Theor Appl Genet (2007) 114:1437–1449
DOI 10.1007/s00122-007-0529-x

ORIGINAL PAPER

Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms

Hiroyoshi Iwata · Yusaku Uga · Yosuke Yoshioka · Kaworu Ebana · Takeshi Hayashi

Received: 6 August 2006 / Accepted: 16 February 2007 / Published online: 14 March 2007

Abstract One way to use a crop germplasm collection directly to map QTLs without using line-crossing experiments is the whole genome association mapping. A major problem with association mapping is the presence of population structure, which can lead to both false positives and failure to detect genuine associations (i.e., false negatives). Particularly in highly selfing species such as Asian cultivated rice, high levels of population structure are expected and therefore the efficiency of association mapping remains almost unknown. Here, we propose an approach that combines a Bayesian method for mapping multiple QTLs with a regression method that directly incorporates estimates of population structure. That is, the effects due to both multiple QTLs and population structure were included in our statistical model. We evaluated the efficiency of our approach in simulated- and real-trait analyses of a rice germplasm collection. Simulation analyses based on real marker data showed that our model could suppress both false-positive and false-negative rates and the error of estimation of genetic effects over single QTL models, indicating that our model has statistically desirable attributes over single QTL models. As real traits, we analyzed the size and shape of milled rice grains and found significant markers that may be linked to QTLs reported previously. Association mapping should have good prospects in highly selfing species such as rice if proper methods are adopted. Our approach will be useful for the whole genome association mapping of various selfing crop species.

Communicated by J.-L. Jannink.

H. Iwata (✉)
Data Mining and Grid Research Team,
National Agricultural Research Center,
3-1-1 Kannondai, Tsukuba,
Ibaraki 305-8666, Japan
e-mail: iwatah@affrc.go.jp

Y. Uga · K. Ebana
QTL Genomics Research Center,
National Institute of Agrobiological Sciences,
2-1-2 Kannondai, Tsukuba,
Ibaraki 305-8602, Japan

Y. Yoshioka
Graduate School of Life and Environmental Sciences,
University of Tsukuba, 1-1-1 Tennohdai,
Tsukuba, Ibaraki 305-8572, Japan

T. Hayashi
Laboratory of Animal Genome,
National Institute of Agrobiological Sciences,
2 Ikenodai, Tsukuba, Ibaraki 305-0901, Japan

Introduction

To enable efficient management and utilization of crop genetic resources, systematic germplasm collections, such as core collections (Brown 1989), have been developed and maintained in various crop species. Such collections aim to capture most of the genetic variability in entire genetic resources and thus harbor large amounts of genetic variation. Although such variations in agronomic traits as well as morphological and physiological ones have been evaluated and accumulated for these collections, the quantitative trait loci (QTLs) responsible for the variations are largely unexplored, mainly because of a lack of appropriate statistical methods. QTL mapping based on line-crossing experiments may be one of the most efficient methods of exploring QTLs. To harvest entire variations in collections by this method, however, one might need to assess a large number of segregating families that would include most of

the accessions in the collection as parents. In practice, this would be quite difficult in terms of space, time, and funds.

One way to use accessions in a collection directly to map QTLs without using line-crossing experiments is association mapping, or linkage-disequilibrium (LD) mapping. Association mapping has been used to dissect the genetic basis of human disease (e.g. Kerem et al. 1989; Corder et al. 1994), and has recently been extended to plants (Thornsberry et al. 2001; Parisseaux and Bernardo 2004; Kraakman et al. 2004; Zhang et al. 2005a, b; Yu et al. 2006; Breseghello and Sorrels 2006). One major obstacle in applying association mapping to crop species is that the complex breeding histories of many important crops have created complex population structures within the germplasm (Flint-Garcia et al. 2003). The presence of population structure and unequal distribution of alleles within subpopulations can result in nonfunctional, spurious associations between a phenotype and unlinked candidate gene (Knowler et al. 1988; Lander and Schork 1994). To deal with this problem, several methods have been proposed. Pritchard et al. (2000b) proposed a method of testing association that conditions on the inferred ancestries of individuals. Ancestries were inferred by a Bayesian method proposed by Pritchard et al. (2000a). Thornsberry et al. (2001) extended this method to deal with a quantitative trait, and studied a candidate gene for the control of flowering time in maize. Recently, Yu et al. (2006) proposed a mixed-linear-model method in which effects caused by population structure and background polygenic effects are included as independent variables, and demonstrated that their method can control both false-positive and false-negative rates.

The methods described above can deal with the effect of population structure, but they do not simultaneously take multiple QTLs into account. For a complex trait governed by multiple QTLs, it is reasonable to include the effects of multiple QTLs in the model in order to correctly estimate the number, locations, and genetic effects of QTLs. Recently, Bayesian methods based on the Markov chain Monte Carlo (MCMC) algorithm have been developed for mapping multiple QTLs at the same time (e.g. Satagopan et al. 1996; Uimari and Hoeschele 1997; Sillanpää and Arjas 1998, 1999; Yi et al. 2003; Kilpikari and Sillanpää 2003; Yi 2004; Sillanpää and Bhattacharjee 2005). Among these methods, those based on Bayesian variable selection (Yi et al. 2003; Yi 2004; Sillanpää and Bhattacharjee 2005) are advantageous in that they can be implemented via a simple and easy-to-use Gibbs sampler. As proposed by Kilpikari and Sillanpää (2003), these Bayesian methods can be extended to the whole genome association mapping. Their practical application to the whole genome association mapping, however, has been rarely attempted.

Asian cultivated rice, *Oryza sativa* L., is an important crop and staple food for half of the world's population. Rice landraces show a broader range of phenotypic variation than do modern rice cultivars. Among the variations left unused in modern cultivars, there are many that would be valuable for breeding programs. To use these variations more actively in future rice breeding, it is necessary to establish an efficient statistical method for detecting the QTLs responsible for the variations. As rice is a highly selfing species and is expected to have high levels of population structure because of the nature of its breeding history, it remains almost unknown whether the association mapping approach would be efficient for mapping QTLs in rice (Yu et al. 2006).

We propose an approach that combines the Bayesian variable selection method for mapping multiple QTLs with an association mapping method that directly incorporates estimates of population structure. We evaluated the efficiency of our approach in the whole genome association mapping of a rice germplasm collection. We performed simulation analyses based on real marker data. We also performed analyses of real trait data, i.e. the size and shape of milled rice grains, which are typical traits that show moderate to high heritability. Finally, we discuss the prospects for the application of our approach to the whole genome association mapping of germplasm collections.

Materials and methods

Plant materials

Recently, 332 rice accessions were selected as representatives of the rice germplasm maintained at the National Institute of Agrobiological Sciences (NIAS) Genebank and were genotyped for 179 restriction fragment length polymorphism (RFLP) markers (Kojima et al. 2005). The 332 accessions originate from 23 countries and include 281 landraces and 51 modern cultivars (Table 1 in Kojima et al. 2005). The 179 RFLP markers have been located on the high-density genetic linkage map of rice (Kurata et al. 1994; Harushima et al. 1998) and distributed as landmarker RFLP sets from the NIAS DNA Bank (http://www.dna.affrc.go.jp/). We used the 332 accessions in this study.

Among the 332 accessions, the alleles of two reference cultivars Nipponbare (ssp. *japonica*) and Kasalath (ssp. *indica*) dominated in most of the 179 RFLP markers. The number of alleles observed ranged from 2 to 8 (mean 3.1) per locus (Kojima et al. 2005). The average frequencies of the Nipponbare and Kasalath alleles over all markers were 0.53 and 0.37, respectively (Kojima et al. 2005).

Population structure

The population structure among the 332 accessions was inferred by model-based Bayesian clustering analysis with

the 179 RFLP markers by using the program Structure (Pritchard et al. 2000a). Because rice is a highly selfing species and the accessions were almost homozygous, we treated each accession as a haploid in the model of the Bayesian clustering analysis. MCMC cycles were repeated 1×10^6 times after 1×10^4 cycles of a burn-in period. In the analyses, we tested the admixture models with two to eight populations. The model in which the number of populations (J) was six showed higher log-likelihood values than the other models. Thus, we chose $J = 6$ and obtained estimates for the proportion of accession i's genome that originated from population j, q_{ij}. The $\mathbf{Q}$ matrix whose (i, j)-th element was q_{ij} was further incorporated into the model of Bayesian association mapping of multiple QTLs.

Statistical model

We considered only marker positions as putative QTLs in our association mapping method. Each marker position k ($k = 1, 2, ..., K$) has its own indicator variable γ_k, where the value one ($\gamma_k = 1$) corresponds to the case in which the marker is included in the model as a QTL representative, and the value zero ($\gamma_k = 0$) implies exclusion. Here, we considered the marker loci as bi-allelic. Each marker position has its own genetic effect coefficient β_k, where the effects associated with two homozygous genotypes of marker k are β_k and $-\beta_k$, respectively. The observed phenotypic value of individual i ($i = 1, 2, ..., N$), y_i, can then be described by the linear model,

$$y_i = \sum_{j=1}^{J} q_{ij}\alpha_j + \sum_{k=1}^{K} x_{ik}\gamma_k\beta_k + e_i, \tag{1}$$

where q_{ij} is the (i, j)-th element of matrix $\mathbf{Q}$, α_j is the population effect associated with population j ($j = 1, 2, ..., J$), x_{ik} denotes the genotype of marker k for individual i, and is defined by 1 or –1 for the two genotypes, and e_i is the residual error assumed to follow $N(0, \sigma_e^2)$. Because rice is a highly selfing species, the dominance effect was not included. Epistatic effects can be included in the model theoretically, but we excluded them for simplicity. The model (1) can be formulated in matrix notation as

$$\mathbf{y} = \mathbf{Q}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\eta} + \mathbf{e}, \tag{2}$$

where $\mathbf{y}$ is an $N \times 1$ vector whose i-th element is y_i, $\boldsymbol{\alpha}$ is a $J \times 1$ vector whose j-th element is α_j, $\mathbf{X}$ is an $N \times K$ matrix whose (i, k)-th element is x_{ik}, $\boldsymbol{\eta}$ is a $K \times 1$ vector whose k-th element is $\gamma_k\ \beta_k$, and $\mathbf{e}$ is an $N \times 1$ vector whose i-th element is e_i.

MCMC algorithm for parameter estimation

Our method is based on a variable selection method developed by Kuo and Mallick (1998). The method is similar to, but simpler than, the method developed by George and McCulloch (1993), which has been utilized in multiple QTL mapping (Yi et al. 2003; Yi 2004).

Prior and posterior distribution of parameters

In our method, we considered the prior distributions of the parameters $\boldsymbol{\beta}, \gamma_k$, and σ_e^2 as

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{I}\sigma_\beta^2),$$

$$\gamma_k \sim B(1, p_k),$$

and

$$\sigma_e^2 \sim \upsilon_e s_e^2 \chi_{\upsilon_e}^{-2},$$

where σ_β^2, p_k, υ_e, and s_e^2 are hyperparameters for the distributions. We considered a flat prior (i.e., an improper uniform distribution) for the parameter $\boldsymbol{\alpha}$. That is, $p(\boldsymbol{\alpha}) \propto$ constant.

Now, let

$$\mathbf{X}^* = [\gamma_1\mathbf{x}_1, \ldots, \gamma_K\mathbf{x}_K], \tag{3}$$

where $\mathbf{X} = [\mathbf{x}_1, ...,\mathbf{x}_K]$. Then, Eq. 2 can be written as

$$\mathbf{y} = \mathbf{Q}\boldsymbol{\alpha} + \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}.$$

Here, let

$$\mathbf{Q}\boldsymbol{\alpha} + \mathbf{X}^*\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\theta},$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2/\sigma_\beta^2 \end{bmatrix}, \tag{4}$$

$$\mathbf{C} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \boldsymbol{\Sigma}, \tag{5}$$

and

$$\mathbf{r} = \mathbf{W}^{\mathrm{T}}\mathbf{y}, \tag{6}$$

where

$$\mathbf{W} = [\mathbf{Q}\ \mathbf{X}^*] \tag{7}$$

and $\boldsymbol{\theta} = [\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}]^{\mathrm{T}}$. Then the conditional posterior distribution of the i-th element of $\boldsymbol{\theta}$ is

$$\theta_i|\boldsymbol{\theta}_{-i}, \boldsymbol{\gamma}, \sigma_e^2, \mathbf{y} \sim N(\tilde{\theta}_i, \sigma_e^2/c_{i,i}), \tag{8}$$

where $\boldsymbol{\gamma}$ is a vector whose k-th element is $\gamma_k, \tilde{\theta}_i = (r_i - \mathbf{C}_{i,-i}\boldsymbol{\theta}_{-i})/c_{i,i}, c_{i,i}$ is the i-th diagonal element of the matrix $\mathbf{C}$, r_i is the i-th element of the vector $\mathbf{r}$, $\mathbf{C}_{i,-i}$ is a row vector obtained by deleting element i from the i-th row of the matrix $\mathbf{C}$, and $\boldsymbol{\theta}_{-i}$ is a vector obtained by element i from the vector $\boldsymbol{\theta}$ (Sorensen and Gianola 2002).

The fully conditional posterior distribution of σ_e^2 is given by

$$\sigma_e^2|\boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y} \sim \tilde{\upsilon}_e \tilde{s}_e^2 \chi_{\tilde{\upsilon}_e}^{-2}, \tag{9}$$

where $\tilde{\upsilon}_e = n + \upsilon_e$ and $\tilde{s}_e^2 = [(\mathbf{y} - \mathbf{W}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - \mathbf{W}\boldsymbol{\theta}) + \tilde{\upsilon}_e s_e^2]/\tilde{\upsilon}_e$.

The fully conditional posterior distribution of γ_k is given by

$$\gamma_k|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{-k}, \sigma_e^2, \mathbf{y} \sim B(1, \tilde{p}_k), \tag{10}$$

where $\boldsymbol{\gamma}_{-k}$ is a vector obtained by element k from the vector $\boldsymbol{\gamma}, \tilde{p}_k = a_k/(a_k + b_k)$,

$$a_k = p_k \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^*)^{\mathrm{T}}(\mathbf{y} - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^*)\right\} \tag{11}$$

$$b_k = (1 - p_k)\exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{y} - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^{**})^{\mathrm{T}}(\mathbf{y} - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^{**})\right\}. \tag{12}$$

The vector $\boldsymbol{\eta}_k^*$ in Eq. 11 is the column vector of $\boldsymbol{\eta}$ with the k-th entry replaced by β_k. Similarly, $\boldsymbol{\eta}_k^{**}$ in Eq. 12 is obtained from $\boldsymbol{\eta}$ with the k-th entry replaced by 0.

MCMC sampling

On the basis of the above equations for prior and posterior distributions, we can use the Gibbs sampler to generate MCMC samples from the posterior distribution of the model parameters. In the sampling, we set hyperparameters for the prior distributions as $\sigma_\beta^2 = 16$, $p_k = 0.5$, $\upsilon_e = -2$, and $s_e^2 = 0$. The hyperparameter σ_β^2 was determined by evaluating the influence of σ_β^2 on the MCMC estimation with three different settings (i.e., $\sigma_\beta^2 = 1$, 16, *or* 100), as described in the Results. Setting the initial values of the parameters as $\sigma_e^2 = 1, \boldsymbol{\alpha} = \mathbf{0}, \boldsymbol{\beta} = \mathbf{0}$, and $\boldsymbol{\gamma} = \mathbf{0}$, the Gibbs sampler proceeds as follows:

1. Update $\mathbf{W}$ and $\boldsymbol{\Sigma}$ with Eqs. 4 and 7, and then update $\mathbf{C}$ and $\mathbf{r}$ with Eqs. 5 and 6.
2. Sample $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (i.e., $\boldsymbol{\theta}$) from the full conditional posterior distribution described in Eq. 8.
3. Sample σ_e^2 from the full conditional posterior distribution described in Eq. 9.
4. Sample $\boldsymbol{\gamma}$ from the full conditional posterior distribution described in Eq. 10.

The above process was repeated many times (see ''Data analysis procedure'') to obtain MCMC samples.

Simulated datasets

In our simulation studies, we used simulated datasets with the 332 rice accessions. We used the observed genotypes of 179 RFLP markers of the 332 rice accessions to generate the simulated datasets. The marker genotypes remained the same as those in the real data. We then simulated 10 QTLs at 10 different positions randomly selected from the 179 RFLP markers. We simulated the genotypes of the QTLs according to the RFLP markers at the same positions. For half (i.e., five) of the QTLs, we simulated the QTL genotype as *QQ* if the marker genotype was homozygous for the Nipponbare allele, and as *qq* otherwise. For the other half of the QTLs, we simulated the QTL genotype in the opposite way (i.e., as *qq* if the marker genotype was homozygous for the Nipponbare allele, and as *QQ* otherwise). We then simulated the genotypic values of the QTLs according to the true parameter values. The true parameter values of the QTL were set as 0.5 for the *QQ* genotype and –0.5 for the *qq* genotype. Next, we simulated a residual variance set at $\sigma_e^2 = 1$ to generate the phenotypic values of all the accessions. Finally, we simulated population effects and added them to the phenotypic values as follows. First, we calculated the phenotypic variance σ_y^2 at this point. Next, we sampled the population effect α_j (j = 1, 2, ..., J) from $N(0, 0.25\sigma_y^2)$. Then, we added the population effect α_j to the phenotypic value of accession i, weighted by q_{ij}. The proportion of variance due to population effect was scaled as 20% (=0.25/1.25), reflecting population effects estimated from the real datasets (see Results). The process described above was performed 100 times to generate 100 simulated datasets.

Real dataset

Of the 332 rice accessions, 296 were cultivated in an experimental field at NIAS (Tsukuba, Ibaraki, Japan) in the 2003 cropping season and were used in the real-dataset study. Six milled rice grains were randomly selected from each accession and photographed by digital camera (EOS 10D, Canon, Japan) at 0.002 mm/pixel resolution. The length (LEN) and width (WID) of the grains in millimeters were measured by image analysis. The length to width ratio (LWR), which represents grain shape, was given by LEN/WID. For each of these traits, first, we performed a one-way ANOVA to test the variation among accessions. Then, the average for the six grains was used as the phenotypic value of each accession.

Data analysis procedure

In our model, we considered the marker loci to be bi-allelic. As described above, in our data, the Nipponbare and Kasalath alleles dominated among the 332 accessions at most loci. Thus, in the association mapping analyses, we regarded all loci as bi-allelic by scoring them by the presence (1) or absence (–1) of the Nipponbare allele.

For each dataset, MCMC cycles were repeated 1.5×10^5 times, and the first 5×10^4 cycles (burn-in) were not used for estimating the parameter values. Sampling was carried out every ten cycles to reduce serial correlation, so that the total number of samples kept was 1×10^4. This sampling scheme was based on the evaluation of the convergence of MCMC cycles using QTL occupancy probability (Heath 1997; Uimari and Sillanpää 2001; Hayashi and Awata 2005), as described in the Results.

We regarded a marker as significant when the mean of the posterior distribution of γ_k was larger than a specified threshold. Two different thresholds (0.5 and 0.9) were tested. That is, a marker was regarded as significant when it was included in the model as a QTL representative (i.e., $\gamma_k = 1$) in over half (in the case of the 0.5 threshold) or 90% (in the case of the 0.9 threshold) of MCMC samples. In the following sentences, we refer the thresholds of 0.5 and 0.9 as 'moderate' and 'strict' thresholds, respectively.

For the simulated datasets, we also performed analyses based on the following reduced models, as well as on the full model described in Eq. 1:

Model R1: Single QTL model without population effects. The model equation was

$$y_i = \mu + x_{ik}\beta_k + e_i, \tag{13}$$

where μ was the overall mean.

Model R2: Single QTL model with population effects. The model equation was

$$y_i = \sum_{j=1}^{J} q_{ij}\alpha_j + x_{ik}\beta_k + e_i. \tag{14}$$

Model R3: Multiple QTL model without population effects. The model equation was

$$y_i = \mu + \sum_{k=1}^{K} x_{ik}\gamma_k\beta_k + e_i. \tag{15}$$

Parameters in the models R1 and R2 could be estimated by simple regression analysis and multiple linear regression analysis, respectively. For both models, the statistical significance of each marker could be determined by t test for the significance of the regression coefficient. We regarded a marker as significant when the P value was less than 0.001. Parameters in model R3 could be estimated in the same way as in the full model. In the MCMC sampling, we considered a flat prior for the parameter μ. The significance of each marker was determined as in the case of the full model.

To compare the performances (i.e., the accuracies of estimation) of the full and reduced models, we calculated the following indices for each model in each simulated dataset:

False-negative rate (FNR):

$$\mathrm{FNR} = \frac{n_{\mathrm{fn}}}{n_{\mathrm{tp}} + n_{\mathrm{fn}}},$$

where n_{tp} was the number of loci regarded as significant when they were QTLs (i.e., the number of true positives) and n_{fn} was the number of loci mistakenly regarded as non-significant when in fact they were QTLs (i.e., the number of false negatives).

False-positive rate (FPR):

$$\mathrm{FPR} = \frac{n_{\mathrm{fp}}}{n_{\mathrm{fp}} + n_{\mathrm{tn}}},$$

where n_{fp} was the number of loci mistakenly regarded as significant when in fact they were not QTLs (i.e., the number of false positives), and n_{tn} was the number of loci regarded as non-significant when they were not QTLs (i.e., the number of true negatives).

False-discovery rate (FDR):

$$\mathrm{FDR} = \frac{n_{\mathrm{fp}}}{n_{\mathrm{fp}} + n_{\mathrm{tp}}}.$$

We also calculated the following index for each model using all simulated datasets.

The root-mean-square error (RMSE) between the estimated and true values of the genetic effect of the QTL, scaled by the true value:

$$\mathrm{RMSE} = \sqrt{\frac{1}{D}\sum_{i}^{N}\sum_{k}^{K}\delta_{k,i}\left(\frac{\hat{\beta}_{k,i} - \beta_{k,i}}{\beta_{k,i}}\right)^2},$$

where N is the number of simulated datasets. $\hat{\beta}_{k,i}$ and $\beta_{k,i}$ are the estimate and true value of the genetic effect of the k-th marker locus in the i-th dataset, respectively. $\delta_{k,i}$ is an indicator variable in which the value one (i.e., $\delta_{k,i} = 1$) corresponds to the case in which the k-th locus is a true positive in the i-th dataset, and the value zero (i.e., $\delta_{k,i} = 0$) corresponds to the remainder of possibilities (i.e., the k-th locus is a false negative, a false positive, or a true negative). D is the number of true positives over all simulated datasets; that is, this index is the RMSE of true positives over all simulated datasets.

For the real datasets, we also calculated the expected FDR (Benjamini and Hochberg 1995) as follows. First, P values were calculated for all K marker loci under the null hypothesis ''there is no QTL at the specified locus''. Next, we determined the largest P value, $P_{\max}$, among the P values of all n_p significant loci. Finally, the expected FDR (EFDR) was calculated as

$$\mathrm{EFDR} = \frac{\mathrm{KP}_{\max}}{n_p}.$$

To determine the P value under the null hypothesis, we empirically obtained the null distribution of the mean of the posterior distribution of γ_k from the results of simulation analyses with the full model. That is, we gathered 16,900 values (169 markers $\times$ 100 simulated datasets) of the mean of the posterior distribution of γ_k of markers that did not have a QTL in the simulation analyses, and we considered them as an empirical distribution of the mean of the posterior distribution of γ_k. We used this empirical distribution as the null distribution, and we determined P values corresponding to the means of the posterior distribution of γ_k of all marker loci in the real data analyses.

Results

Simulated data analysis

The mixing property of the MCMC and the convergence of MCMC estimation were evaluated by analyzing one simulated dataset with the full model. The number of cycles required for the convergence of MCMC estimation was evaluated on the basis of a plot of cumulative QTL occupation probabilities (Fig. 1b) as a function of cycle number, that is, Pr([number of QTLs] $\leq l$ | [cycle number] = k) (where $10 \leq l \leq 18$, $0 \leq k \leq 1.5 \times 10^5$, respectively) following Heath (1997). The MCMC mixed well between models with different numbers of QTLs (Fig. 1a), and the cumulative QTL occupation probabilities became stable by the 5×10^4 th cycle (Fig. 1b). Analysis of the same dataset with model R3 revealed the same tendency (data not shown). Thus, we determined that a chain of 1.5×10^5

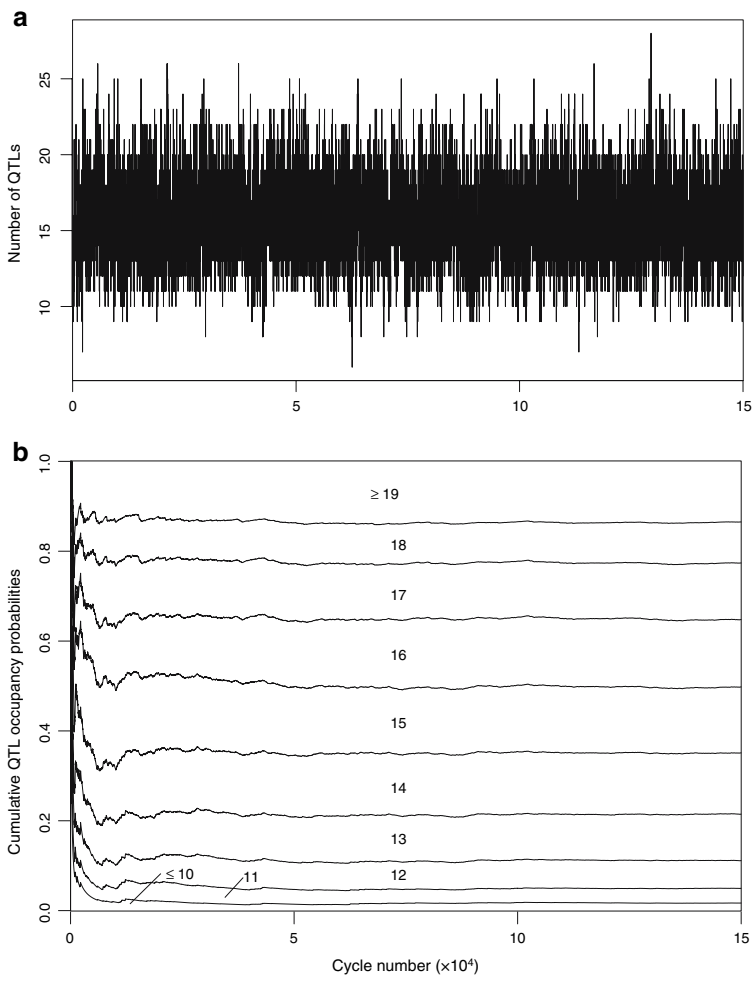

Fig. 1 **a** Sampled values for the number of QTLs over 150,000 MCMC cycles, and **b** cumulative QTL occupancy probabilities as a function of number of cycles, obtained in the analyses of one simulated dataset. The 10 areas in **b** (one at the top, eight between the two lines, and one at the bottom) indicate the probabilities of the number of fitted QTLs, as denoted by the numeral in each area

cycles and a burn-in period of 5×10^4 cycles were sufficient to achieve convergence of MCMC estimation.

To evaluate the influence of the prior variance of a QTL effect (i.e., σ_β^2) on the MCMC estimation, we analyzed one simulated dataset with the full model by setting σ_β^2 as 1, 16, or 100. Although the mean of the posterior distribution of γ_k was generally smaller for larger σ_β^2 (i.e., the mean of the posterior distribution of γ_k was 0.22, 0.10, and 0.07 on average for $\sigma_\beta^2 = 1$, 16, and 100, respectively), the significance of each marker under the strict threshold was nearly identical among the three settings (i.e., discordance between the settings was observed in only 2 out of 179 markers). The mean of the posterior distribution of β_k was highly correlated between settings (i.e., $r = 0.94$ for $\sigma_\beta^2 = 16$ vs. 1, and $r = 0.98$ for $\sigma_\beta^2 = 16$ vs. 100). These results indicate that the prior variance of a QTL effect did not have a large influence on the MCMC estimation. Therefore, we set σ_β^2 as 16 in the subsequent analyses.

In the 100 simulated datasets, the mean proportion of phenotypic variance explained by each QTL (i.e., heritability) was 0.085, and the mean joint heritability of all QTLs was 0.472 (Table 1).

The histograms in Fig. 2 show the numbers of simulated datasets (out of 100) that fell into specified intervals for FNR, FPR, and FDR. For FNR, the full model with the moderate threshold (i.e., 0.5) tended to show smaller values than the full model with the strict threshold (i.e., 0.9) and models R1, R2, and R3 (Fig. 2a). In increasing order, the average FNR was smallest in the full model with the moderate threshold, followed by model R3 with the moderate threshold, the full model with the strict threshold, model R3 with the strict threshold, and model R1 (Table 2). The average FNR of the full model with the strict threshold was a little smaller than those of models R1 and R2.

Table 1 Heritability of each QTL and joint heritability of all QTLs in 100 simulated datasets

	Value
Heritability of each QTL[a]	
Min.	0.002
Mean	0.085
Max.	0.155
SD	0.031
Joint heritability of all QTLs[b]	
Min.	0.228
Mean	0.472
Max.	0.831
SD	0.126

[a] Proportion of phenotypic variance explained by each QTL (i.e., heritability of each QTL). Minimum (*Min.*), mean, maximum (*Max.*), and standard deviation (*SD*) were calculated over 1,000 QTLs (i.e., 10 QTLs × 100 simulated datasets)

[b] Proportion of phenotypic variance explained by all QTLs (i.e., joint heritability of all QTLs). Minimum (*Min.*), mean, maximum (*Max.*), and standard deviation (*SD*) were calculated over 100 simulated datasets

For FPR, model R1 tended to show considerably larger values than the other models (Fig. 2b). The average FPR of model R1 reached 45%, whereas the average FPR of the other models was less than 2% (Table 2). In increasing order, the average FPR was smallest in the full model with the strict threshold, followed by model R3 with the strict threshold, the full model with the moderate threshold, model R2, and model R3 with the moderate threshold.

For FDR, model R1 tended to show considerably larger values than the other models (Fig. 2c). In model R1, 75 out of 100 datasets had FDRs greater than 90%. In contrast, in the full model with the strict threshold, 91 out of 100 datasets had FDRs equal to 0%. The average FDR of model R1 was as high as 89%, whereas the average FDR of the full model with the strict threshold was less than 2% (Table 2). In increasing order, the average FDR was smallest in the full model with the strict threshold, followed by model R3 with the strict threshold, the full model with the moderate threshold, model R2, and model R3 with the moderate threshold.

To evaluate the accuracy of our estimates of genetic effects, we also calculated the RMSE for accurately detected QTLs (i.e., true positives). In increasing order, the RMSE was smallest in the full model with the strict threshold, the full model with the moderate threshold, model R3 with the strict threshold, model R3 with the moderate threshold, and model R2 (Table 2). The RMSE of model R1 was about three times those of the other models.

Real data analysis

ANOVA revealed that the differences between accessions were highly significant ($P < 0.001$) for all traits, suggesting that both the size and shape of the milled rice grains were heritable. Association mapping with the full model showed that the proportion of variance due to population effects of LEN, WID, and LWR was 4.35, 41.0, and 7.97%, respectively. With the moderate threshold (i.e., 0.5), we found six, three, and eight significant markers for LEN, WID, and LWR, respectively (Fig. 3). With this threshold, the EFDRs for LEN, WID, and LWR were 13.5, 8.83, and 11.0%, respectively. Four markers were significant for both LEN and LWR, whereas one was significant for both WID and LWR. No overlap of significant markers was observed between LEN and WID. With the strict threshold (i.e., 0.9), we found three, two, and one significant markers for LEN, WID, and LWR, respectively (Table 3). With this threshold, the EFDRs for LEN, WID, and LWR were 2.12, 1.06,

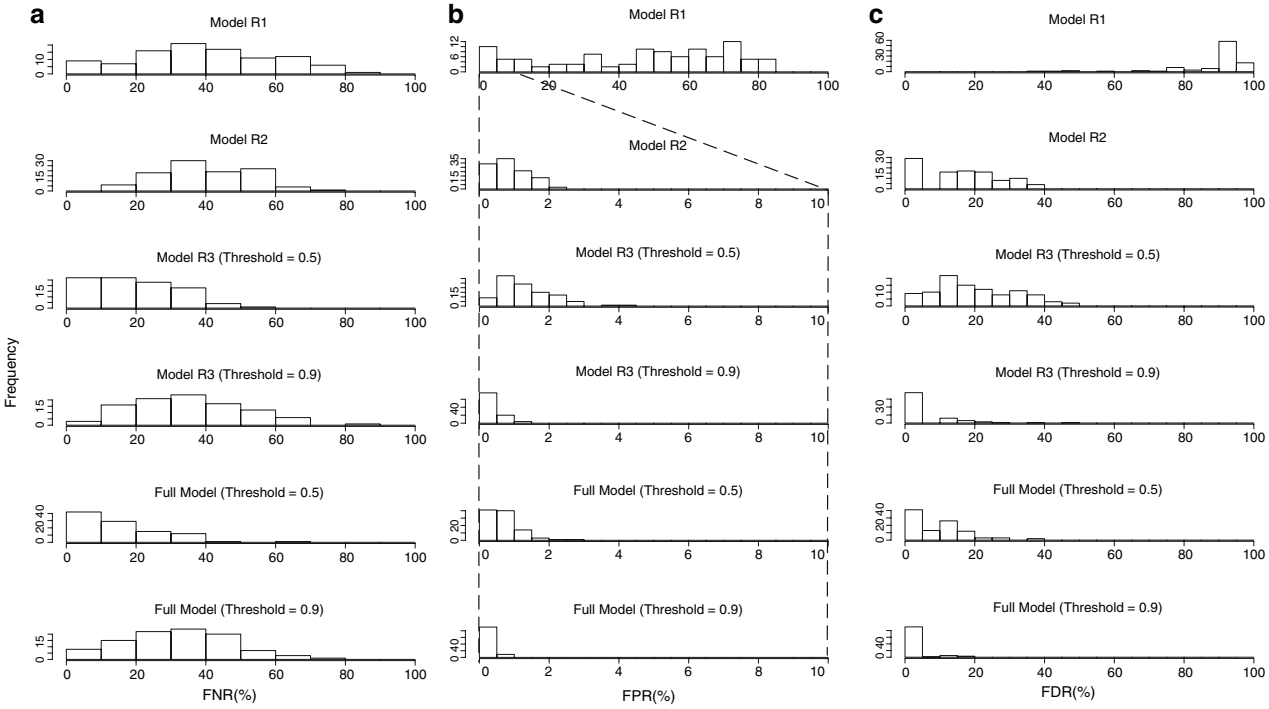

Fig. 2 Histograms of **a** false-negative rate (*FNR*), **b** false-positive rate (*FPR*), and **c** false-discovery rate (*FDR*), obtained from 100 simulated datasets generated by using real marker data. Each simulated dataset was analyzed by different statistical models, and FNR, FPR, and FDR were calculated individually for each model. 'Model R1' is a single QTL model ignoring population structure. 'Model R2' is a single QTL model considering population structure. 'Model R3' is a multiple QTL model ignoring population structure. 'Full Model' is a multiple QTL model considering population structure. For 'Model R3' and 'Full Model', significant markers were decided on two different thresholds, 0.5 and 0.9. For details, see text

Table 2 False-negative, false-positive, and false-discovery rates and root-mean-square error between estimated and true values of QTL effects

Model	R1	R2	R3 (M)	R3 (S)	Full (M)	Full (S)
Average FNR (%)[a]	44.7	44.9	24.1	40.2	19.4	37.1
Average FPR (%)[b]	45.7	0.734	1.25	0.166	0.509	0.0533
Average FDR (%)[c]	89.0	16.5	20.5	4.56	8.97	1.23
RMSE (%)[d]	96.1	32.2	26.8	26.8	24.1	24.0

R1 Single QTL model ignoring effects due to population structure, *R2* single QTL model considering effects due to population structure, *R3 (M)* multiple QTL model ignoring effects due to population structure with a moderate (i.e., 0.5) threshold, *R3 (S)* multiple QTL model ignoring effects due to population structure with a strict (i.e., 0.9) threshold; *Full (M)* multiple QTL model considering effects due to population structure with a moderate threshold, *Full (S)* multiple QTL model considering effects due to population structure with a strict threshold

[a] Average false-negative rate (%). The average was calculated over 100 simulated datasets

[b] Average false-positive rate (%)

[c] Average false-discovery rate (%)

[d] Root-mean-square error between estimated and true values of genetic effects of false-positive QTL, scaled by true values

and 3.18%, respectively. For LEN, the three significant markers were found on chromosomes 3, 5, and 11. For WID, the two significant markers were found on chromosomes 1 and 5. The one significant marker detected for LWR was identical to one detected on chromosome 5 for LEN. For LEN, the estimated effects of the significant markers were all negative, indicating that the allele linked to the Nipponbare allele shortened the length of the rice grains. For WID, one significant marker showed a positive effect and one showed a negative effect. For LWR, the significant marker showed a negative effect (as expected from its negative effect in LEN).

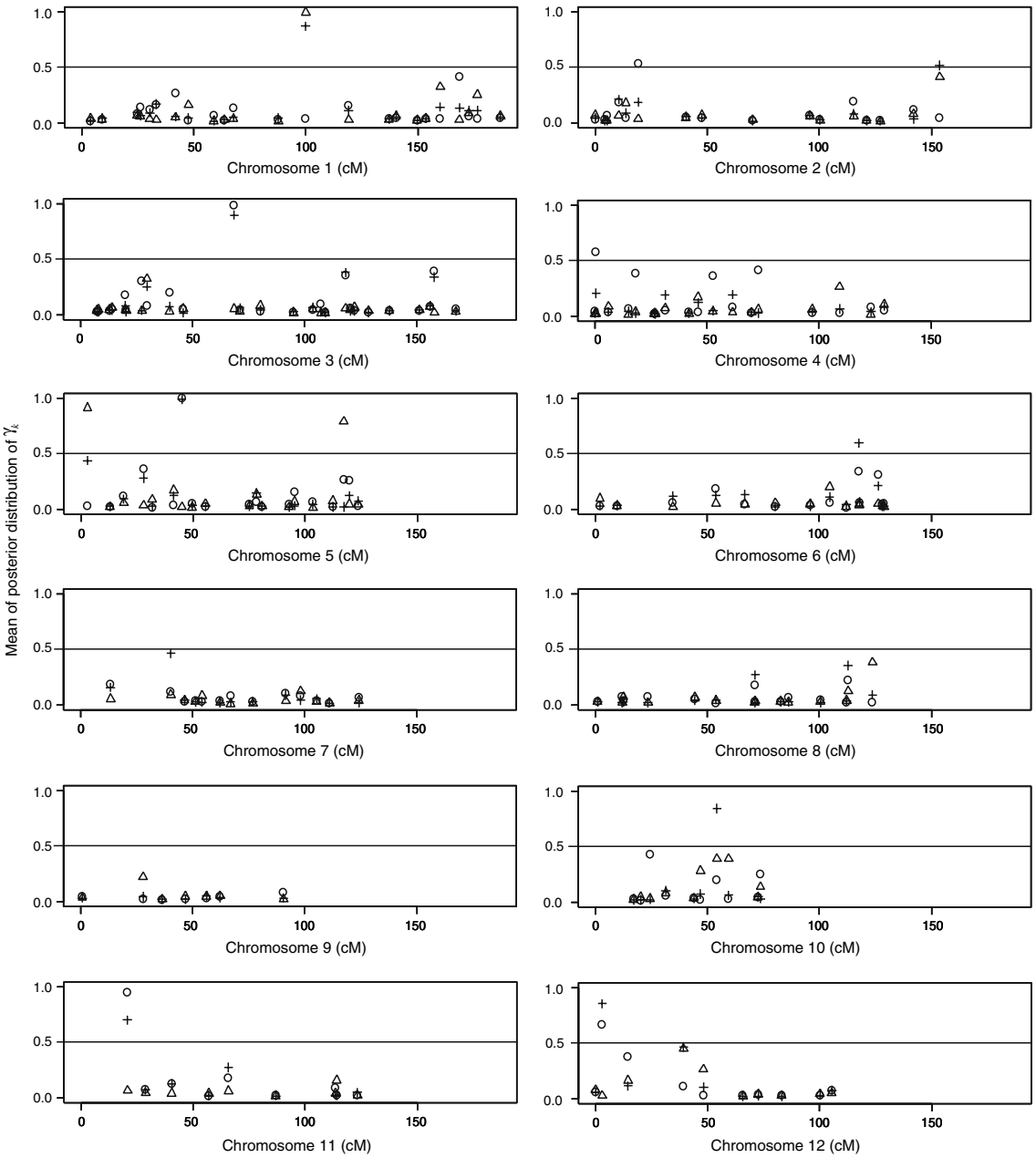

Fig. 3 Mean of posterior distributions of γ_k of 179 RFLP markers, estimated for the length (*circles*), width (*triangles*), and length to width ratio (*crosses*) of milled rice grains. Each marker position k ($k = 1, 2, ..., K$) has its own indicator variable γ_k, where the value one ($\gamma_k = 1$) corresponds to the case in which the marker is included in the model as a QTL representative, and the value zero ($\gamma_k = 0$) implies exclusion. Mean of posterior distributions of γ_k were plotted against marker locations estimated on the genetic linkage map constructed by Kurata et al. (1994)

Discussion

In association mapping, it is crucial to control for the false positives caused by population structure as well as to enhance statistical power in detecting QTLs (i.e., control for false negatives). Our simulation studies indicated that the single QTL model ignoring population structure (i.e., model R1) induced false positives more commonly than the other models (i.e., the full model and models R2 and R3). The average FPR and FDR of model R1 were over 40 and 80%, respectively, suggesting that a single QTL model that ignores population structure is useless in association mapping of the rice germplasm collection. Moreover, the error of estimation of genetic effects (i.e., RMSE) of model R1 was about three times those of the other models. The results indicate that effects due to population structure

Table 3 Locations and estimated parameters of significant markers for traits related to the size and shape of milled rice grains

Trait	Chromosome	Location[a]	Marker	γ[b]	β[c]
LEN	3	68.6	R250	0.982	–0.27 ± 0.06
	5	45.0	R569	0.996	–0.19 ± 0.04
	11	20.7	G1465	0.946	–0.18 ± 0.05
WID	1	99.9	R1928	0.995	0.13 ± 0.03
	5	2.9	C597	0.914	–0.10 ± 0.03
LWR	5	45.0	R569	0.987	–0.11 ± 0.02

LEN length of milled rice grain, *WID* width of milled rice grain, *LWR* length to width ratio of milled rice grain

[a] Marker location estimated on the genetic linkage map constructed by Kurata et al. (1994)

[b] Mean of posterior distribution of γ_k

[c] Mean and standard deviation of posterior distribution of β_k. In the calculation of mean and standard deviation, we took into account MCMC samples in which $\gamma_k = 1$

should be carefully taken into account when complex population structure is present within a germplasm collection.

The multiple QTL model that considered population structure (i.e., the full model) could suppress both false positives and false negatives more effectively than the single QTL model that considered population structure (i.e., model R2). With either the moderate (i.e., 0.5) or strict (i.e., 0.9) threshold, the average FPR and FDR of the full model were less than those of model R2. With the strict threshold in particular, false positives were barely detected (i.e., only 9 false positives in 100 simulated datasets). This indicates that the multiple QTL model can control the false positives better than the single QTL model. Meanwhile, the FNR of the multiple QTL model with either threshold was also smaller than that of model R2, indicating that the statistical power in detecting QTLs can also be enhanced by simultaneously taking multiple QTLs into account in the model. Moreover, the error of estimation of genetic effects (i.e., RMSE) was smaller in the multiple QTL model with either threshold than in the single QTL models. These results indicate that the multiple QTL model has statistically desirable attributes over the single QTL model for application to the whole genome association mapping. Thus, Bayesian methods for identifying multiple QTLs can be powerful tools for the whole genome association mapping even when complex population structure is present in data.

The multiple QTL model that considered population structure (i.e., the full model) could suppress FNR, FPR, FDR, and RMSE to a greater extent than the multiple QTL model that ignored population structure (i.e., model R3). This indicates that population effects should be included even in the multiple QTL model. In the application to actual data, the superiority of the full model over model R3 may be larger than that in the simulation studies. In the simulation studies, all simulated QTLs were located just on the marker loci: that is, the effects of the QTLs could be fully explained by the effects associated with the marker genotypes. In the application to actual data, however, the effect of QTLs cannot always be fully explained by marker genotypes. Moreover, in the actual data there may be true polygenic effects. That is, there may be many loci that have effects too small for their detection as QTLs, but that nevertheless lead populations to have different means. In the full model, genetic variance that is not associated with the marker loci (i.e., polygenic variance and/or genetic variance due to QTLs that are not closely linked to the marker loci) can be absorbed by the population effects included in the model. In our simulation studies, we also simulated the population effects explaining about 20% of phenotypic variance. The variance due to population effects, however, may be much larger in the actual data than in our simulated datasets, as is the case for WID in this study.

We tested both moderate (i.e., 0.5) and strict (i.e., 0.9) thresholds in deciding on significant markers on the basis of MCMC samples. As described above, the average FPR and FDR were less than 2% with the strict threshold. Because of the trade-off relationship between false-positive and false-negative rates, the average FNR was not as small (37.1%) with the strict threshold. With the moderate threshold, in contrast, the average FNR was 19.4%, but the average FPR and FDR were larger than with the strict threshold. In the practical application of our model, an appropriate threshold, including an intermediate one, can be chosen in accordance with the intended purpose in consideration of the trade-off relationship between false-positive and false-negative rates. As described later, the procedure for controlling FDR (Benjamini and Hochberg 1995) may be used for choosing an appropriate threshold.

A trade-off relationship between false-positive and false-negative rates also exists for single QTL models. When we regarded a marker as significant at $P < 0.01$, the average FNR became small (29.1%) in model R2 (data not shown). This value was less than the average FNR of the full model with the strict threshold. The average FPR and FDR, however, became 3.49 and 43.0%, respectively—obviously larger than those of the full model. When the FDR reaches 43.0%, nearly half of the positives are false. In such a situation, QTL detection cannot be reliable. In contrast, when we regarded a marker as significant at $P < 0.0001$, the average FPR and FDR became small (0.160 and 5.55%, respectively) in model R2, although the values were still larger than those of the full model with the strict threshold (data not shown). In this case, the average FNR became large (60.6%) in model 2, indicating that over half of the QTLs would be overlooked.

In the real data analyses, we estimated EFDR by determining a P value corresponding to a specified value of the mean of the posterior distribution of γ_k based on the empirical null distribution obtained from the simulation analyses. With this procedure, we can also control FDR at a specified level as described by Benjamini and Hochberg (1995). First, P values are computed from the mean of the posterior distribution of γ_k for all K markers. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(K)}$ be the ordered P values. Next, the largest i is determined for which $P_{(i)} \leq \frac{iq^*}{K}$, where q^* is the FDR to be controlled at. Finally, the mean of the posterior distribution of γ_k corresponding to $P_{(i)}$ is obtained on the basis of the empirical null distribution. This value is then the threshold that controls FDR at q^*. This procedure may be a good criterion for choosing an appropriate threshold for our model.

As described by Yu et al. (2006), a model that considers both relatedness among accessions and population structure may further suppress false-positive and false-negative rates, since the relatedness is expected to account for finer-scale variations caused by different genetic backgrounds than variations caused by the population structure. A mixed-model approach in which the pedigree-based relatedness is taken into account has been successfully applied to the whole genome association mapping studies in maize (e.g. Parisseaux and Bernardo 2004; Zhang et al. 2005a). As described by Yu et al. (2006), relatedness can be also estimated on the basis of marker polymorphisms by using methods such as that proposed by Ritland (1996). In this study, however, we did not incorporate relatedness in our model. In our data, the allele frequencies of each marker locus differ from population to population; it is therefore necessary to take into account the difference in allele frequencies when we estimate relatedness between accessions. In the Bayesian clustering analysis, the allele frequencies of each locus in each population were also estimated. It may be possible to estimate the relatedness in the presence of population structure by using the information on allele frequencies and population structure (i.e., **Q** matrix) estimated by the Bayesian clustering analysis. Future work will be needed to develop a model that considers both relatedness and population structure in the whole genome association mapping of the rice germplasm collection.

In our model, only marker positions were considered as putative QTLs. In practice, however, linkage disequilibrium between marker and QTL alleles is not complete, and this incompleteness may cause errors in association mapping analyses. In our model, there is also an implicit assumption that the genetic effect of each marker position can be modeled with the same parameter in all populations. In practice, however, the effect may differ between populations, since both the pattern and degree of linkage disequilibrium between marker alleles and QTL alleles may also differ between populations. A more complex model that can deal with these discrepancies between assumption and practice may further suppress false-positive and false-negative rates, and should be addressed in future.

We analyzed the size and shape of milled rice grains as real trait data. We found three, two, and one significant markers, respectively, for LEN, WID, and LWR of milled rice grains. Of these, marker R569 on chromosome 5, which was significant in terms of LEN and LWR, may be linked to a QTL reported previously (Wan et al. 2005, 2006). In this region, we have also found a QTL for grain shape in a QTL analysis using BC_1F_{10} lines of Koshihikari/Kasalath//Koshihikari (unpublished data). Marker R250 on chromosome 3, which was significant in terms of LEN, may also be linked to a QTL that has been detected around the centromeric region of chromosome 3 (Huang et al. 1997; Redona and Mackill 1998; Tan et al. 2000; Kubo et al. 2001; Aluko et al. 2004; Li et al. 2004; Wan et al. 2005). The location of R250, however, does not completely correspond to the genomic region in which the candidate gene of the QTL was narrowed down (Wan et al. 2006, Fan et al. 2006). The discrepancy in the estimated location may be due to incomplete linkage disequilibrium between marker alleles and QTL alleles, as described above. The resolution power in the location estimation may be improved by the use of highly dense markers. Recently, a large number of single nucleotide polymorphisms (SNPs) have been positioned on the rice genome; thus, the level of linkage disequilibrium between marker alleles and QTL alleles may become reasonably high in the near future.

To perform a genome-wide QTL scan in association mapping studies, it is generally necessary to have a large number of markers distributed across the entire genome. This study, however, covered only 179 RFLP markers, which may be much less than the desired number of markers for a moderate-scale association mapping study. To clarify the usefulness of our approach in a larger-scale association mapping study, we analyzed one simulated dataset of 1,000 markers, in which randomly generated genotype data on 821 markers were added to the real data on 179 RFLP markers, and 10 QTLs were simulated at random positions on the 179 RFLP markers. As result, the full model with the moderate threshold detected 6 out of 10 simulated QTLs while achieving low FPR (0.8%) (data not shown), indicating that our approach could work on a dataset of 1,000 markers. At this analytical scale, however, our approach required a long time. For example, with our system (J2SE v5.0 + Red Hat Enterprise Linux ES + Intel Xeon 3.0 GHz), it took about 140 h to analyze the dataset of 1,000 markers, whereas it took about 5 h to analyze the dataset of 179 markers. To speed up the analysis time, we need to develop a more efficient MCMC sampling algo-

rithm. At the present time, a practical solution for this problem is to use the following two-step approach. First, an analysis with model R2 is performed over all markers. Second, an analysis with the full model is performed over the markers that are regarded as significant at the first step. Since the first step requires little time but can reduce the number of markers tested in the second step, it may be possible to analyze a dataset of tens of thousands of markers by this two-step approach within a practical time period. In order not to miss true QTLs in the first step, a moderate significance level (say, $P < 0.01$) may be appropriate in the first step analysis. False positives in the first-step can be eliminated in the second-step analysis.

Association mapping has generally not been attempted in rice, with a few exceptions (e.g., Zhang et al. 2005b). This reluctance to use association mapping in rice may be largely due to the fact that rice is a highly selfing species and is expected to have high levels of population structure in light of its breeding history. Our results indicate that association mapping would have good prospects in a highly selfing species such as rice if a proper method were to be adopted. Since accessions in the rice germplasm collection are inbred, they can be easily propagated and shared by many researchers. This enables us to accumulate large amounts of phenotypic data on various traits observed in various environments. Thus, the whole genome association mapping using the rice germplasm collection will be a useful tool, not only for detecting candidate genes, but also for detecting pleiotropic genes and genes showing interaction with the environment. There are many highly selfing crop species other than rice. The Bayesian method proposed here will be useful for the whole genome association mapping of them too.

Acknowledgments The authors are grateful to Jean-Luc Jannink and the two anonymous reviewers for their valuable comments and suggestions. We thank Akifumi Imada for assistance in digital photography. This work was supported by a grant from the Green Technology Project (QT1001) of the Ministry of Agriculture, Forestry and Fisheries of Japan.

References

Aluko G, Martinez C, Tohme J, Castano C, Bergman C, Oard JH (2004) QTL mapping of grain quality traits from the interspecific cross *Oryza sativa* × *O. glaberrima*. Theor Appl Genet 109:630–639

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165–1177

Brown AHD (1989) Core collections: a practical approach to genetic resources management. Genome 31:818–824

Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC, Rimmler JB, Lacke PA, Conneally PM, Schmader KE, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat Genet 7:180–183

Fan C, Xing Y, Mao H, Lu T, Han B, Xu C, Li X, Xhang Q (2006) *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. Theor Appl Genet 112:1164–1171

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Ann Rev Plant Biol 54:357–374

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. J Am Stat Assoc 88:881–889

Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin SY, Antonio BA, Parco A, Kajiya H, Huang N, Yamamoto K, Nagamura Y, Kurata N, Khush GS, Sasaki T (1998) A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics 148:479–494

Hayashi T, Awata T (2005) Bayesian mapping of QTL in outbred F2 families allowing inference about whether F0 grandparents are homozygous or heterozygous at QTL. Heredity 94:326–337

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760

Huang N, Parco A, Mew T, Magpantay G, McCouch S, Guiderdoni E, Xu JC, Subudhi P, Angeles ER, Khush GS (1997) RFLP mapping of isozymes, RAPD, and QTLs for grain shape, brown planthopper resistance in a doubled-haploid rice population. Mol Breed 3:105–113

Kerem B, Rommens J, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M and Tsui L-C (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kilpikari R, Sillanpää MJ (2003) Bayesian analysis of multilocus association in quantitative and qualitative traits. Genet Epidemiol 25:122–135

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) $Gm^{3;5,13,14}$ and Type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43:520–526

Kojima Y, Ebana K, Fukuoka S, Nagamine T, Kawase M (2005) Development of an RFLP-based rice diversity research set of germplasm. Breed Sci 55:431–440

Kraakman ATW, Niks RE, van den Berg PMMM, Stam P, van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446

Kubo T, Takano-Kai N, Yoshimura A (2001) RFLP mapping of genes for long kernel and awn on chromosome 3 in rice. Rice Genet Newslett 18:26–28

Kuo L, Mallick B (1998) Variable selection for regression models. Sankhya Ser B 60:65–81

Kurata N, Nagamura Y, Yamamoto K, Harushima Y, Sue N, Wu J, Antonio BA, Shomura A, Shimizu T, Lin SY, Inoue T, Fukuda A, Shimano T, Kuboki Y, Toyama T, Miyamoto Y, Kirihara T, Hayasaka K, Miyao A, Monna L, Zhong HS, Tamura Y, Wang ZX, Momma T, Umehara Y, Yano M, Sasaki T, Minobe Y (1994) A 300 kilobase interval genetic map of rice including 883 expressed sequences. Nat Genet 8:365–372

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Li J, Xiao J, Grandillo S, Jiang L, Wan Y, Deng Q, Yuan L, McCouch SR (2004) QTL detection for rice grain quality traits using an interspecific backcross population derived from cultivated Asian (*O. sativa* L.) and African (*O. glaberrima* S.) rice. Genome 47:697–704

Parisseaux B, Bernardo R (2004) In silico mapping of quantitative trait loci in maize. Theor Appl Genet 109:508–514

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67:170–181

Redona ED, Mackill DJ (1998) Quantitative trait locus analysis for rice panicle and grain characteristics. Theor Appl Genet 96:957–963

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res 67:175–186

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805–816

Sillanpää MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373–1388

Sillanpää MJ, Arjas E (1999) Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics 151:1605–1619

Sillanpää MJ, Bhattacharjee M (2005) Bayesian association-based fine mapping in small chromosomal segments. Genetics 169:427–439

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, Heidelberg

Tan YF, Xing YZ, Li JX, Yu SB, Xu CG, Zhang Q (2000) Genetic bases of appearance quality of rice grain characteristics. Theor Appl Genet 96:957–963

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) *Dwarf8* polymorphisms associate with variation with flowering time. Nat Genet 28:286–289

Uimari P, Hoeschele I (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics 146:735–743

Uimari P, Sillanpää MJ (2001) Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genet Epidemiol 21:224–242

Wan XY, Wan JM, Weng JF, Jiang L, Bi JC, Wang CM, Zhai HQ (2005) Stability of QTLs for rice grain dimension and endosperm chalkiness characteristics across eight environments. Theor Appl Genet 110:1334–1346

Wan XY, Wan JM, Jiang L, Wang JK, Zhai HQ, Weng JF, Wang HL, Lei CL, Wang JL, Zhang X, Cheng ZJ, Guo XP (2006) QTL analysis for rice grain length and fine mapping of an identified QTL with stable and major effects. Theor Appl Genet 112:1258–1270

Yi N (2004) A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. Genetics 167:967–975

Yi N, George V, Allison DB (2003) Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics 164:1129–1138

Yu J, Pressoir G, Briggs W, Bi, IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–207

Zhang Y-M, Mao Y, Xie C, Smith H, Luo L, Xu S (2005a) Mapping quantitative trait loci using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). Genetics 169:2267–2275

Zhang N, Xu Y, Akash M, McCouch S, Oard JH (2005b) Identification of candidate markers associated with agronomic traits in rice using discriminant analysis. Theor Appl Genet 110:721–729