence of QTL in a region, for marker-aided selection, or for determining QTL location to within a resolution of the distance between markers.

OVERVIEW: BAYESIAN QTL MAPPING USING MODEL SELECTION

Our approach is to relate trait values directly to marker genotypes, using multiple linear regression. Since there are many markers in a typical cross, most of these will not be near to a QTL. Therefore it is necessary to choose a model or models with subsets of markers selected. Broman (1997) advocated a stepwise regression approach for choosing the "best" model. However, we shall see that particularly with small sample sizes there can be a multiplicity of models that are compatible with the data, and these alternative models need to be considered along with their probabilities (Raftery 1995, Raftery *et al.* 1997).

Our strategy is to build on methods for model selection in linear models from the statistical literature, starting with the Bayesian information criterion (BIC; Schwartz 1978) in this article, and a modification of the Bayesian method for model selection in *hierarchical linear models* (a whole family of linear models combined in an overarching single model) of George and McCulloch (1993) in future work. Our methodology is to the full Bayesian approach as the regression method of Haley and Knott (1992) is to interval mapping—a simpler more easily calculated method that nevertheless captures (at least approximately) the important aspects of the full Bayesian analysis, *i.e.*, the posterior probability that a QTL is located in a given region, or the marginal distribution for the size of effects.

There is a major difference between our approach and others that consider multiple models. Previous approaches consider the model as specifying only the number of QTL on each linkage group, while the locations of the QTL are parameters in the model. In our approach the QTL are at fixed locations. In other words a QTL in our model really is a quantitative trait *locus.* A QTL at a different position is considered a different QTL and is represented by a different model. This increases the number of models but simplifies the analysis of a given model.

Statistical inference is carried out by combining information from each model according to its posterior probability. The following quantities of interest are estimated below: (i) marginal probabilities of selection of one or more markers in a linkage group or region, (ii) model-averaged effect of a marker, and (iii) model-averaged effect of allelic substitution of a marker. These quantities have the following interpretations: (i) probability of existence of one or more QTL in the linkage group or region, (ii) posterior expected gain attributable to QTL in the immediate vicinity of the marker (defined as the region closer to the marker than to any other marker), and (iii) posterior expected gain from marker-aided selection using the marker, resulting from all QTL linked to the marker, respectively. The need for model averaging is demonstrated by estimation of the selection bias that results when the effects of markers and the effects of allelic substitution are estimated conditional on selection. Further discussion of selection bias is given below and a small simulation study demonstrating the effect of model averaging on selection bias is given in the APPENDIX.

In this article approximate posterior probabilities for models are obtained using a modification of the easily calculated BIC (SCHWARTZ 1978). The probabilities are approximate but good enough to give a rough indication. Moreover, it may be possible, by adjusting a single parameter, to fine tune the method. We apply the method to QTL mapping data for wood density in *Pinus radiata* and demonstrate the need to consider multiple models in assessing the probability of existence of a QTL and obtaining estimates of the effects of allelic substitution at a marker free of selection bias.

DATA AND METHODS

The method is demonstrated on analysis of associations between wood density and markers from two linkage groups in *P. radiata*.

Marker and trait data were obtained from a single full-sib family with parents 850.55 and 850.96, for the purpose of QTL detection.

Trait data: Two 5-mm pith-to-bark cores were taken from each tree. Usable data were available from 93 trees. Wood density was assessed from each of the cores by X-ray densitometry (COWN and CLEMENT 1983). The traits considered here were juvenile wood density at ages 1–5 years (estimated as the area weighted average of rings 1–5) as the average outerwood density (estimated as the average density of the outer 5 cm from each core), adjusted for site and replicate differences, and standardized. These traits are similar to the traits WD1_5, WD14 (outerwood density, not standardized) analyzed by KUMAR *et al.* (2000).

Marker data: There were 126 markers of various types [randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), and simple sequence repeat (SSR)] in 23 linkage groups with from 2 to 16 markers per linkage group, which were segregating in pseudobackcross configuration, for the 850.55 parent. There were 1171 missing marker values (10% of the marker data). These values were due to indistinct bands or PCR failure and are assumed missing at random. All 93 trees had one or more missing marker values. For further information on the study see KUMAR *et al.* (2000).

Note that

1. with markers in pseudobackcross configuration, the analysis is equivalent to the analysis of a backcross, and
2. the method of this article, in particular multiple imputation for missing markers, does not apply to selectively genotyped data, where individuals genotyped are selected from the tails of the trait distribution.

Two linkage groups (linkage groups 1 and 3) were chosen for illustration of our method. Linkage group 3 was chosen because it contained statistically significant associations found from previous (non-Bayesian) analyses (KUMAR *et al.* 2000). Linkage group 1 was chosen as simply another linkage group where no significant association had previously been found.

A non-Bayesian analysis is given for comparison. This is based on *t*-tests for the regression coefficient of a marker in a single-marker model for the trait and repeated for each available marker. Missing data were removed (rather than using multiple imputation) before testing each marker. Genome-wise thresholds for the *t*-statistics were estimated using a permutation test (*cf.* CHURCHILL and DOERGE 1994).

Bayesian information criterion: Approximate probabilities for models can be obtained from the BIC, given by

$$\mathrm{BIC} = n\log(1 - R^2) + k\log(n), \tag{1}$$

where k is the number of parameters fitted in the model, n is the number of observations, and R^2 is the proportion of variance explained by the model. SCHWARTZ (1978) showed that with no prior (all models *a priori* equally likely), the posterior probability (p) of a model is approximately proportional to

$$p \propto \exp(-\,\mathrm{BIC}/2). \tag{2}$$

This approach has the advantage of relative simplicity and consequent ease of computation. It has the disadvantage that the relationship (2) between BIC and probabilities of models is *asymptotic*; *i.e.*, it relies on large sample sizes. To allow for this BROMAN (1997) advocates a modification, BIC-δ,

$$\mathrm{BIC}\text{-}\delta = n\log(1 - R^2) + k\delta\log(n), \tag{3}$$

where δ is a constant. Broman recommends $\delta = 2$ or $\delta = 3$; however, the best value to use depends on sample size and other factors and is a topic for future research.

To handle missing values, multiple imputation (RUBIN and SCHENKER 1986) was used to generate multiple instances of the data sets with missing markers randomly estimated according to the values and proximity of flanking markers. This is an alternative to the HALEY and KNOTT (1992) method of assigning a weighted average of flanking marker values to the missing markers. The multiple imputation approach allows for uncertainty in the imputed values.

Ten imputations were used. Rubin and Schenker report good results with 2–3 imputations for estimating means and confidence intervals for continuous random variables. However, since we apply imputation to marker values that take on discrete values (0 or 1) we may need more imputations. As a check on the effectiveness of multiple imputations the model-averaged effects of allelic substitution for outerwood density at markers RAPD.192 and A47.c were reestimated for each of the 10 imputations separately.

Rather than repeating model fits for each imputation, the data from each imputation were combined, giving a data set with $n \times n_I$ points, where n_I denotes the number of imputations used. The multiple imputation estimate of BIC is given by applying (1), with n being the number of observations in the original (unimputed) data and R^2 the value of R^2 from the model fitted to the combined data set. This can be justified by considering the likelihood function for an analysis with each of the n_I imputations of an observation given weight $1/n_I$; *i.e.*, the likelihood contribution for all imputations of a data point is the same as the likelihood contribution for the data point in the unimputed data set, if there is no missing marker, and the average of the likelihood contributions from the various imputations if there is a missing marker.

Calculations for this work used Splus version 3.4 for Unix (Becker *et al.* 1988). The Splus function bicreg.qtl, a modification of the function bicreg (Raftery 1995), was used to search through possible models and calculate the BIC criterion and associated quantities. The search procedure used by bicreg is essentially an exhaustive search using the all subsets regression function leaps, returning the value of R^2, for each model, from which BIC is calculated. Backward elimination is used to reduce the number of variables to the limit of 30 prior to calling leaps. Our function contains modifications to allow for adjustment for multiple imputations, prior distributions, and the parameter δ. The function bicreg gives estimates of average effects for variable *conditional on selection* (*i.e.*, averaged over models in which the variable is selected). We also give unconditional estimates, where the effect of a variable (marker effect) is defined to be zero in models where the marker is not selected, and model-averaged effects of allelic substitution (*i.e.*, estimates of the difference between marker classes). Several options from bicreg can be adjusted to control the amount of computing done. These are Occam's razor constant OR and the lower limit on number of models nbest considered for each model size. Initially at least nbest models of each size are considered, then models less likely than the most likely model by the factor OR or more are eliminated from consideration. The calculations for Tables 2–4 all used OR = 10,000 and nbest = 100.

Incorporating prior probabilities: The BIC criterion does not incorporate prior information. For any given prior, the effect of the prior will be negligible for a sufficiently large sample size. For BIC-δ increasing δ may compensate for a lower prior probability of linkage. We prefer, however, to explicitly incorporate the prior and leave the parameter δ to compensate for the effects of finite sample size on the asymptotic approximation involved in (2).

We now modify the probabilities of (2) by considering the prior probability for a QTL to be present at a given locus. If there are k QTL present, and markers are spaced on average at s cM in a genome of length L cM, then the probability that a QTL is present in the immediate neighborhood of a marker is

$$\pi \approx 1 - \left(1 - \frac{s}{L}\right)^k. \qquad (4)$$

Our markers are spaced at ∼12–15 cM on a genome of estimated length 2000 cM (Wilcox 1997). A genetic architecture with many small QTL seems likely since large effect QTL should have been detected by previous studies (Wilcox *et al.* 1997). If there are 20 QTL we have $k = 20$, $s = 12$, $L = 2000$ cM, giving $\pi \approx 0.1$. To show sensitivity to π, and to accommodate readers who believe there are fewer QTL likely to be present, we also calculate posterior probabilities for model size for $\pi = 0.03$ or 0.06 corresponding to 5 or 10 QTL, respectively, in the analysis below.

Ideally one would like to have a marker at each QTL location. What we *can* guarantee is a marker at each QTL location *to within the resolution of the marker linkage map.* So for each QTL configuration, the "true" model (our best approximation) will be a model such that each marker is selected (in the model) if and only if it is the closest marker to some QTL or, equivalently, there is a QTL in the region around a marker extending halfway to each of the adjacent flanking markers. QTL are assumed to occur randomly in any s-cM interval, with occurrences in various intervals being mutually independent. So the prior probability that each marker is selected is π, and the events of selection or nonselection of the various markers are *a priori* mutually independent. The prior probability that a given model with k markers selected is the true model is therefore

$$\pi^k(1 - \pi)^{(n-k)}. \qquad (5)$$

The combined prior probability for all models with k markers linked to QTL is therefore the binomial probability

$$\frac{q!}{(q - k)!k!}\pi^k(1 - \pi)^{(q-k)}, \qquad (6)$$

where q is the total number or markers being considered.

In calculating the posterior probability of a model the estimate from (2) is multiplied by the probability from (5).

Marginal probabilities of QTL location: The marginal probability that a QTL is in a region is estimated as the marginal probability that one or more of the markers in the region are selected, which is obtained from the BIC calculation by summing posterior probabilities for models containing one or more of the markers.

Models, effects of markers, and effects of allelic sub-

stitution: Let m denote the number of markers and n the number of progeny. We consider all possible models, corresponding to all 2^m possible subsets of markers. Each model is characterized by a set of markers that are *selected.* Let $\mathcal{M}_k$ denote the kth model, let p_k denote the posterior probability of $\mathcal{M}_k$, let $\mathcal{M}_{k(i)}$ denote the model where only marker i is selected, and let S_i denote the set of models with marker i selected.

Let M_j be the jth marker, with alleles labeled 1 and 2, and y_i and $M_{j(i)}$ the trait value and value of the jth marker for the ith individual, respectively. Let $V(M_j)$ (the *vicinity* of marker j), be defined as the set of points closer to marker j than to any other marker.

The models fitted are of the form

$$y_i = \sum_{j=1}^{m} b_{j,k}x_{ij} + e_{ij}, \quad i = 1, \ldots, n, \tag{7}$$

where

$$x_{ij} = \begin{cases} +1/2 & \text{if } M_{j(i)} = 2 \\ -1/2 & \text{if } M_{j(i)} = 1, \end{cases} \tag{8}$$

and the errors e_{ij} are assumed to be normally distributed.

Effects of markers: The regression coefficients for $\mathcal{M}_k$ are denoted by $b_{j,k}$ or simply b_j when there is no need to distinguish models. We refer to $b_{j,k}$ as *the effect of the jth marker in* $\mathcal{M}_k$. The coefficients $b_{j,k}$ for unselected markers are set to zero by convention, so that the sum in (7) is effectively over selected markers.

Effects of allelic substitution: The effect of allelic substitution, $d_{i,k}$, for M_i in $\mathcal{M}_k$ is defined as the difference in population means between the two marker classes if $\mathcal{M}_k$ is the true model.

Note that

1. the effect of allelic substitution $d_{i,k(i)}$ in the model $\mathcal{M}_{k(i)}$ where only marker i is selected and is the same as the conventional effect of allelic substitution, and
2. the effects of allelic substitution $d_{i,k}$ are not the same as the effects of markers $b_{i,k}$, except in the model $\mathcal{M}_{k(i)}$, in which case

$$b_{i,k(i)} = d_{i,k(i)}$$

and the observed difference between the averages of marker classes is an unbiased estimate of both quantities.

Estimation: Let $\hat{b}_{i,k}$, $\hat{d}_{i,k}$ be the conventional maximum-likelihood estimates of estimates of $b_{i,k}$, $d_{i,k}$, respectively, in model $\mathcal{M}_k$.

Let $\hat{b}_{i,s}$ be the estimated effect for the ith marker conditional on selection (*i.e.*, the effect, averaged over models, in which the marker is selected) given by

$$\hat{b}_{i,s} = \frac{\Sigma_{\mathcal{M}_k \in S_i} p_k \hat{b}_{i,k}}{\Sigma_{\mathcal{M}_k \in S_i} p_k}, \tag{9}$$

and let $\hat{b}_{i,av}$, be the unconditional estimate (averaged over all models with the effect set to zero in models where the marker is not selected) of the marker effect for marker i given by

$$\hat{b}_{i,av} = \sum_k p_k \hat{b}_{i,k}. \tag{10}$$

Similarly, let $\hat{d}_{i,av}$ be the model-averaged estimate of allelic substitution for the ith marker given by

$$\hat{d}_{i,av} = \sum_k p_k \hat{d}_{i,k}. \tag{11}$$

Selection bias: Selection bias is a well-known phenomenon that occurs when using a model selection method, such as stepwise regression, to select the variables in a model, and the same data used for model selection are used to estimate the effects. Conditional on selection, the estimates of effects are greater in absolute value than the true values, because in the sampling distribution of effects, the values larger than some threshold are selected. An implicit model selection step is being carried out when selecting markers using the conventional t-test or interval mapping methods.

Our estimate of selection bias is obtained by comparing the model-averaged estimates, which we argue have no problem with selection bias (*cf.* APPENDIX), to corresponding estimates conditional on selection, *i.e.*, estimates averaged over models in which the marker is selected.

Selection bias in the effect of allelic substitution is estimated as

$$\text{selection bias} \approx \frac{\hat{d}_{i,k(i)} - \hat{d}_{i,av}}{\hat{d}_{i,av}} \tag{12}$$

and is given to indicate the bias likely to result from commonly used methods that both select a marker or markers on the basis of some test (effectively selecting a model) and estimate the effect of the marker using the same data. The actual selection bias when using the non-Bayesian methods depends on the threshold used for the test statistic (it is higher with higher threshold or lower P value).

The concepts of this selection and their interpretations are summarized in Table 1. Further discussion of selection bias and a small simulation study are given in the APPENDIX, showing the selection bias in the t-test method, that Bayesian estimates have negative bias (shrinkage toward zero), in the case of model uncertainty, and that even conditional on selection using the t-test at quite low thresholds, the model-averaged estimates did not show selection bias.

RESULTS

Results are given for $\delta = 1, 1.5, 2$, and 3. For simplicity of discussion, unless otherwise stated, all comments below refer to the case $\delta = 1$. Higher values are more conservative, leading to lower probabilities for a QTL

TABLE 1

Notation, terminology, and interpretations for models, effects of markers, effects of allelic substitution, and associated estimates

Symbol	Concept	Practical interpretation
M_i	ith marker	
$V(M_i)$	Vicinity of ith marker	Region closer to ith marker than any other
$\mathcal{M}_k$	kth model	
p_k	Posterior probability of $\mathcal{M}_k$	
—	Marginal probability of selection of markers(s) in a region	Probability of existence of one or more QTL in the region
$\hat{b}_{i,s}$	Estimated effect of marker M_i conditional on selection	Estimate of QTL effect assuming QTL exists in $V(M_i)$
$\hat{b}_{i,av}$	Unconditional (*i.e.*, model averaged) effect of M_i	Posterior expected gain from QTL in $V(M_i)$
$\hat{d}_{i,k(i)}$	Conditional effect of allelic substitution	Estimated gain from selection on M_i assuming QTL exists only in $V(M_i)$
$\hat{d}_{i,av}$	Unconditional (*i.e.*, model averaged) effect of allelic substitution	Posterior expected gain if using M_i for marker-aided selection

on a linkage group, larger amounts of selection bias, and smaller estimates of effects of allelic substitution.

Table 2 shows the marginal probabilities for models of various sizes for two linkage groups, linkage groups 1 and 3. These probabilities are obtained by amalgamating probabilities of all models of each given size.

The marginal probability that one or more QTL are present on a linkage group is the probability that the model size is 1 or more or 1 minus the probability that the model size is zero.

For juvenile wood, from Table 2 with $\delta = 1$, the probability that a QTL is present on linkage groups 1, 3 is 0.17, 0.997, respectively. On linkage group 3 the probability of model size 2 was 0.26, indicating the possibility of two QTL separated by one or more markers. Higher values of δ are more conservative, giving lower probabilities for a QTL; *e.g.*, with $\delta = 2, 3$ the probability that a QTL is present on linkage group 3 is 0.97, 0.76, respectively. Evidence for a QTL, albeit not strong, persists to $\delta = 3$.

For outerwood, from Table 2 with $\delta = 1$, the probability that a QTL is present on linkage groups 1, 3 is 0.38, 0.88, respectively. On linkage group 3 the probability of model size 2 was 0.28, indicating the possibility of two QTL separated by one or more markers. With $\delta \geq 2$ the probability that a QTL is present is <0.5 on each linkage group.

Marginal probabilities for model size for various values of the prior probability $\pi = 0.03, 0.06, 0.1$, corresponding to a prior expectation of 5, 10, 20 QTL, respectively, are shown in Table 3.

Table 4 shows the probability of selection, estimated

TABLE 2

Marginal probabilities for model size estimated using the BIC-δ criterion for various values of δ

	Linkage group 1				Linkage group 3			
k	$\delta = 1$	$\delta = 1.5$	$\delta = 2$	$\delta = 3$	$\delta = 1$	$\delta = 1.5$	$\delta = 2$	$\delta = 3$
	Juvenile wood density (ages 1–5 years)							
0	0.832	0.940	0.979	0.998	0.003	0.010	0.033	0.240
1	0.155	0.058	0.020	0.002	0.710	0.877	0.928	0.757
2	0.013	0.002	<0.001	<0.001	0.263	0.109	0.038	0.003
3	<0.001	<0.001	<0.001	<0.001	0.024	0.003	<0.001	<0.001
	Outerwood density (age 14 years)							
0	0.618	0.840	0.943	0.994	0.123	0.369	0.668	0.953
1	0.359	0.157	0.057	0.006	0.562	0.542	0.316	0.047
2	0.022	0.003	<0.001	<0.001	0.275	0.085	0.016	<0.001
3	<0.001	<0.001	<0.001	<0.001	0.038	0.004	<0.001	<0.001
4	<0.001	<0.001	<0.001	<0.001	0.003	<0.001	<0.001	<0.001

Marginal probabilities for model size (k) for linkage groups 1 and 3, for juvenile wood density (ages 1–5 years), and outerwood density (age 14 years). Probabilities were estimated using BIC-δ, with $\delta = 1, 1.5, 2, 3$, and a prior probability of 0.1 for each marker to be in the model.

TABLE 3

Marginal probabilities for model size estimated using the BIC-δ criterion for various prior probabilities of selection of markers

	Linkage group 1			Linkage group 3		
k	π = 0.03	π = 0.06	π = 0.1	π = 0.03	π = 0.06	π = 0.1
			Juvenile wood density (ages 1–5 years)			
0	0.94	0.90	0.83	0.013	0.006	0.003
1	0.049	0.096	0.154	0.893	0.812	0.710
2	0.001	0.005	0.013	0.092	0.173	0.263
3	<0.001	<0.001	<0.001	0.002	0.009	0.024
			Outerwood density (age 14 years)			
0	0.859	0.743	0.618	0.408	0.226	0.123
1	0.139	0.248	0.359	0.518	0.593	0.562
2	0.002	0.009	0.022	0.070	0.167	0.274
3	<0.001	<0.001	<0.001	0.002	0.013	0.038
4	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Marginal probabilities for model size (*k*) for linkage groups 1 and 3, for juvenile wood density (ages 1–5 years), and outerwood density (age 14 years). Probabilities were estimated using BIC-δ, with δ = 1, and a prior probability of π = 0.03, 0.06, and 0.1 for each marker to be in the model, corresponding to 5, 10, and 20 QTL, respectively.

effects, and standard errors for markers, obtained by combining effects across models according to their probabilities. Effects ($\hat{b}_{i,s}$) are shown conditional on selection (averaged over models, in which the marker is selected, corresponding to estimated QTL effects assuming a QTL is present) or unconditionally ($\hat{b}_{i,av}$, averaged over all models with the effect set to zero in models where the marker is not selected, corresponding to the posterior mean of estimated QTL effects for QTL in the vicinity of marker *i*).

Note that no single marker has a high posterior probability. This reflects uncertainty in the positional location of a QTL. For example, for juvenile wood density the markers RAPD.38 to A297.b2 had probabilities of 12–42%. Outside this region posterior probabilities dropped off to low values. This suggests that a QTL if present is most likely to be in the region between these two markers or possibly in the closer one-half of the adjoining intervals beyond this region. The marginal probability that a QTL is in this region is estimated at 0.995 and contains practically all of the posterior probability of models of nonzero size.

Table 5 gives the conventional *t*-test/LOD score analysis one marker at a time for linkage group 3 plus model-averaged estimates of the effect of allelic substitution and estimates of selection bias in the non-model-averaged estimates of allelic substitution. The conventional estimate of allelic substitution $\hat{d}_{i,k(i)}$ for the *i*th marker is subject to selection bias—it was obtained from the same data that were used to select the marker. The model averaged estimate $\hat{d}_{i,av}$ in Table 5 is not subject to selection bias.

For juvenile wood density, the markers RAPD.38 and A113.a1 had comparison-wise *P* values of ∼0.00001 (genome-wise $P < 0.01$) and markers A329.c3 and A297.b2 had comparison-wise *P* values of ∼0.0001 (genome-wise $P < 0.05$).

For outerwood density, RAPD.192 and A47.c have comparison-wise *P* values of ∼0.002.

Note the selection bias of 27, 25, and 75% for the three largest effect markers for juvenile wood density and 85 and 77% for the two largest effect markers. If higher values of δ are used these estimates will increase.

The calculations for outerwood density for linkage group 3 with 16 markers took ∼9 min in Splus on a Silicon Graphics Indigo Impact 10,000 running Irix 6.2, finding probabilities for a total of 332 models.

Calculations of just the model probabilities in Table 2 with δ = 1 and nbest = 10 took 6 sec.

The effects of allelic substitution for outerwood density at markers RAPD.192 and A47.c were reestimated for each of the 10 imputations separately. The standard deviation between imputations was 0.075, giving a standard error of the mean for 10 imputations of ∼0.02, which is acceptable.

DISCUSSION

The BIC method with multiple imputations for missing values gives estimates of posterior probabilities that can be easily and rapidly calculated for a linkage group. With 10 imputations the standard error of the mean of the effects of allelic substitution estimated separately for each imputation was only about one-eighth of the standard error of the non-model-averaged estimate of the effect of allelic substitution. Therefore, more imputations would not significantly decrease the error of the model-averaged estimate.

TABLE 4

Effects of markers and probabilities of selection for linkage group 3

			Juvenile wood density (ages 1–5 years)					Outerwood density (age 14 years)				
				Conditional		Unconditional			Conditional		Unconditional	
i	Marker	Position (cM)	Probability selected (%)	$\hat{b}_{i,s}$	SE$(\hat{b}_{i,s})$	$\hat{b}_{i,av}$	SE$(\hat{b}_{i,av})$	Probability selected (%)	$b_{i,s}$	SE$(\hat{b}_{i,s})$	$\hat{b}_{i,av}$	SE$(\hat{b}_{i,av})$
1	A93.b3	0	1.7	4.1	1.7	0.07	0.58	0.9	0.05	0.05	<0.001	0.007
2	A56.A	15	2.4	−5.2	1.6	−0.12	0.82	2.2	−0.19	0.05	−0.004	0.029
3	A184.b3	25	1.0	−0.2	1.6	−0.00	0.16	4.0	0.24	0.06	0.010	0.049
4	RAPD.59	48	1.5	−3.7	1.6	−0.05	0.48	3.5	−0.24	0.06	−0.009	0.046
5	RAPD.38	66	19.6	14.3	3.6	2.80	5.89	2.0	0.18	0.07	0.003	0.027
6	A329.c3	83	35.4	16.7	2.7	5.92	8.16	11.4	0.32	0.05	0.037	0.106
7	A113.a1	91	42.1	17.5	2.3	7.35	8.76	6.3	0.29	0.07	0.018	0.071
8	A297.b2	99	12.4	14.5	4.0	1.79	4.96	3.1	0.24	0.09	0.007	0.044
9	RAPD.270	106	1.2	1.8	2.6	0.02	0.35	4.8	0.27	0.06	0.013	0.059
10	A338.A1	115	1.3	−2.4	3.5	−0.03	0.49	4.8	−0.30	0.09	−0.014	0.067
11	A219.b2	123	2.4	5.7	2.3	0.14	0.96	2.3	−0.21	0.12	−0.005	0.037
12	A140.C2	130	1.6	−3.9	2.5	−0.06	0.58	7.6	−0.35	0.09	−0.026	0.095
13	RAPD.192	140	3.6	6.4	1.5	0.23	1.22	23.3	0.43	0.07	0.100	0.184
14	A47.c	153	2.3	5.2	1.6	0.12	0.82	39.8	0.44	0.06	0.177	0.220
15	RAPD.209	170	1.4	−2.9	1.5	−0.04	0.38	6.2	−0.32	0.08	−0.020	0.080
16	A72.A	187	1.0	−0.6	1.6	−0.00	0.18	1.2	−0.10	0.10	−0.001	0.015

Probability of selection, estimated effects, and standard errors of effects for markers from linkage group 3 for juvenile wood density (ages 1–5 years) and outerwood density (age 14 years) in *kg*/*m*^{3}. Effects and standard errors are shown conditional on selection ($\hat{b}_{i,s}$, columns 3 and 4) and unconditionally ($\hat{b}_i$, columns 5 and 6). Estimates were obtained using BIC-δ, with $\delta = 1$ and a prior probability of $\pi = 0.1$ for each marker to be in the model.

TABLE 5

Single-marker *t*-test/LOD score analysis and selection bias for linkage group 3

		Juvenile wood density (ages 1–5 years)						Outerwood density (age 14 years)					
i	Marker	$\hat{d}_{i,k(i)}$	$\mathrm{SE}(\hat{d}_{i,k(i)})$	P	LOD	$\hat{d}_{i,av}$	Selection bias (%)	$\hat{d}_{i,k(i)}$	$\mathrm{SE}(\hat{d}_{i,k(i)})$	P	LOD	$\hat{d}_{i,av}$	Selection bias (%)
1	A93.b3	5.4	4.7	0.263	0.3	1.3	315	0.10	0.16	0.525	0.1	0.03	233
2	A56.A	−8.0	4.5	0.081	0.7	−4.1	95	−0.25	0.15	0.101	0.6	−0.07	257
3	A184.b3	5.9	5.2	0.262	0.3	4.3	37	0.26	0.16	0.094	0.6	−0.05	−620
4	RAPD.59	−13.7	4.7	0.005	1.8	−5.0	174	−0.40	0.16	0.014	1.4	−0.09	344
5	RAPD.38	21.0	4.4	1.1×10^{-5}	4.3	12.0	75	0.35	0.17	0.039	1.0	0.09	289
6	A329.c3	18.6	4.6	1.2×10^{-4}	3.3	14.9	25	0.26	0.16	0.103	0.6	0.11	136
7	A113.a1	20.0	4.2	8.6×10^{-6}	4.4	15.8	27	0.35	0.15	0.022	1.2	0.12	192
8	A297.b2	17.4	4.3	1.3×10^{-4}	3.3	13.9	25	0.31	0.15	0.046	0.9	0.15	107
9	RAPD.270	10.7	5.0	0.038	1.0	8.7	23	0.40	0.19	0.036	1.0	0.11	264
10	A338.A1	−11.1	4.3	0.013	1.4	−9.1	22	−0.37	0.15	0.013	1.4	−0.18	106
11	A219.b2	11.7	4.6	0.014	1.4	6.8	72	0.21	0.15	0.172	0.4	0.16	31
12	A140.C2	−10.3	4.5	0.026	1.1	−6.9	49	−0.44	0.14	0.003	2.0	−0.21	110
13	RAPD.192	10.9	5.0	0.032	1.0	4.1	166	0.46	0.16	0.004	1.8	0.26	77
14	A47.c	8.7	4.8	0.075	0.7	3.1	181	0.50	0.15	0.002	2.2	0.27	85
15	RAPD.209	−7.2	4.4	0.104	0.6	−1.5	380	−0.42	0.16	0.010	1.5	−0.19	121
16	A72.A	0.9	4.6	0.838	0.0	1.2	−25	−0.18	0.15	0.231	0.3	−0.11	63

Single marker *t*-test/LOD score analysis for linkage group 3, model-averaged effect of allelic substitution, and selection bias for juvenile wood density (ages 1–5 years) and outerwood density (age 14 years) in *kg/m*^{3}. The columns $\hat{d}_{i,k(i)}$, $\mathrm{SE}(\hat{d}_{i,k(i)})$, P, LOD, and $\hat{d}_{i,av}$ denote the conventional (*i.e.*, non-model-averaged estimate, estimated in the model with marker i only selected) estimate of allelic substitution, its standard error, the P value for the null hypothesis of no QTL effect, LOD score, and a model-averaged effect of allelic substitution for the ith marker, respectively. Model averaging is based on $\delta = 1$ and prior probability of $\pi = 0.1$ for a QTL to be in the neighborhood of a marker. Thresholds for P corresponding to $P < 0.05$ and $P < 0.01$ genome-wise are, respectively, 4.4×10^{-4} and 9.5×10^{-5} for juvenile wood and 4.1×10^{-4} and 7.8×10^{-5} for outerwood.

For juvenile wood density the posterior probability that there is a QTL on linkage group 1 is ∼0.17. The posterior probability that there is a QTL is high for linkage group 3 with a posterior probability >0.95 for $\delta \leq 2$.

The putative QTL for juvenile wood density on linkage group 3 was (or were) located in the region between or adjoining the markers RAPD.38 and A297.b2 with probability 0.995. By comparison, Kumar *et al.* (2000), applying the bootstrap method of Visscher *et al.* (1996), obtained a 95% confidence interval of 56–96 cM, the region from approximately midway between RAPD.59 and RAPD.38 to just before marker A297.b2. This is comparable to our result, although a 95% confidence interval is not the same as a 95% posterior interval, and the posterior probabilities decrease rapidly as one moves beyond this interval so that a 95% probability interval is not much different in size to a 99% or higher probability interval.

For outerwood density, the posterior probability that there is a QTL on linkage group 1 is ∼0.38 and on linkage group 3 the probability is ∼0.88. This is evidence against a QTL on linkage group 1 and evidence for a QTL on linkage group 3. The evidence is not strong, however, so we should not be surprised, if, as was the case, the QTL association was not subsequently verified on linkage group 3. Nor does the evidence rule out the existence of a small undetected QTL on linkage group 1. A larger sample size is recommended.

Although no single marker for outerwood density attained the experiment-wise level of $P < 0.05$, Kumar *et al.* (2000) obtained a P value of 0.002 (experiment-wise $P < 0.05$) for the F-test when seven equally spaced markers from linkage group 3 were jointly regressed on outerwood density. Their experiment-wise P value of just <0.05 can be compared to our marginal probability for model size greater than zero of 0.88 for linkage group 3 with $\delta = 1$. The experiment-wise P value is about one-half the posterior probability of model size zero in this case. If $\delta = 2$ or $\delta = 1$ and $\pi = 0.03$ corresponding to a prior expectation of only five QTL then the experiment-wise P value is about one-third or one-eighth of the posterior probability of model size zero.

These results demonstrate the difference between the results of the Bayesian approach to QTL mapping and the non-Bayesian or frequentist approaches, where, to the naive user, the evidence in the form of P values appears stronger. If using P values, controlling for multiple comparisons is certainly necessary in this case—the comparison-wise P values were orders of magnitude less than the probability of model size zero. The experiment-wise P values were somewhat less than but of the same order of magnitude as the probabilities of model size zero in this example.

As has been demonstrated in other applications (see, *e.g.*, Berger and Berry 1988), the evidence for a QTL can be weaker than appears to be the case with P values: A genome-wise significance level of $\alpha = 0.05$ (comparison-wise 4.4×10^{-4}) can correspond to only weak evidence for a real effect. This is to be expected because the strength of evidence implied by a given P value decreases with sample size. This occurs because the P value is measuring evidence that H_0 is not the true model under which the data are generated. In practice, any model is only an approximation for the process generating the observed data, and hence as the sample size gets large the probability of observing the data under H_0 tends to zero. The problem is that the probability of observing the data under the alternative hypothesis H_1 of a real effect also tends to zero (for the same reason) and may be equally small; *i.e.*, the data do not favor H_1 over H_0 just because the P value is small. Therefore, with larger sample sizes the differences between the P value and the posterior probability of H_0 are likely to increase.

The problem remains with approaches using P values, whether comparison-wise, chromosome-wise, or experiment-wise: what threshold to use and how to interpret the results. Is the evidence strong, fair, or weak if we get an experiment-wise P value or 0.05 or 0.01? There is no relation between the experiment-wise P value and posterior probabilities that is independent of sample size and problem setup.

The 2 linkage groups analyzed here were preselected from 23 linkage groups. For the Bayesian approaches this poses no problem—the results of a Bayesian analysis depend only on the data analyzed and not other information such as what other data had been, or might have been, or will be analyzed. This avoids complexities such as whether to use comparison-wise, genome-wise, or experiment-wise thresholds for QTL detection, and the difficulties of interpretation and use of any of these quantities for decisions.

Using a high threshold reduces the number of false positives but also reduces the probability of detection of real QTL. To be "wrong" only 5% of the time when there is no effect may be comforting for an experimenter, but this is no comfort to decision makers who may be presented with only the 5% of "significant" results, which could well be all wrong. Decision makers need to know the posterior probability of presence of a QTL in a region *vs.* the cost of using more markers or carrying out further, or larger, QTL mapping experiments. Decision makers also need unbiased estimates of effects of allelic substitution at a marker putatively associated to a QTL. Our best estimate, given the data and prior knowledge, is one-half the model-averaged effect of allelic substitution. This is unbiased in the sense that the model-averaged effect is the posterior mean for the effect. To obtain unbiased estimates with the non-Bayesian QTL mapping methods requires separate data for selection (QTL "detection") and estimation (or "verification"). For these reasons we recommend readers adopt the Bayesian approach.

There are two major differences between our approach and non-Bayesian methods—considering the prior probability for a QTL to be present and the use of multiple models. These are important to avoid exaggerating the evidence for a QTL, for estimating the gain from marker-aided selection, and for avoiding the problems with selection bias (Miller 1990). Selection bias is shown to be a problem that cannot be ignored for the data of this article. The method of this article explains, and asymptotically (to within the accuracy of the estimates of probabilities based on BIC) overcomes, the problems of selection bias and QTL being frequently detected but not verified (see, *e.g.*, Beavis 1994; Wilcox *et al.* 1997; Melchinger *et al.* 1998).

We compared the proposed Bayesian approach with standard non-Bayesian QTL mapping methods, as commonly used. In the context of non-Bayesian QTL mapping, other techniques have been suggested such as cross-validation (Utz *et al.* 2000) and bootstrapping (Beavis 1994; Visscher *et al.* 1996), which could potentially be used to ameliorate problems of selection bias.

Cross-validation is a technique where the analysis is repeated with various disjoint subsets of the data left out and results combined. Cross-validation is generally used, in the context of a single model, to obtain an estimate of prediction error. Utz *et al.* (2000) point out that estimates of QTL effects are often inflated (we suggest mainly because of selection bias). They use cross-validation to obtain unbiased estimates of the magnitude of QTL effects. However, they had already eliminated selection bias prior to cross-validation, by using "test data sets" for the cross-validation separate from their "estimation data" that were used to select the markers. A common problem with cross-validation is the inaccuracy in cross-validation estimates of error. To use cross-validation with a method such as interval mapping, which selects models, the cross-validation subsets would have to be chosen to be sufficiently large to result in a range of different models being selected. Therefore the overall sample size would have to be large to give a reasonable number of cross validations. Bootstrap bagging (Breiman 1996) or boosting (Freund 1995; Freund and Schapire 1996) may be better. See, *e.g.*, Dudoit *et al.* (2000) for an application to microarray data analysis. These methods should be more robust than single-model methods and may give acceptable results for some purposes—if one is interested solely in a "black box" type of model for prediction only and not inferences about individual loci. However, they will be effective in reducing selection bias only to the extent that they effectively average over a set of possible models approximately proportional to their probabilities. Note that bootstrapping (Efron 1982) was invented for use as a general method; however, considerable sophistication is needed to rigorously justify applications of the bootstrap to other than the simplest setups. For a discussion see Young (1994).

In the Bayesian context an alternative approach to analyzing and averaging over multiple models according to their probabilities is to fit one large model with all possible predictors, with the appropriate prior correlation structure. Determining the "appropriate" correlation structure is the difficulty with this approach. A generic method that effectively does this is ridge regression, where predictors are shrunk toward zero, predictors with less support from the data being shrunk more. Ridge regression has a Bayesian interpretation (Frank and Friedman 1993) corresponding to a uniform prior distribution on directions in the vector space spanned by predictors. The ridge parameter controlling the overall shrinkage can be chosen by cross-validation (see, *e.g.*, Ball *et al.* 1998, for an example relating chemical analysis to sensory perception).

We prefer the approach of this article over the above alternatives on philosophical grounds because our prior relates more naturally to prior expectations about the number of QTL than the Bayesian alternative and follows logically from the natural prior using probability theory, *i.e.*, does not involve the *ad hoc*-ery of the frequentist alternatives. Assuming our prior distributions are a reasonable representation of our prior knowledge, Bayesian theory guarantees the full Bayesian approach is optimal.

The probabilities calculated in this article depend on the values of the parameter δ and the value of the prior probability $\pi = 0.1$. Higher values of δ correspond to lower probabilities for presence of a QTL and higher selection bias. The adjustment factor δ was proposed by Broman (1997) to correct for the finite sample size, which may not be large enough to rely on the asymptotic approximation in (2). We expect the appropriate value to use depends on sample size, with $\delta = 1$ being the appropriate choice for large sample size, $\delta = 2$ being fairly conservative. Broman recommends $\delta = 2$ or $\delta = 3$. However, since we, unlike Broman, adjust for the prior, we may not need as high values for δ as Broman recommends.

The appropriate value(s) of δ could be determined by comparison with simulation consistent estimates for the hierarchical model obtained from a Markov chain Monte Carlo (MCMC) method. Another possible approach is to compare results with analytical calculations of Bayes factors (*cf.* Smith and Kohn 1996) for one or more single-marker models compared with the null model, with a suitable prior on the size of marker effects.

The reader can and should try other values of π. The marginal probabilities for model sizes in Table 2 can be adjusted for different values of π using (6). More generally a continuous prior distribution for π, rather than a single value as we have used here, can be approximated by combining results from several different values of π.

There are several options available to cut down on computing time:

1. Reduce the number of markers analyzed. For example, if seven markers spaced at ~20 cM are analyzed the time is reduced to 1 min.
2. Cut down on the number of models to be found for each model size (option nbest of bicreg.qtl) and/or limit the largest size of models considered.
3. Reduce the Occam's razor threshold (option OR of bicreg). Models less likely than the most likely model by the factor OR or more are eliminated from consideration.
4. Program the method in a lower level language such as C.

The method of this article can be applied one chromosome or linkage group at a time, with a substantial reduction in computation. Since linkage groups are independent, this makes little difference to inference and reduces the number of models that need to be considered to a reasonable number. If enough QTL have been found (by an initial iteration of the method) to significantly reduce the error variance, then it will be advantageous to add covariates to control for the QTL from other linkage groups to the models for the linkage group being considered. The covariates would always be selected: *i.e.*, models being compared would consist of covariates plus selected subsets of markers in the linkage group being considered.

The marker density required depends on the sample size and the accuracy with which it is desired to estimate QTL location. The latter is, however, limited by sample size. A dense marker map is not required. We would not recommend having much more than about five markers within a region in which a QTL could be located with probability 95%.

A future direction for development of QTL mapping methods based on model selection is to allow for finer structure mapping, using grids of points between markers. Marker values at these grids would be regarded as missing data, which, if known, would allow the application of the method. Missing marker values could be estimated using multiple imputation or (preferably) as additional parameters in an MCMC algorithm. Our current marker density is adequate, given the sample size and consequent uncertainty in QTL location.

For outbred crosses, considering one parent at a time, markers are either segregating in pseudobackcross configuration (as analyzed in this article) for the parent or not segregating, in which case they contribute no information. The method as described here will find additive effects of loci in each parent separately. The analysis may be extended to the general case by combining loci from the two parents and allowing for interactions (corresponding to dominance and epistasis) between loci.

The author thanks Phillip Wilcox and Rowland Burdon for valuable feedback on various versions of the manuscript from the perspective of molecular and quantitative geneticists; Mark Kimberley and the referees for useful comments on the manuscript; John Lee for the phenotypic data; and Geoff Corbett and Paul Fisher for the genotypic data. Comments from Phillip Wilcox, Gary Churchill, and a referee have resulted in a substantially expanded discussion and justification of the results on selection bias. This work was funded by the New Zealand Foundation for Research Science and Technology.

LITERATURE CITED

Ball, R. D., S. H. Murray, H. Young and J. M. Gilbert, 1998 Statistical analysis relating analytical and consumer panel assessments of kiwifruit flavour compounds in a model juice base. Food Quality Pref. **9**(4): 255–266.

Beavis, W. D., 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies. Proceedings of the 49th Annual Corn and Sorghum Industry Research Conference. American Seed Trade Assoc., Washington, DC, pp. 250–266.

Becker, R. A., J. M. Chambers and A. R. Wilks, 1988 *The New S Language, a Programming Environment for Data Analysis and Graphics.* Wadsworth & Brooks/Cole Advance Books & Software, Pacific Grove, CA.

Berger, J., and D. Berry, 1988 Statistical analysis and the illusion of objectivity. Am. Scientist **76:** 159–165.

Breiman, L., 1996 Bagging predictors. Mach. Learning **24**(2): 123–140.

Broman, K. W., 1997 Identifying quantitative trait loci in experimental crosses. Ph.D. Thesis, University of California, Berkeley, CA.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Cown, D. J., and B. C. Clement, 1983 A wood densitometer using direct scanning with x-rays. Wood Sci. Tech. **17:** 91–99.

Doerge, R. W., Z-B. Zeng and B. S. Weir, 1997 Statistical issues in the search for genes affecting quantitative traits in experimental populations. Stat. Sci. **12**(3): 195–219.

Dudoit, S., J. Fridlyand and T. Speed, 2000 Comparison of discrimination methods for the classification of tumors using gene expression data. Technical report 576, Department of Statistics, University of California, Berkeley, CA. http://www.stat.berkeley.edu/users/terry/zarray/Html/discr.html

Efron, B., 1982 *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

Frank, I. E., and J. H. Friedman, 1993 A statistical view of some chemometrics regression tools. Technometrics **35:** 109–135.

Freund, Y., 1995 Boosting a weak learning algorithm by majority. Inf. Comput. **121**(2): 256–285.

Freund, Y., and R. E. Schapire, 1996 Experiments with a new boosting algorithm, pp. 148–156 in *Proceedings of the 13th International conference on Machine Learning.* Morgan Kaufmann, San Francisco.

George, E. I., and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**(423): 881–889.

Green, P. J., 1995 Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Haley, S. C., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Hoeschele, I., P. Uimari, F. E. Grignola, Q. Zhang and K. M. Gage, 1997 Advances in statistical methods to map outbred populations. Genetics **147:** 1445–1457.

Kumar, S., R. J. Spelman, D. J. Garrick, T. E. Richardson, M. Lausberg *et al.*, 2000 Multiple marker mapping of wood density loci in an outbred pedigree of radiata pine. Theor. Appl. Genet. **100:** 926–933.

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

Melchinger, A. E., H. F. Utz and C. C. Schön, 1998 Quantitative trait locus (QTL) mapping using different testers and independent population samples in Maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics **149:** 383–403.

Miller, A. J., 1990 *Subset Selection in Regression* (Monographs on Statistics and Applied Probability 40), Chapman & Hall, London.

Paterson, A. H., 1995 Molecular dissection of quantitative traits: progress and prospects. Genome Res. **5:** 321–333.

Raftery, A. E., 1995 Bayesian model selection in social research (with discussion), pp. 111–196 in *Sociological Methodology*, edited by P. V. Marsden. Blackwell, Cambridge, MA.

Raftery, A. E., D. Madigan and J. A. Hoeting, 1997 Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. **92** (437): 179–191.

Rubin, D. B., and N. Schenker, 1986 Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J. Am. Stat. Assoc. **81:** 366–374.

Satagopan, J. M., and B. S. Yandell, 1996 Estimating the number of quantitative trait loci by Bayesian model determination. Special contributed paper session on genetic analysis of quantitative traits and complex diseases, Biometrics Session, Joint Statistical Meetings, Chicago (ftp://ftp.stat.wisc.edu/pub/yandell/reujump.html).

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.

Schwartz, G., 1978 Estimating the dimension of a model. Ann. Stat. **6**(2): 461–464.

Sillanpää, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics **148:** 1373–1388.

Sillanpää, M. J., and E. Arjas, 1999 Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics **151:** 1605–1619.

Smith, M., and R. Kohn, 1996 Nonparametric regression using Bayesian variable selection. J. Econ. **75**(2): 317–343.

Stephens, D. A., and R. D. Fisch, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics **54**(4): 1334–1347.

Utz, H. F., A. E. Melchinger and C. C. Schön, 2000 Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. Genetics **154:** 1839–1849.

Visscher, P. M., R. Thompson and C. S. Haley, 1996 Confidence intervals in QTL mapping by bootstrapping. Genetics **143:** 1013–1020.

Wilcox, P. L., 1997 Linkage groups and map length in Pinus radiata. p. 79 in *Proceedings of IUFRO '97: Genetics of Radiata Pine*. Dec 1–5, 1997, edited by R. D. Burdon and J. M. Moore. FRI Bulletin No. 203, Rotorua, New Zealand.

Wilcox, P. L., T. E. Richardson and S. D. Carson, 1997 Nature of quantitative trait variation in *Pinus radiata*: insights from QTL detection experiments, pp. 304–312 in *Proceedings of IUFRO '97: Genetics of Radiata Pine*. Dec 1–5, 1997, edited by R. D. Burdon and J. M. Moore. FRI Bulletin No. 203, Rotorua, New Zealand.

Young, G. A., 1994 Bootstrap: more than a stab in the dark? (with discussion). Stat. Sci. **9:** 382–415.

Communicating editor: G. A. Churchill

APPENDIX: SELECTION BIAS AND MODEL AVERAGING

We explain why there is no problem with selection bias in model-averaged estimates of allelic substitution, derive an approximate relationship between selection bias and uncertainty in QTL location, then give the results of a small simulation study comparing selection bias from the *t*-test method using several thresholds to results from model averaging, which are given both unconditionally and conditional on selection for each threshold.

Discussion of selection bias and model averaging: The model-averaged estimate $\hat{d}_{i,av}$ of the effect of allelic substitution d_i for marker *i* is the same as the expectation d_i under the posterior distribution of the overarching hierarchical model, as can be seen by a straightforward calculation,

$$\hat{d}_{i,av} = \sum_k p_k d_{i,k}$$

$$= \int d_i(\theta_k) \ \sum p_k f_k(\theta_k|\text{data})\, d\theta_i \qquad \text{(A1)}$$

$$= E_f(d), \qquad \text{(A2)}$$

where $\mathcal{M}_k$ are the various models, p_k their posterior probabilities, θ_k and $f_k(\cdot)$ denote the parameters and posterior density function of model $\mathcal{M}_k$, $d_i(\theta_k)$ is the value of d_i as a function of θ_k, and f denotes the combined posterior distribution, whose density is given by the second term in (A1).

Hence, the model-averaged estimate is the mean of the marginal posterior distribution for $d_{i,av}$. The marginal posterior distribution represents our knowledge of $d_{i,av}$, and according to standard Bayesian theory, optimal decisions are obtained by minimizing the expected loss (or maximizing the expected gain) over this distribution. Thus the model-averaged estimate is our best estimate, to which we compare estimates conditional on selection.

Note: This argument applies to any quantity that can be calculated as a function of the parameters in any model and has a well-defined interpretation.

Relationship between selection bias and uncertainty in QTL location: To better understand selection bias in the context of hypothesis test approaches to QTL mapping, we consider what happens to the estimates of allelic substitution at marker *i* under various single-marker models and the null model and derive an approximate relationship between selection bias in the effect of allelic substitution at a marker and uncertainty in the location of a QTL.

Let $\mathcal{M}_{k(i)}$ denote the single-marker model with marker *i* selected, $\mathcal{M}_0$ be the null model with no marker selected, r_{ij} be the recombination distance between markers *i* and *j*, and δr be the average intermarker spacing.

If the true model is known to be $\mathcal{M}_{k(i)}$ then the estimate of the effect of allelic substitution d_i under $\mathcal{M}_{k(i)}$ is the standard least-squares estimate, which is unbiased.

Suppose the true model is $\mathcal{M}_{k(j)}$ but $\mathcal{M}_{k(i)}$ has been selected [in preference to $\mathcal{M}_{k(j)}$ and other markers] by a single-marker hypothesis testing procedure. Then the estimate $\hat{d}_i$ of d_i under $\mathcal{M}_{k(i)}$ is greater than the estimate $\hat{d}_j$ of d_j under $\mathcal{M}_{k(j)}$. The latter estimate is unbiased so $\hat{d}_i$ is likely to be comparable to or greater than d_j.

Since $\mathcal{M}_{k(j)}$ is the true model, d_j is the QTL effect [up to a factor of $(1 - \delta r/2)$].

Then the true effect of allelic substitution at marker *i* is

TABLE A1

Selection bias in estimation of effects of allelic substitution, for single marker *t*-test analysis and model averaging using BIC-δ

	$h_Q^2 = 0.05,\ a = 0.45$			$h_Q^2 = 0.20,\ a = 0.89$		
Threshold	$P = 0.01$	$P = 0.001$	$P = 0.0001$	$P = 0.01$	$P = 0.001$	$P = 0.0001$
t-test	0.265	0.377	0.444	0.096	0.127	0.160
BIC-1 (no selection)	−0.098	−0.098	−0.098	−0.199	−0.199	−0.1997
BIC-1 (selected according to *t*-test)	−0.254	−0.294	−0.328	−0.567	−0.631	−0.6726

Bias in estimation of effects of allelic substitution in units of one phenotypic standard deviation from 2000 simulations, each with 100 progeny with six markers on a single chromosome of length 120 cM, is shown. A QTL was present with probability 0.53, and if present had QTL heritability $h_Q^2 = 0.05$ or $h_Q^2 = 0.2$, is given for the *t*-test, model averaging using BIC, and model averaging conditional on selection by the *t*-test. Selection was for various thresholds corresponding to $P = 0.01$, 0.001, and 0.0001.

$$d_i = d_{i,j} = d_j \times (1 - 2r_{ij}).$$

Thus selection bias in $\hat{d}_i$ is $\sim 1/(1 - 2r_{ij})$. Furthermore, assuming model $\mathcal{M}_{k(i)}$ we estimate

$$\hat{d}_j = \hat{d}_i \times (1 - 2r_{ij}).$$

Thus assuming $\mathcal{M}_{k(i)}$ when $\mathcal{M}_{k(j)}$ is the true model induces a relative bias of the order of $(1 - 2r_{ij})^2$ in the ratio of $\hat{d}_i/\hat{d}_j$.

If the model selected is the null model $\mathcal{M}_0$ with no QTL, the estimated effect is zero.

Selection bias and model averaging—a simulation study: For each of $h_Q^2 = 0.05$ and $h_Q^2 = 0.2$, 2000 simulated data sets were generated each with 100 backcross progeny, with 0 (probability 0.53) or 1 QTL explaining proportion h_Q^2 of total variance, randomly placed on a chromosome of length 120 cM with six markers evenly spaced at positions 10, 30, 50, 70, 90, and 110 cM. The size of the QTL effect if present always had a positive sign, corresponding to a QTL effect of $a = 0.45$ (or $a = 0.89$ in units of 1 phenotypic standard deviation). The analysis used model averaging with BIC-δ, with $\delta = 1$ and prior probability $\pi = 0.1$ for a marker to be selected.

Table A1 gives bias in units of 1 phenotypic standard deviation for the single-marker *t*-test and model averaging using BIC.

The fourth row of Table A1 is the average "bias" using model averaging with BIC-1 ($\delta = 1$). A negative bias (shrinkage toward zero) is expected with the Bayesian method if there is any model uncertainty.

The fifth row of Table A1 is the average bias using model averaging conditional on markers being selected by the *t*-test method (not something we would necessarily advocate). Interestingly, while the bias for the *t*-test method as used increases with selection intensity, as expected, the bias for BIC actually decreases (increases in magnitude) with selection intensity; *i.e.*, model averaging is more than compensating for selection of significant markers only. This may be an artifact of the fact that as the threshold increases, the simulated QTL, if selected, is more likely to be at or near the marker under consideration, so the effect is larger; hence the potential shrinkage toward zero is larger.

Selection bias and model averaging—summary:

1. Selection bias occurs when the same data are used to select a regression model and to estimate the coefficients (here marker effects) in the model.
2. Selection bias occurs because expected values of estimates of the effect of allelic substitution at a marker are always greater under the model with that marker selected than under any other single-marker model or the null model.
3. Estimates of allelic substitution at a marker are unbiased, if the true model is known, or selected with independent data.
4. Selection bias is not a problem if Bayesian model averaging is used. The model-averaged Bayesian estimates are not unbiased under any assumed QTL configuration but are the average of unbiased estimates under various models, averaged according to the posterior probability that each model is the true model. From observation (2) it follows that model-averaging estimates of QTL effects are shrunk toward zero. The amount of shrinkage reduces with the precision with which the QTL location can be determined.
5. The Bayesian method can be applied even if a previous hypothesis test or tests have selected a chromosome or region (as was the case for linkage group 3 in this study). This is because the posterior probability distribution for that chromosome or region depends only on the data through the likelihood function and the prior probability for that chromosome or region.
6. Model averaging can overcome selection bias even if markers are selected using non-Bayesian tests.