Theor Appl Genet (2009) 118:865–880
DOI 10.1007/s00122-008-0945-6

ORIGINAL PAPER

Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms

Hiroyoshi Iwata · Kaworu Ebana · Shuichi Fukuoka · Jean-Luc Jannink · Takeshi Hayashi

Received: 18 September 2008 / Accepted: 2 December 2008 / Published online: 9 January 2009

Abstract Association mapping can be a powerful tool for detecting quantitative trait loci (QTLs) without requiring line-crossing experiments. We previously proposed a Bayesian approach for simultaneously mapping multiple QTLs by a regression method that directly incorporates estimates of the population structure. In the present study, we extended our method to analyze ordinal and censored traits, since both types of traits are common in the evaluation of germplasm collections. Ordinal-probit and tobit models were employed to analyze ordinal and censored traits, respectively. In both models, we postulated the existence of a latent continuous variable associated with the observable data, and we used a Markov-chain Monte Carlo algorithm to sample the latent variable and determine the model parameters. We evaluated the efficiency of our approach by using simulated- and real-trait analyses of a rice germplasm collection. Simulation analyses based on real marker data showed that our models could reduce both false-positive and false-negative rates in detecting QTLs to reasonable levels. Simulation analyses based on highly polymorphic marker data, which were generated by coalescent simulations, showed that our models could be applied to genotype data based on highly polymorphic marker systems, like simple sequence repeats. For the real traits, we analyzed heading date as a censored trait and amylose content and the shape of milled rice grains as ordinal traits. We found significant markers that may be linked to previously reported QTLs. Our approach will be useful for whole-genome association mapping of ordinal and censored traits in rice germplasm collections.

Communicated by M. Sillanpää.

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

K. Ebana · S. Fukuoka
QTL Genomics Research Center, National Institute of Agrobiological Sciences, Kannondai, Tsukuba, Ibaraki 305-8602, Japan

J.-L. Jannink
Plant, Soil and Nutrition Research Unit, Robert W. Holly Center for Agriculture and Health, USDA-ARS, Cornell Univ., Ithaca, NY 14853, USA

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

Introduction

Landraces and primitive cultivars preserved in gene banks generally show a broader range of phenotypic variation than do advanced cultivars. Among the polymorphisms left unused in advanced cultivars, there are many that would be valuable for future breeding programs (Hawks 1983). Utilization of such variation for further crop improvement will depend on a better understanding of the genetic basis for the phenotypic variation. Nowadays, association mapping (also known as linkage disequilibrium mapping) has emerged as a powerful tool for detecting loci or genes responsible for the phenotypic variation in crops (e.g., Yu and Buckler 2006; Oraguzie et al. 2007). Since association mapping can be done without requiring line-crossing experiments, it is suitable for directly detecting QTLs by using accessions in a germplasm collection. In gene banks,

the characteristics of germplasms are systematically evaluated for characteristics such as morphology, flowering phenology, disease resistance, and grain quality, and the results are stored in database systems. If these data can be analyzed by using DNA polymorphisms for the accessions, then the loci or genes responsible for various characteristics can be detected by association-mapping approaches.

The evaluation data for crop germplasms include a range of data types in addition to continuous data, such as *ordinal* and *censored* data. Because of the difficulty or laboriousness of quantitative evaluations, quantitative traits are sometimes measured in an ordinal manner. For example, morphological characteristics are often scored in several ordered categories on the basis of visual judgments. Similarly, the degree of disease resistance is often scored on the basis of the magnitude of the disease symptoms. Censored observations also sometimes arise when the range of measurements is limited. For example, flowering dates may be censored because of the limited duration of an evaluation, and root depth may be censored because of the limited depth of the pots used in the evaluation. For these types of data, the statistical methods that are used for more familiar continuous forms of data are not optimal because the normality assumption is violated (for all observations in ordinal data and for the censored observations in censored data); furthermore, information loss resulting from categorizing or censoring the continuous traits greatly reduces the statistical power of such methods. Therefore, association mapping for ordinal and censored data requires special methods to replace the methods used for continuous data.

Recently, various Bayesian methods based on the Markov-chain Monte Carlo (MCMC) algorithm have been developed for QTL mapping and association mapping (Satagopan et al. 1996; Uimari and Hoeschele 1997; Sillanpää and Arjas 1998, 1999; Uimari and Sillanpää 2001; Sillanpää et al. 2001; Yi et al. 2003; Yi 2004; Sillanpää and Bhattacharjee 2005; Yi et al. 2005; Iwata et al. 2007; Sillanpää and Hoti 2007; Huang et al. 2007). Bayesian mapping methods based on the MCMC algorithm can generally map multiple QTLs simultaneously; thus they can correctly estimate the number, locations, and genetic effects of QTLs for a complex trait governed by multiple QTLs. These methods also have the advantage of being extensible. For example, we can extend a model for continuous traits to models for binary, ordinal, and categorical traits by postulating a latent variable that underlies the generation of the binary, ordinal, or categorical responses (Yi and Xu 2000; Kilpikari and Sillanpää 2003; Yi et al. 2004; Sillanpää and Bhattacharjee 2006; Yi et al. 2007; Huang et al. 2007). A model for continuous traits can also be extended to models for censored traits by a latent variable approach (Sorensen et al. 1998; Sillanpää and Hoti 2007). That is, the mapping of multiple QTLs for various types of traits can be implemented through a latent variable approach based on the Bayesian mapping methods.

Asian cultivated rice, *Oryza sativa* L., is an important crop and staple food for half of the world's population. A complete and high-quality map-based sequence for rice (International Rice Genome Sequencing Project 2005) and other genomic resources such as mapped and annotated cDNA clones (Kikuchi et al. 2003) have paved the way for association mapping of rice at a whole-genome scale. Since rice is expected to have high levels of population structure because of self fertilization and the nature of its breeding history, association mapping of rice creates a potential problem: the presence of a population structure can mimic the signal from an association and lead to more false positives or to missed real effects (i.e., false negatives) (Lander and Schork 1994). A typical statistical model that deals with the influence of population structure includes a term for the DNA polymorphism itself and a term for the genetic background of the individual (Thornsberry et al. 2001; Yu et al. 2006; Malosetti et al. 2007; Jannink 2007; Zhao et al. 2007; Weber et al. 2007; Agrama et al. 2007). In our previous study (Iwata et al. 2007), we proposed an approach that combined a Bayesian method for mapping multiple QTLs with a model that deals with the influence of population structure, and we evaluated the efficiency of our approach in simulated- and real-trait analyses of a rice germplasm collection. Our results indicated that the multiple QTL model with population effect could reduce the incidence of both false positives and false negatives compared with single QTL models (with and without population effect) and a multiple QTL model without population effect, and suggested that association mapping has good prospects in rice if proper methods are adopted.

In the present study, we extended our previous method to deal with ordinal and censored data. We evaluated the efficiency of the method for analyzing the variation in a rice germplasm collection. We performed simulation analyses based on real marker data to assess the accuracy of the estimation. We also performed analyses of real trait data (i.e., heading date as a censored trait, and amylose content and the shape of milled rice grain on a five-point scale as ordinal traits), and we compared the results with QTLs previously reported for these traits. Finally, we discuss the prospects for the application of our approach to the association mapping of ordinal and censored traits in a rice germplasm collection.

Materials and methods

Plant materials

As in our previous study (Iwata et al. 2007), we used 332 rice accessions. The 332 rice accessions were selected as

representatives of the rice germplasm maintained in 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 in 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/). The population structure of the 332 accessions was inferred by means of model-based Bayesian clustering analysis (Pritchard et al. 2000) using the 179 RFLP markers as described in Iwata et al. (2007). In the analyses, we tested the admixture models with two to eight populations, and the model in which the number of populations (J) was six showed higher posterior probability 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} (Iwata et al. 2007). The differentiation among populations was so large that most accessions originated mainly from a single population (i.e., the maximum q_{ij} of each accession is larger than 0.9 in 278 out of 332 accessions). The **Q** matrix, the elements of which (i.e., q_{ij}) represent estimates of the genetic backgrounds of the accessions, was further incorporated into the statistical models used for association mapping.

Statistical models

For the censored data, we employed the tobit (censored regression) model. The tobit model supposes that there is a latent (i.e., unobservable) variable y_i^* that underlies the generation of observable censored response y_i for the ith sample ($i = 1, 2,\ldots, N$). The observable response y_i is defined to be equal to the latent variable y_i^* whenever the latent variable is below the specific threshold y_T, and equal to y_T otherwise. That is, when values greater than y_T are censored,

$$y_i = \begin{cases} y_T & \text{if } y_i^* > y_T \\ y_i^* & \text{if } y_i^* \le y_T \end{cases}. \quad (1)$$

It is also possible to equate the observable response to the latent variable when values less than y_T are censored, in which case the observable response is equal to the latent variable whenever the latent variable is above the threshold, and equal to the threshold otherwise.

For ordinal data, we employed the ordinal probit (cumulative probit) model. The ordinal probit model also supposes a latent (i.e., unobservable) variable y_i^* that underlies the generation of observable ordinal response y_i for the ith sample. That is, the value of each y_i^* falls into one of M contiguous bins on the real line demarcated by the cut-points $\kappa_0, \kappa_1,\ldots, \kappa_M$, and the observed values of y_i are determined by the following relationship:

$$y_i = m \quad \text{if } \kappa_{m-1} < y_1^* \le \kappa_m (m = 1, 2, \ldots, M). \quad (2)$$

Since the cut-points are also unobservable, the values of κ_m are sampled a posteriori, but the first, second, and last cut-points were fixed as $\kappa_0 = -\infty$, $\kappa_1 = 0$, and $\kappa_M = \infty$. When the number of bins is two, the model corresponds to the probit model for binary data.

The method employed here is based on a variable selection method developed by Kuo and Mallick (1998). In both the tobit and the ordinal probit models, we described the latent variables of individual i by the following linear model:

$$y_i^* = \sum_{j=1}^{J} q_{ij}\alpha_j + \sum_{k=1}^{K}\sum_{l=1}^{L_k} x_{ikl}\gamma_k\beta_{kl} + e_i, \quad (3)$$

where q_{ij} is the (i, j)th element of matrix **Q**; α_j the population effect associated with population j ($j = 1, 2,\ldots, J$); L_k the number of alleles of marker k ($k = 1, 2,\ldots, K$); x_{ikl}

Table 1 Setting for the QTL genotypes and effects used to generate the simulated datasets, and average variance of each QTL

	Simulated QTLs					
	1	2	3	4	5	6
QTL genotypes[a]						
NN	*QQ*	*QQ*	*QQ*	*qq*	*qq*	*qq*
KK	*qq*	*qq*	*qq*	*QQ*	*QQ*	*QQ*
Others	*qq*	*qq*	*qq*	*qq*	*qq*	*qq*
QTL effects						
QQ	1.00	0.75	0.50	1.00	0.75	0.50
qq	−1.00	−0.75	−0.50	−1.00	−0.75	−0.50
Average variance of each QTL	0.821	0.470	0.198	0.811	0.434	0.194

[a] NN, homozygous for the 'Nipponbare' allele in RFLP markers at the same positions; KK, homozygous for the 'Kasalath' allele in RFLP markers at the same positions; others, the other genotypes in RFLP markers at the same positions

denotes the genotypes at marker k for individual i and equals 1 if the genotype is homozygous for allele l ($l = 1, 2,\ldots, L_k$) and equals 0 otherwise; γ_k the indicator variable, and $\gamma_k = 1$ corresponds to the case in which the marker is included in the model as a QTL representative, and $\gamma_k = 0$ implies exclusion; β_{kl} the effect associated with the homozygous genotype of allele l for marker k; and e_i the residual error, which is assumed to follow $N(0, \sigma_e^2)$. Here, we considered only marker positions as putative QTLs in this model, although it is usually a false assumption in practice, as discussed below. Because rice is a species with a high degree of selfing, the dominance effect was not included. Although epistatic effects can theoretically be included in the model, we excluded them from our analysis for simplicity. The linear model in Eq. 3 is similar to the one used in our previous study (Iwata et al. 2007), but considers marker loci to be multi-allelic rather than bi-allelic.

In order to reduce the number of parameters sampled in the MCMC estimation, we fixed the genetic effect of the first allele of the marker k (i.e., β_{k1}) at 0. Now, let

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_K]$$

and

$$\boldsymbol{\eta} = \left[\gamma_1 \boldsymbol{\beta}_1^T, \gamma_2 \boldsymbol{\beta}_2^T, \ldots, \gamma_K \boldsymbol{\beta}_K^T\right]^T,$$

where $\mathbf{X}_k$ is an $N \times (L_k - 1)$ matrix whose (i, j)th element is $x_{ik(j+1)}$, and $\boldsymbol{\beta}_k$ is a column vector for the genetic effects of marker k without the effect of the first allele; that is, $\left(\beta_{k2}, \ldots, \beta_{kL_k}\right)^T$. Then, the matrix notation for the model in Eq. 3 is

$$\mathbf{y}^* = \mathbf{Q}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\eta} + \mathbf{e}, \tag{4}$$

where $\mathbf{y}^*$ is an $N \times 1$ vector whose ith element is y_i^*, $\boldsymbol{\alpha}$ is a $J \times 1$ vector whose jth element is α_j, and $\mathbf{e}$ is an $N \times 1$ vector whose ith element is e_i.

Bayesian estimation of parameters in the models was carried out on the basis of prior and posterior distributions of the parameters. The prior and posterior distributions are described in Appendix A. MCMC sampling was used for Bayesian inference about each of the parameters. The details of MCMC sampling adopted in this study are described in Appendix B.

Simulated datasets

Simulation A

As in our previous study (Iwata et al. 2007), we used the observed genotypes of 179 RFLP markers from the 332 rice accessions to generate the simulated datasets. The marker genotypes remained the same as those in the real data, and six QTLs were simulated at different positions randomly selected from the 179 RFLP markers. The genotypes of the QTLs were simulated according to the genotypes of the RFLP markers at the same positions. For half (i.e., three) of the QTLs, we simulated the QTL genotypes as *QQ* if the marker genotype was homozygous for the allele of 'Nipponbare' (a *japonica* cultivar), and as *qq* otherwise. For the other half of the QTLs, we simulated the QTL genotype as *QQ* if the marker genotype was homozygous for the allele of 'Kasalath' (an *indica* cultivar), and as *qq* otherwise (Table 1). The genetic effects were set as 1, 0.75, and 0.5 for the *QQ* genotype and -1, -0.75, and -0.5 for the *qq* genotype. We then simulated the genotypic values of the lines by summing up the genetic effects of the six QTLs. Next, we simulated a residual variance set at $\sigma_e^2 = 1$ to generate the value of the latent variable for all the accessions. We calculated the variance of the latent variable, $\sigma_{y^*}^2$, at this point. Next, we sampled the population effect α_j ($j = 1, 2, \ldots, J$) from $N(0, 0.25\sigma_{y^*}^2)$. We then added the population effect α_j to the value of the latent variable of accession i, weighted by q_{ij}. The proportion of the variance due to population effects was scaled as 20% (=0.25/1.25) to reflect population effects estimated from the real datasets in our previous study (Iwata et al. 2007). In this simulation, estimated population structure, i.e., q_{ij}, was used as a true value. In practice, however, the estimated population structure is also subject to estimation error, and the error may cause decline in statistical power. The process described above was repeated 100 times to generate 100 sets of values for the latent variable.

From the 100 sets of values for the latent variable, we generated 100 simulated datasets for each of three types of trait: continuous, censored, and ordinal traits. The phenotypic values of these traits were simulated according to the values of the latent variable. For a continuous trait, the phenotypic values were the same as the values of latent variables. For a censored trait, we set the threshold value y_T to equal the median value for the latent variable, and we simulated the phenotypic values according to the model in Eq. 1. For an ordinal trait, we used two sets of bin cut-points. In the first set, we chose cut-points at the 20, 40, 60, and 80 percentiles. In the second set, we chose cut-points at the 10, 20, 40, and 60 percentiles. Then, we converted the values of the latent variables into ordinal scores ranging from 1 (for the lowest bin) to 5 (for the highest bin) according to these bin cut-points. In the remainder of the text, we refer to the ordinal datasets based on the former and latter sets of cut-points as the *balanced* and *unbalanced* datasets, respectively.

Simulation B

To evaluate the performance of the methods in the case where genotypic data are obtained from highly polymorphic markers, like simple sequence repeats (SSRs), we generated another set of simulated datasets by using coalescent

simulation (Kingman 1982; Donnelly and Tavaré 1995). The simple demographic scenario used in the simulations was not intended to simulate the historical demography of rice populations but simply to simulate highly polymorphic datasets that have population stratification. We used SIMCOAL2 (Laval and Excoffier 2004) to simulate three populations with migration. The three populations had 500 haploid individuals each, and were diverged from a population of 500 haploid individuals 1,000 generations ago. The migration rate between the three populations was 0.001 per generation in both directions. Each individual had six chromosomes each carrying 51 SSR markers. The recombination rate between adjacent markers was 0.02 per generation. We also simulated 11 SSR markers that were not linked to other markers (i.e., recombination rate was 0.5) in order to simulate background genetic effect (i.e., genetic effect that cannot be captured by mapped markers). Mutation rate of the SSR markers was 0.002 per generation. From 1,500 simulated individuals, 300 individuals (100 for each population) were randomly selected for the subsequent process. We located four QTLs on the 20th markers of each of first four chromosomes. We also located 11 QTLs on the 11 unlinked markers to simulate genetic background effect. Genetic effects of the QTLs were simulated according to the genotypes of the SSR markers at the same positions. The genetic effect and a residual variance was sampled from $N(0, 1)$, and the sizes of the genetic effects were adjusted to make QTLs have specific heritability. The heritability of the four QTLs located on chromosomes and one QTL located on unlinked markers were 0.1, while the heritability of remaining QTL was 0.02. The phenotype of continuous data type was converted to the phenotype of censored and ordinal data in the same way as for simulated datasets A, except that we did not generate the unbalanced ordinal data in these datasets. We conducted Bayesian clustering analysis with the program Structure (Pritchard et al. 2000) with the admixture model in which the number of population was three. MCMC cycles for the Bayesian clustering analysis were repeated 1×10^5 times after 1×10^4 cycles of a burn-in period. The **Q** matrix estimated was further incorporated into the model of Bayesian association mapping. We generated 100 datasets and conducted the Bayesian association mapping in the same way as for simulated datasets A. The genotypes of all 306 linked markers were used in the Bayesian clustering analysis and the Bayesian association mapping.

Real datasets

As real data, we analyzed heading date as the censored trait, and glutinousness and the shape of milled rice grain on a five-point scale as the ordinal traits.

The seeds of all accessions were sown in seedling cases in the middle of April 2002 and transplanted into the experimental field at NIAS (Tsukuba, Ibaraki, Japan) in the middle of May. Heading date was measured as the number of days after sowing until the first panicle of each individual appeared under natural field conditions. The heading date data was averaged over 25 individuals for each accession. In this experiment, the duration of the evaluation was limited to 160 days. That is, heading dates later than 160 days were censored. The data of 69 out of 332 accessions (ca. 20%) were censored.

Amylose content (used as a measure of glutinousness) was measured by means of colorimetric analysis of the starch–iodine reaction detected with an Autoanalyzer II (Bran + Luebe). Three grains of each accession were degraded overnight with 2 ml 2 N KOH before measurement with the Autoanalyzer II. After calculating the apparent amylose content, we divided the data into five categories: waxy (0–3% amylose content), dull (3–10%), low (10–15%), medium (15–25%), and high (more than 25%). We then scored these categories on a five-point scale ranging from 1 (waxy) to 5 (high amylose content).

The shape of the milled rice grains was scored on a five-point scale ranging from 1 (broad) to 5 (narrow) on the basis of visual judgments. In the visual judgments, we used the digital images from our previous study (Iwata et al. 2007). In the previous study, 296 out of the 332 accessions were cultivated in an experimental field at NIAS in the 2003 cropping season, and six randomly selected milled rice grains from each accession were photographed with a digital camera.

Data analysis procedure

In the statistical models used here, the marker loci can be treated as multi-allelic. In our data, however, alleles from the Japanese 'Nipponbare' cultivar and the Indian 'Kasalath' cultivar dominated the 332 accessions at most loci (Kojima et al. 2005). Thus, in the association mapping analysis, alleles other than the 'Nipponbare' and 'Kasalath' alleles were merged into a single allele class. That is, we regarded the marker loci as bi-allelic (when only 'Nipponbare' and 'Kasalath' alleles were present) or tri-allelic (when there were alleles other than the 'Nipponbare' and 'Kasalath' alleles). Of the 179 RFLP loci, 51 were bi-allelic and 128 were tri-allelic.

For each dataset, MCMC cycles were repeated 13×10^4 times, and the first 3×10^4 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 we retained was 1×10^4.

This sampling scheme was based on the previously described evaluation of the convergence of MCMC cycles (Iwata et al. 2007).

We regarded a marker as "significant" when the mean of the posterior distribution of γ_k was larger than a specified threshold of 0.5. 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 more than 50% of the MCMC samples. This threshold corresponds to the "moderate" threshold in our previous study (Iwata et al. 2007).

To evaluate the performance (i.e., the estimation accuracy) of our approach, we calculated the false-negative rate (FNR), the false-positive rate (FPR), and the false-discovery rate (FDR). FNR is the proportion of total loci mistakenly regarded as non-significant when in fact they were QTLs. FPR is the proportion of total loci mistakenly regarded as significant when in fact they were not QTLs. FDR is the ratio of FPR to all significant loci. The equations that define these indices can be found in Iwata et al. (2007). In order to show a tradeoff relationship between FPR and true-positive rate (TPR; 1-FNR), a receiver operating characteristic (ROC) curve (Zweig and Campbell 1993) was obtained based on TPR, FPR and FDR. TPR, FPR and FDR were calculated for 1,001 different levels (0, 0.001,…, 1) of the γ_k threshold. For each type of trait, TPR, FPR and FDR were averaged over all datasets contained in each of simulations A and B, and ROC curves were drawn based on the averaged TPR, FPR and FDR.

For the datasets of simulation A, in order to assess the accuracy of estimation of the genetic effect of a QTL, we compared the estimated and true values and calculated the root-mean-square error (RMSE) between the estimated and true values, scaled using the true value:

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

where N is the number of simulated datasets and D the number of true positives over all simulated datasets. $\delta_{k,i}$ an indicator variable in which a value of 1 corresponds to the case in which the kth locus is a true positive in the ith dataset, and a value of 0 corresponds to the remainder of the possibilities. Since we fixed the genetic effect of the 'Nipponbare' allele of all markers at 0 in our statistical model (see "Statistical models"), we compared the estimated and true values on the basis of the differences in genetic effects between the 'Nipponbare' and 'Kasalath' alleles. That is, $\hat{\zeta}_{k,i}$ and $\zeta_{k,i}$ are the estimated and true values of the difference in the genetic effects of the 'Nipponbare' and 'Kasalath' alleles, respectively. In ordinal data, the genetic effect of a QTL was estimated on a scale relative to the residual variance, since the residual variance was fixed at 1 in the statistical model for ordinal data (see "Statistical models"). However, we could directly compare the estimated and true values of the genetic effects even for ordinal data, because we set the residual variance at 1 in our simulation study.

Results

Simulation A

The proportion of the phenotypic variance explained by the simulated QTLs (i.e., heritability) in the datasets of simulation A was calculated for the continuous data. In the 100 datasets, the average heritability of each QTL was 0.150 and the average joint heritability of all six QTLs was 0.540 (Fig. 1). The average heritability of each QTL was 0.248, 0.140, and 0.061, respectively, for QTLs that had large (i.e., 1 or −1), moderate (i.e., 0.75 or −0.75), and small (i.e., 0.5 or −0.5) effects, respectively.

The histograms in Fig. 2 show the number of datasets in simulation A (out of 100) that fell into specified intervals for FNR, FPR, and FDR. For FNR, the censored data tended to show larger values than the continuous and ordinal data (Fig. 2a). The average FNR was smallest in the continuous data, followed by the balanced ordinal, unbalanced ordinal, and censored data. The average FNR

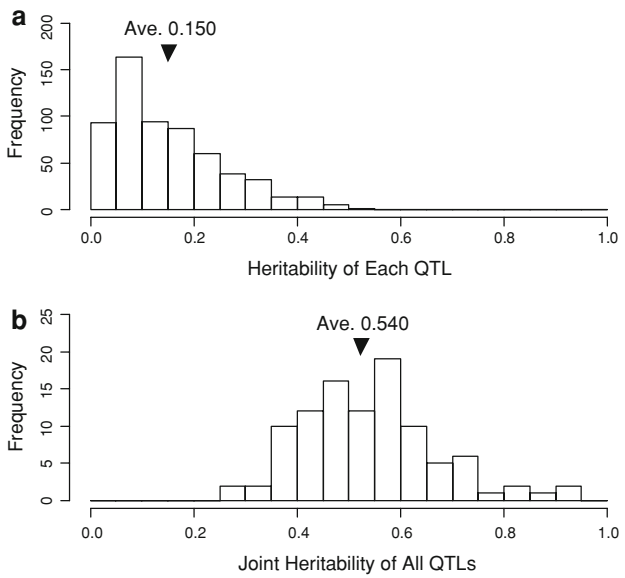

Fig. 1 **a** Heritability of each QTL. The proportion of phenotypic variance explained by each QTL (i.e., heritability of each QTL) was calculated over 600 QTLs (i.e., 6 QTLs × 100 simulated datasets) for continuous data. **b** Joint heritability of all QTLs. The proportion of phenotypic variance explained by all QTLs (i.e., the joint heritability of all QTLs) was calculated over 100 simulated datasets for continuous data

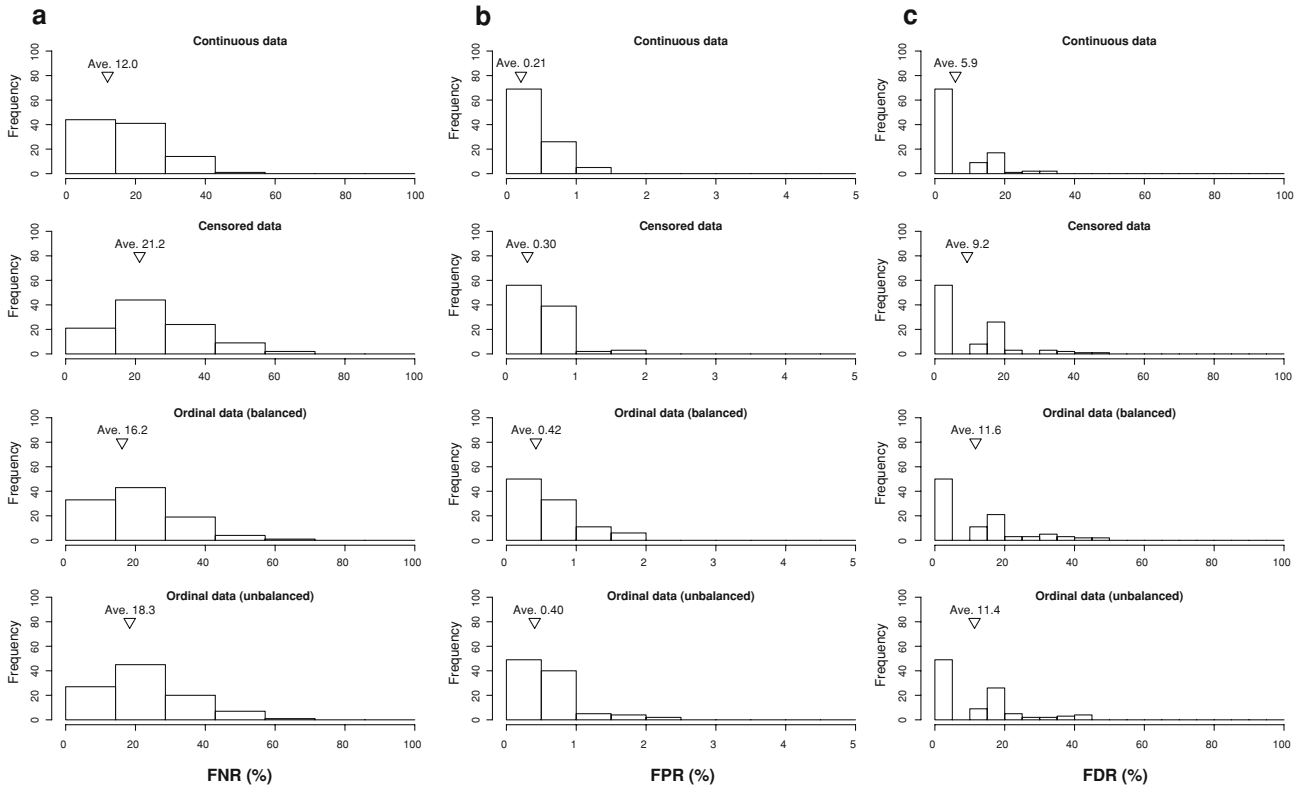

Fig. 2 Histograms for (**a**) the false-negative rate (FNR), (**b**) the false-positive rate (FPR), and (**c**) the false-discovery rate (FDR), obtained from 100 simulated datasets A for each data type. The simulated datasets for continuous data were first generated using real marker data, and were then converted into the datasets for censored and ordinal data. For the ordinal data conversion, we generated two different sets (i.e., balanced and unbalanced) of data. For details, see the text

Table 2 False-negative rates for the three QTL effect sizes

Size of QTL effect	FNR(%)			
	Continuous data	Censored data	Balanced ordinal data	Unbalanced ordinal data
Small (0.5, −0.5)	27.5	45.0	34.5	40.5
Middle (0.75, −0.75)	5.0	12.5	9.0	9.5
Large (1.0, −1.0)	3.5	6.0	5.0	5.0

was larger for QTLs that had smaller effects (Table 2). In particular, the average FNR for the QTLs that had small effects reached 45% in the censored data, whereas the corresponding average FNR for the QTLs that had large effects was 6%.

For FPR, the ordinal data tended to show larger values than the continuous and censored data (Fig. 2b). The average FPR was smallest in the continuous data, followed by the censored, unbalanced ordinal, and balanced ordinal data. In all data types, FPR was less than 1% in most datasets. The proportions of datasets in which no false positives were observed (i.e., in which FPR = 0%) were 0.69, 0.56, 0.50, and 0.49 in the continuous, censored, balanced ordinal, and unbalanced ordinal data, respectively.

For FDR, the ordinal data tended to show larger values than the continuous and censored data (Fig. 2c). The average FDR was smallest in the continuous data, followed by the censored, unbalanced ordinal, and balanced ordinal data. In all data types, FDR was less than 20% in most datasets.

ROC curves of four different data types were shown in Fig. 3. In Fig. 3b, we also drew ROC curves based on TPR and FDR in addition to normal ROC curves (i.e., ROC curves based on TPR and FPR). In these graphs, perfect discrimination (zero false negatives and zero false positives, or zero false negatives and zero false discovery) corresponds to a point at the upper left corner of each graph, and the closer the ROC curve is to the upper left corner, the higher the performance of detection methods. In

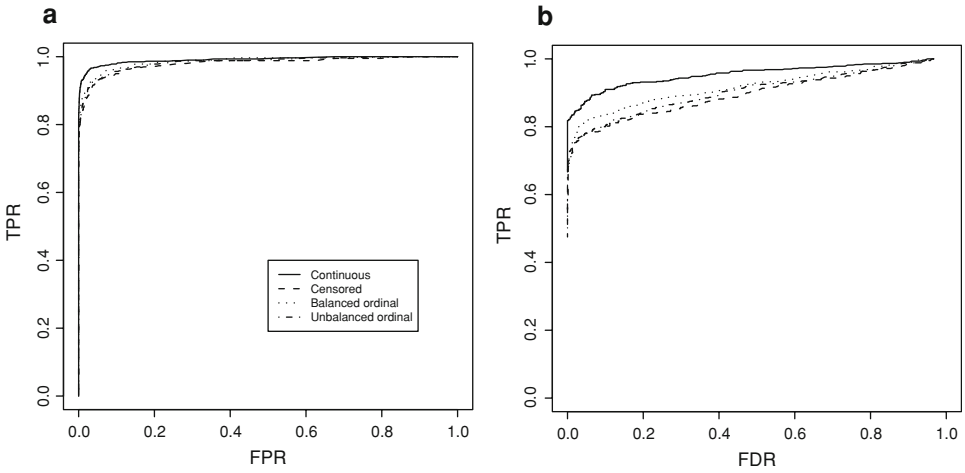

Fig. 3 Receiver operating characteristic (ROC) curves of QTL detection in association mapping for four different data types in simulated datasets A. True positive rate [TPR; 1-false negative rate (FNR)] were plotted against (**a**) false positive rate (FPR) and (**b**) false discovery rate (FDR). TPR, FPR and FDR were obtained for the different levels of threshold γ_k and averaged over 100 datasets of each data type. For details, see the text

Table 3 Root-mean-square error values between the estimated and true values of QTL effects

	RMSE[a] (%)
Continuous data	19.8
Censored data	24.6
Balanced ordinal data	22.8
Unbalanced ordinal data	25.4

[a] RMSE between the estimated and true values of genetic effects of true-positive QTLs, scaled using the true values

both Fig. 3a and b, the curves were closest to the upper left corner in the continuous data, followed by the balanced ordinal, unbalanced ordinal, and censored data, indicating the highest performance in QTL detection for these data types in this order.

To evaluate the accuracy of our estimates of the genetic effects, we also calculated the RMSE for correctly detected QTLs (i.e., for true positives). RMSE was smallest in the continuous data, followed by the balanced ordinal, censored, and unbalanced ordinal data (Table 3), but all values were less than 26%.

Posterior averages of the number of 1 s in γ, i.e., the number of QTLs included in the model (N_Q), showed a tendency to be larger in censored and ordinal data than continuous data (Fig. 4). In all data types, the posterior averages of the number of QTLs were larger than the number of simulated QTLs, i.e., 6.

Simulation B

The number of alleles of simulated markers ranged from 3 to 14 and its average was 6.6. Nei's gene diversity (Nei 1973) of the simulated markers ranged from 0.21 to 0.88 and its average was 0.72.

The histograms in Fig. 5 show the number of datasets of simulation B (out of 100) that fell into specified intervals for FNR, FPR, and FDR. Histograms of FNR, FPR and FDR showed the same tendency with ones obtained in the datasets of simulation A. That is, the average FNR was smallest in the continuous data, followed by the ordinal, and censored data. The average FPR was smallest in the continuous data, followed by the censored, and ordinal data. The average FDR was smallest in the continuous data, followed by the censored, and ordinal data.

ROC curves of three different data types also showed the same tendency with ones obtained in the simulation A (data not shown). That is, curves were closest to the upper left corner in the continuous data, followed by the ordinal and censored data in both ROC curves for TPR versus FPR and TPR versus FDR. Posterior averages of the number of QTLs included in the model also showed the same tendency with ones obtained in the simulation A (data not shown). That is, the posterior averages were larger in censored and ordinal data than continuous.

Real data analysis

We found five, five, and six significant markers for heading date, amylose content, and the shape of milled rice grains, respectively (Table 4). The posterior average of the number of QTLs included in the model (N_Q) was 11.5, 13.7 and 12.5 for heading date, amylose content, and the shape of milled rice grains, respectively. With a stricter threshold (0.9), we found two significant markers for each trait. One significant marker (i.e., R2869) was observed for both heading date and amylose content. One overlap (i.e., C962) was observed between amylose content and the shape of milled rice grains. There was no overlap between heading date and grain shape.

For the heading date, the estimated effects of the 'Kasalath' allele for the significant markers were all negative except for the R3226 marker, indicating that QTL allele link to the 'Kasalath' allele generally shortened the

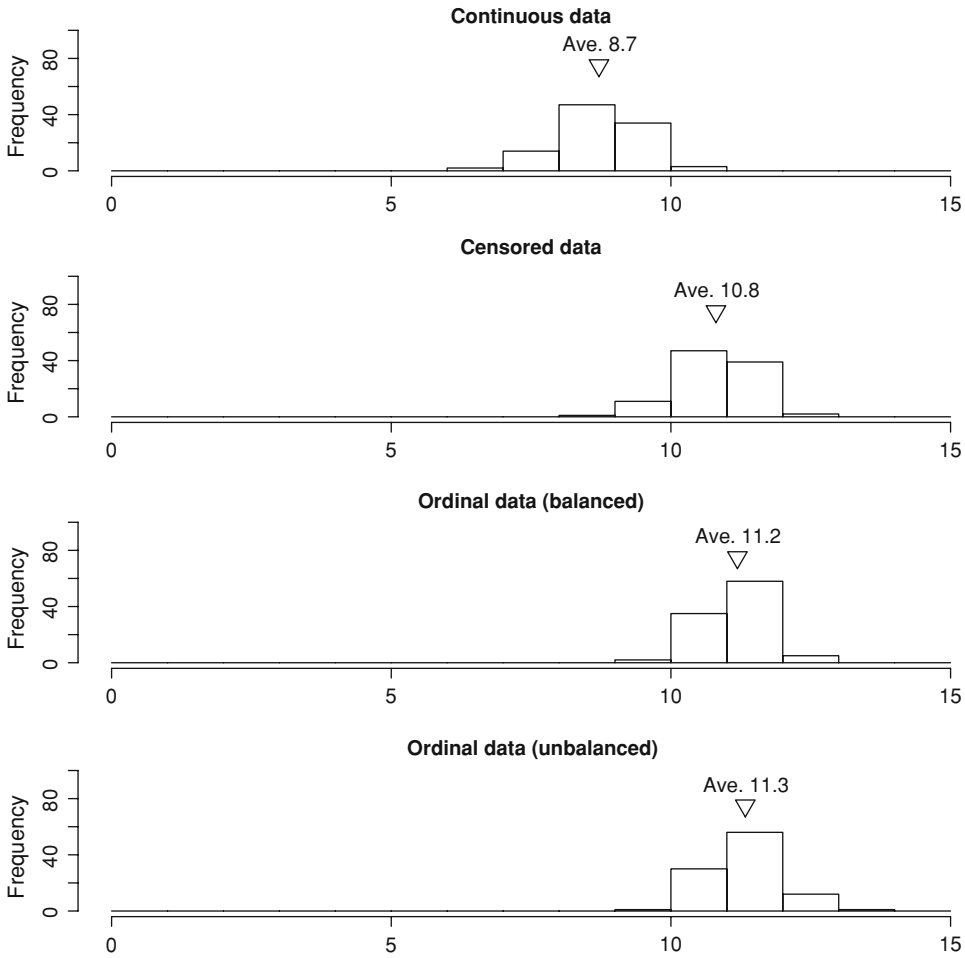

Fig. 4 Histograms of the posterior average of the number of 1 s in γ, i.e., the number of QTLs included in the model, for each data type in datasets A. The number of simulated QTLs was six in these datasets

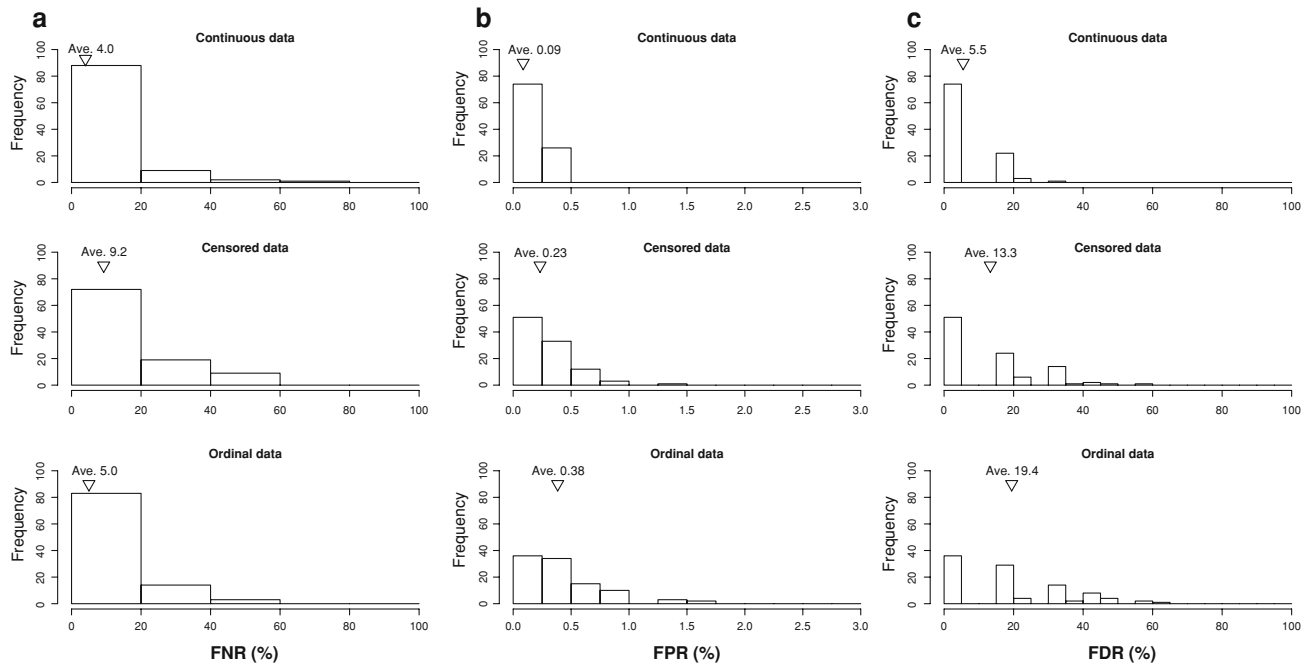

Fig. 5 Histograms for (**a**) the false-negative rate (FNR), (**b**) the false-positive rate (FPR), and (**c**) the false-discovery rate (FDR), obtained from 100 simulated datasets B for each data type

Table 4 Locations and estimated parameters for significant markers in the real datasets

Trait	Chromosome	Location[a]	Marker[b]	γ_k^c	β_{kl}^d
HD	1	87.7	G302*	0.916	−4.02 ± 4.09
	3	20.1	R3226	0.629	12.08 ± 8.68
	6	2.2	R2869	0.790	−11.12 ± 3.58
	6	34.8	R2147*	0.941	−1.00 ± 4.48
	12	3.0	R1684	0.821	−11.63 ± 3.44
AC	4	72.9	C891*	0.988	1.13 ± 0.27
	5	12.9	R830	0.526	−0.30 ± 0.34
	5	117.2	R521	0.664	−1.17 ± 0.36
	6	2.2	R2869*	0.993	1.68 ± 0.45
	6	117.7	C962	0.631	1.19 ± 0.60
GS	3	8.1	R1925*	0.943	0.06 ± 0.34
	3	68.6	R250*	1.000	1.54 ± 0.33
	4	0.6	C107	0.547	0.89 ± 0.35
	5	45.0	R569	0.587	0.60 ± 0.19
	6	117.7	C962	0.728	0.64 ± 0.26
	11	20.7	G1465	0.894	0.80 ± 0.23

HD heading date, *AC* amylose content, *GS* shape of milled rice grains

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

[b] Markers that were significant with strict (i.e., 0.9) thresholds are indicated with an asterisks

[c] Mean of the posterior distribution of γ_k

[d] Mean and SD of the posterior distribution of β_{kl} for the 'Kasalath' allele. In the calculation of mean and SD, we accounted only for MCMC samples in which $\gamma_k = 1$

number of days to heading. For the amylose content, the signs of the estimated effect of the 'Kasalath' allele differed among the significant markers, with no obvious pattern. However, for the highly significant markers (i.e., C891 and R2869), the effects of the 'Kasalath' allele were positive. For the shape of the milled rice grains, the estimated effects of the 'Kasalath' allele were all positive, indicating that the linkage of the alleles to the 'Kasalath' allele was responsible for narrower rice grains.

In analyzing the real data, we used a Poisson prior with mean $\lambda = 1$ on the number of QTL (i.e., the number of 1's in $\boldsymbol{\gamma}$; see Appendix A for details), because we thought that the marker density in this study was not enough to capture many QTLs segregating in the accessions. To evaluate the influence of this hyperparameter setting, we analyzed real data also with $\lambda = 10$. As a result, the posterior averages of γ_k were highly correlated between these two settings (i.e., correlations ranged from 0.974 to 0.995), although the posterior averages of γ_k were significantly larger in $\lambda = 10$ than in $\lambda = 1$ ($P < 0.001$ with the Mann–Whitney U test). Significant markers corresponded between these settings in heading date and the shape of the milled rice grain. In amylose content, six markers were significant when $\lambda = 10$, five of which were also significant when $\lambda = 1$. These analyses suggest that the mean of the Poisson prior does not strongly affect the results of analyses at least in the range from 1 to 10, although the significance of markers whose posterior average of γ_k is around 0.5 may be affected by the mean of the Poisson prior.

Discussion

Combined with the latent variable approach, the Bayesian multilocus association mapping method can be an efficient way to detect true associations between DNA polymorphisms on the one hand and censored or ordinal traits on the other. In our simulation studies, the FNR, FPR, and FDR values were larger in the censored and ordinal data than in the continuous data. The difference in these rates between different data types, however, was less obvious, and all the rates could be controlled at a practical level in censored and ordinal data as well as in continuous data by using the Bayesian method. In association mapping, it is important to reduce the number of false positives caused by the population structure. In our simulation studies, FPR was less than 1% in most datasets in all data types, indicating that the method successfully dealt with the effect of population structure. Although the estimation accuracy for the genetic effects of QTLs was worse in the censored and ordinal data than in the continuous data (Table 2), the RMSE was less than 26% even in the worst case (i.e., the case for the unbalanced ordinal data; Table 3), indicating that the estimation accuracy was also controlled at a practical level. It is noteworthy that we can estimate the genetic effects of QTLs for ordinal data as well as for censored and continuous data, even though we cannot directly observe continuous phenotypic variation hidden behind the scoring phenotype.

The linear model used in this study (i.e., the model in Eq. 3) is similar to the one used in our previous study (Iwata et al. 2007), but considers marker loci to be multiallelic rather than bi-allelic. To evaluate the power of this multiallelic model in the case where genotype data are highly polymorphic (i.e., large number of alleles), we conducted simulation analyses based on genotype data generated from coalescent simulation rather than on the real marker data of the rice germplasm. The results of the simulations (i.e., simulation B) were comparable to ones obtained in the simulations based on the real marker data (i.e., simulation A), indicating that our models could be applied to genotypic data based on highly polymorphic marker systems, like SSRs. In the simulations, however, we assigned a different allelic effect to each marker allele, although this condition is unrealistic in practice. When linkage disequilibrium between QTL alleles and marker

alleles is incomplete, the statistical power of the methods will become lower than obtained in the simulation study. The simulation study was also based on unrealistic demographical settings as described in "Materials and methods". To investigate the power of the analysis more concisely, a more detailed simulation study like Meuwissen et al. (2001) may be necessary.

The size of the genetic effects of the QTLs is expected to affect their FNR. In this study, we simulated three different sizes of QTL effects. The results showed that the FNR of QTLs with small effects showed large values (i.e., 34–45%) in the censored and ordinal data, whereas the FNR for QTLs with large effects were much lower (i.e., 5–6%). To clarify the relationship between heritability and the successful detection of a QTL, we applied logistic regression to this relationship. The resulting contribution of heritability to successful detection of a QTL was highly significant ($P < 0.0001$) in all data types. On the basis of the estimated regression equations, the probability that a QTL will be detected exceeds 0.9 when its heritability is 0.2 or higher (Fig. 6). Thus, our method can detect most of the major QTLs in both censored and ordinal data, although the method will miss a large proportion of the intermediate to minor QTLs. However, it must be noted that major QTLs can also be missed if the linkage disequilibrium between genetic markers and QTLs is not high.

The statistical models used in this study contain population effects (i.e., α_j) as well as the effects of markers distributed genome-wide (i.e., β_{kl}). Although the population effects are expected to absorb effects caused by

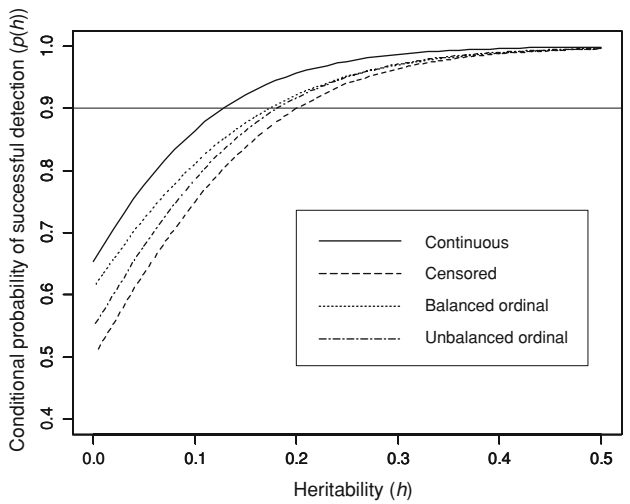

Fig. 6 Conditional probability [$p(h)$] that a true QTL with a given heritability (h) would be successfully detected in the simulation study. The probability calculation was based on the estimated logistic regression equations for successful detection of QTLs as a function of the heritability of the QTLs. We estimated the probability separately for each of four different data types (i.e., continuous, censored, and balanced and unbalanced ordinal)

population stratification, it may possibly cause false negatives. That is, if the distribution of a causal allele correlates strongly with estimated population structure (i.e., **Q** matrix), the effect of the QTL may be absorbed in the population effects and then the QTL will not be detected. When QTL effects are collinear with population structure, they are more likely to be absorbed by population than by marker effects because the former are fixed (not shrunken to zero) while the latter are random (and are shrunken to zero).

We also investigated the impact of biasing the score distribution in the ordinal data by comparing estimation accuracy between balanced and unbalanced ordinal data. In this comparison, the unbalanced ordinal data tended to show larger values for FNR than the balanced ordinal data, whereas the opposite tendency was observed for FPR and FDR (Fig. 2). A paired t test showed that the difference between the balanced and unbalanced ordinal data was significant for FNR ($P = 0.032$) but was not significant for FPR ($P = 0.765$) and FDR ($P = 0.843$). Thus, biasing the score distribution in ordinal data may decrease the statistical power of this approach for detecting QTLs, although it does not result in detecting non-functional spurious associations. In the evaluation of crop germplasms, grading the characteristics of accessions is an efficient way to measure a large number of samples. Our results indicate that the evenness of the score distribution (i.e., equal frequency of accessions among the score classes) should be taken into account when the grading system is determined to increase the statistical power for detecting QTLs.

In this study, we assumed that genotype data were available for censored individuals. In practice, however, genotype data may not be available for censored individuals. For example, we may genotype only individuals showing extreme phenotype (i.e., selective genotyping) or only susceptible or survived individuals. If genotype data are also missing for censored individuals, a more advanced modeling scheme such as Sillanpää and Hoti (2007) is necessary.

In Bayesian QTL mapping, reversible-jump MCMC algorithm (Green 1995) has been used as an almost routine tool (Yi 2004). The reversible-jump procedure can move between models of different dimension so that it can evaluate models of unknown number of QTLs. Although the method has flexibility in applications, it is sometimes subject to poor mixing and slow convergence (Yi 2004). In this study, we used a Bayesian variable selection method proposed by Kuo and Mallick (1998). The method is similar to, but simpler than, the stochastic search variable selection (SSVS) method developed by George and McCulloch (1993), which has been utilized in multiple QTL mapping (e.g., Yi et al. 2003). In Bayesian variable selection methods, the parameter space has constant dimensionality and hence

has the advantage that a Gibbs sampler can be applied directly without concern for the varying dimension aspects caused by the uncertainty in the number of QTLs (Godsill 2001; Yi 2004). In these methods, a varying dimensional parameter space is augmented to a fixed dimensional space to achieve the constant dimensionality. In this study, we sampled parameters $\boldsymbol{\beta}_k$ even when $\gamma_k = 0$ (i.e., $\boldsymbol{\beta}_k$ was not used in the model) for achieving the constant dimensionality. The algorithms proposed by Yi et al. (2005) and Yi et al. (2007) enable us to omit the sampling of $\boldsymbol{\beta}_k$ when $\gamma_k = 0$, and thus they have computational advantages over the algorithm used in this study. Fast and efficient algorithms must be considered when the number of markers gets large. For discussions about relationship between reversible-jump MCMC and Bayesian variable selection methods, see Godsill (2001) and Yi (2004).

We analyzed the heading date, amylose content, and shape of milled rice grain as real data, and found five, five, and six significant markers, respectively. Most of these markers may be linked genes and QTLs that have been previously reported. For heading date, the significant marker R2147 on chromosome 6 may be linked to the photoperiod sensitivity gene *Hd1*, which is a rice ortholog of the *Arabidopsis CONSTANS* gene (Yano et al. 2000). Marker R2869 on chromosome 6 may be linked to the *Hd3a* gene, which is a rice ortholog of the *Arabidopsis FT* gene (Kojima et al. 2002). Marker R3226 on chromosome 3 may be linked to the heading date QTL *Hd6* (Yamamoto et al. 2000). Markers G302 and R1684 may also be linked to previously reported QTLs (Li et al. 2003; Mei et al. 2003).

For amylose content, marker R2869 on chromosome 6 may be linked to the QTL that controls α-amylase activity (Cui et al. 2002). The *Waxy* gene, which controls amylose synthesis in both the endosperm and the pollen of cereal crops, also resides in the region near this marker (e.g., Yamanaka et al. 2004). Marker C891 on chromosome 4, marker R521 on chromosome 5, and marker C962 on chromosome 6 may be linked to the amylose-content QTLs reported by Lanceras et al. (2000), He et al. (1999), and Li et al. (2004), respectively.

For the shape of the milled rice grain, three out of the six significant markers (i.e., R250, R569, and G1465) correspond to the same markers in our previous study, (Iwata et al. 2007), which were highly significant for rice grain length. As described by Iwata et al. (2007), marker R569 on chromosome 5 may be linked to a previously reported QTL for grain length and length–width ratio (Wan et al. 2005). Marker R250 on chromosome 3 may also be linked to a QTL for grain length that has been detected around the centromeric region of chromosome 3 (Huang et al. 1997).

The correspondences between significant markers on the one hand and previously reported genes and QTLs on the other indicate the practical efficiency of our approach. Some of the pairs listed above (i.e., pairs between significant markers detected in this study and previously reported genes and QTLs), however, show a discrepancy in the estimated location. For example, the location of R2147 shows a 9 Mb deviation from the genomic region of *Hd1*. In our statistical model, we considered only marker positions as putative QTLs. In other words, we assumed complete linkage disequilibrium between marker alleles and QTL alleles in our model. In practice, however, linkage disequilibrium between marker alleles and QTL alleles is incomplete, with rare exceptions. Incomplete linkage disequilibrium between marker alleles and QTL alleles may cause the discrepancy between actual and estimated locations. In the future, the resolving power of the location estimation in our approach may be improved by using high-density markers such as genome-wide single-nucleotide polymorphism (SNP) markers.

In the evaluation of crop germplasms, we often encounter censored and ordinal data, since labor- and time-saving measurements are necessary to permit routine evaluation of a large number of accessions for various characteristics. Our results suggest that it will be practical to conduct association mapping of these types of data in order to detect novel QTLs. Our approach will be useful for detecting QTLs of various traits by using a crop germplasm collection and, thus, it will permit more efficient use of the data resources stored in crop germplasm databases in the future.

We have written computer programs in Java language to implement the proposed method. The programs are available from the senior author upon request.

Acknowledgments This work was supported by a grant from the Ministry of Agriculture, Forestry and Fisheries of Japan (Green Technology Project, QT1001, and Genomics for Agricultural Innovation, DD-4050).

Appendix A: Priors and posteriors

We considered the prior distributions of the parameters in the model in Eq. 4 except $\boldsymbol{\gamma}$ as follows:

$$\alpha_j \sim U(-\infty, \infty),$$
$$\boldsymbol{\beta}_k \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{\beta_k}),$$
$$\sigma^2_{\beta_k} \sim \upsilon_\beta s^2_\beta \chi^{-2}_{\upsilon_\beta},$$

and

$$\sigma^2_e \sim \upsilon_e s^2_e \chi^{-2}_{\upsilon_e},$$

where υ_β, s^2_β, υ_e, and s^2_e are hyperparameters for the distributions. That is, $\sigma^2_{\beta_k}$ was sampled from a scaled inverted chi-square distribution with υ_β degree of freedom and scale parameter s^2_β, and σ^2_e from the same distribution with different parameters (i.e., υ_e and s^2_e).

We considered that the prior distribution of the number of 1 s in $\boldsymbol{\gamma}$, i.e., the number of QTLs included in the model (N_Q), follows a truncated Poisson distribution. That is,

$$N_Q = \sum_{k=1}^{K} \gamma_k$$

and

$$p(N_Q = n) = \begin{cases} \frac{\lambda^n \exp(-\lambda)}{n!} \Big/ \sum_{i=1}^{Q_{\max}} \frac{\lambda^i \exp(-\lambda)}{i!} & \text{if } n \le Q_{\max} \\ 0 & \text{if } n > Q_{\max} \end{cases},$$

where λ is a hyperparameter for the distribution and is construed as the expected number of QTLs included in the model, and $Q_{\max}$ is a hyperparameter that determines the upper limit of the number of QTLs that can be included in the model. The Poisson prior on the number of 1 s in $\boldsymbol{\gamma}$, has been proposed by Yi (2004).

The prior of the cut-points in the ordinal probit model in Eq. 2, i.e., κ_m ($m = 2, 3, \ldots, M-1$), has a uniform distribution that respects the ordering constraints. That is,

$$\kappa_m | \kappa_{m-1}, \kappa_{m+1} \sim U[\kappa_{m-1}, \kappa_{m+1}].$$

Bayesian implementations of the tobit model are based on data augmentation, and they have been applied to statistical genetic analysis for censored traits (Sorensen et al. 1998; Sillanpää and Hoti 2007). In the tobit model in Eq. 1, y_i^* values are not observed if y_i^* is greater than or equal to the threshold y_T, although their values are necessary for estimating parameters in the model in Eq. 3. The unobserved y_i^* values are augmented by values sampled from the fully conditional posterior distribution:

$$y_i^* | \boldsymbol{\alpha}, \boldsymbol{\eta}, \sigma_e^2 \sim TN_{[y_T,\infty)}(\mathbf{q}_i\boldsymbol{\alpha} + \mathbf{x}_i\boldsymbol{\eta}, \sigma_e^2) \quad \text{(A1)}$$

where $TN_{[a,b)}(\mu, \sigma^2)$ is a normal distribution $N(\mu, \sigma^2)$ truncated to $[a, b)$, $\mathbf{q}_i$ is the ith row of matrix $\mathbf{Q}$, and $\mathbf{x}_i$ is the ith row of matrix $\mathbf{X}$. The sampling of unobserved y_i^* values was conducted repeatedly in the MCMC sampling procedure, as described in the next section.

Bayesian implementations of the ordinal probit model based on data augmentation were presented by Albert and Chib (1993), and they have been applied to multi-locus QTL mapping for ordinal traits (Yi et al. 2004, 2007). In the model, y_i^* in Eq. 3 is not observed for all samples, and the variance of y_i^* is assumed to be 1 for consistency with the standard normal c.d.f. link function (Cowles 1996). Thus, the y_i^* values are augmented by values sampled from the fully conditional posterior distribution:

$$y_i^* | \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\kappa}, y_i \sim TN_{(\kappa_{y_i-1}, \kappa_{y_i}]}(\boldsymbol{q}_i\boldsymbol{\alpha} + \mathbf{x}_i\boldsymbol{\eta}, 1), \quad \text{(A2)}$$

where $\boldsymbol{\kappa}$ is a vector for the bin cut-points, i.e., $(\kappa_0, \kappa_1, \ldots, \kappa_M)^T$. The fully conditional posterior distribution of κ_m is uniform, as follows:

$$\kappa_m | \mathbf{y}, \mathbf{y}^*, \kappa_{m-1}, \kappa_{m+1} \sim U[\max(y_i^* | y_i = m, \kappa_{m-1}), \min(y_i^* | y_i = m+1, \kappa_{m+1})]. \quad \text{(A3)}$$

Here, let $\mathbf{X}^* = [\gamma_1\mathbf{X}_1, \gamma_2\mathbf{X}_2, \ldots, \gamma_K\mathbf{X}_K]$. Then, Eq. 4 can be rewritten as:

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

where $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_K^T\right]^T$. Here, let

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

where

$$\mathbf{W} = [\mathbf{Q}, \mathbf{X}^*] \quad \text{(A4)}$$

and $\boldsymbol{\theta} = \left[\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T\right]^T$, and let

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2/\sigma_{\beta_1}^2 & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}\sigma_e^2/\sigma_{\beta_K}^2 \end{bmatrix}, \quad \text{(A5)}$$

$$\mathbf{C} = \mathbf{W}^T\mathbf{W} + \boldsymbol{\Sigma}, \quad \text{(A6)}$$

and

$$\mathbf{r} = \mathbf{W}^T\mathbf{y}^*. \quad \text{(A7)}$$

Then, the conditional posterior distribution of the ith element of $\boldsymbol{\theta}$ is:

$$\theta_i | \boldsymbol{\theta}_{-i}, \boldsymbol{\gamma}, \sigma_e^2, \mathbf{y}^* \sim N\left(\tilde{\theta}_i, \sigma_e^2/c_{i,i}\right), \quad \text{(A8)}$$

where $\boldsymbol{\gamma}$ is a vector whose kth element is γ_k, $\tilde{\theta}_i = (r_i - \mathbf{C}_{i,-i}\boldsymbol{\theta}_{-i})/c_{i,i}$, is the ith diagonal element of the matrix $\mathbf{C}$, r_i is the ith element of vector $\mathbf{r}$, $\mathbf{C}_{i,-i}$ is a row vector obtained by deleting element i from the ith row of the matrix $\mathbf{C}$, and $\boldsymbol{\theta}_{-i}$ is a vector obtained by deleting element i from the vector $\boldsymbol{\theta}$ (Sorensen and Gianola 2002, p. 566).

The fully conditional posterior distribution of $\sigma_{\beta_k}^2$ is given by

$$\sigma_{\beta_k}^2 | \boldsymbol{\beta}_k \sim \tilde{\upsilon}_{\beta_k}\tilde{s}_{\beta_k}^2\tilde{\chi}_{\tilde{\upsilon}_{\beta_k}}^{-2}, \quad \text{(A9)}$$

where $\boldsymbol{\beta}_k$ is a column vector for the genetic effects of marker k, i.e., $(\beta_{k2}, \ldots, \beta_{kL_k})^T$, $\tilde{\upsilon}_{\beta_k} = L_k + \upsilon_\beta - 1$, and $\tilde{s}_{\beta_k}^2 = \left[\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k + \upsilon_\beta s_\beta^2\right]/\tilde{\upsilon}_{\beta_k}$.

For the tobit model, the fully conditional posterior distribution of σ_e^2 is given by:

$$\sigma_e^2 | \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y}^* \sim \tilde{\upsilon}_e\tilde{s}_e^2\tilde{\chi}_{\tilde{\upsilon}_e}^{-2}, \quad \text{(A10)}$$

where $\tilde{\upsilon}_e = n + \upsilon_e$ and $\tilde{s}_e^2 = [(\mathbf{y}^* - \mathbf{W}\boldsymbol{\theta})^T(\mathbf{y}^* - \mathbf{W}\boldsymbol{\theta}) + \upsilon_e s_e^2]/\tilde{\upsilon}_e$. For the ordinal probit model, σ_e^2 is fixed at 1 (see Eq. A2).

The fully conditional posterior distribution of γ_k is given by:

$$\gamma_k|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{-k}, \sigma_e^2, \mathbf{y}^* \sim B(1, \tilde{p}_k), \tag{A11}$$

where $\boldsymbol{\gamma}_{-k}$ is a vector obtained by removing element k from the vector $\boldsymbol{\gamma}$,

$$\tilde{p}_k = \begin{cases} a_k/(a_k+b_k) & \text{if } g_k < Q_{\max} \\ 0 & \text{if } g_k \geq Q_{\max} \end{cases},$$

$$a_k = \frac{\lambda^{g_k+1}}{c(g_k+1)!} \times \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{y}^* - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^*)^T(\mathbf{y}^* - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^*) - \lambda\right\}, \text{ and}$$

$$b_k = \frac{\lambda^{g_k+1}}{cg_k!} \times \exp\left\{-\frac{1}{2\sigma_e^2}(\mathbf{y}^* - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^{**})^T(\mathbf{y}^* - \mathbf{Q}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\eta}_k^{**}) - \lambda\right\}$$

where

$$g_k = \sum_{i=1}^{K} \gamma_i - \gamma_k \quad \text{and}$$

$$c = \sum_{i=1}^{Q_{\max}} \frac{\lambda^i \exp(-\lambda)}{i!}.$$

The vector $\boldsymbol{\eta}_k^*$ is the column vector of $\boldsymbol{\eta}$ with the entries corresponding to marker k replaced by $\boldsymbol{\beta}_k$. Similarly, $\boldsymbol{\eta}_k^{**}$ is obtained from $\boldsymbol{\eta}$ with the entries corresponding to marker k replaced by the null vector ($\mathbf{0}$).

In the analyses, we set hyperparameters for the prior distributions as $\upsilon_\beta = 2, s_\beta^2 = 1, \upsilon_e = -2, s_e^2 = 0, \lambda = 1$, and $Q_{\max} = 15$. In this hyperparameter setting, the prior of σ_e^2 becomes a flat prior (i.e., an improper uniform distribution). Thus, the parameters $\boldsymbol{\alpha}$, σ_e^2 and $\boldsymbol{\kappa}$ had improper distributions in this study. When improper priors are assigned, the posterior distributions may not always be proper (Hobert and Casella 1996). One way to avoid an improper posterior distribution caused by an improper prior distribution is to specify upper and/or lower limits of parameters. In this study, however, we did not specify such limits for the parameters $\boldsymbol{\alpha}, \sigma_e^2$ and $\boldsymbol{\kappa}$, because the improper priors of these parameters seemed to work well in the simulation studies.

Appendix B: 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 of y_i^* and κ_m in the model in Eq. 2, however, we used the multivariate Hastings-within-Gibbs algorithm proposed by Cowles (1996), since the latter algorithm substantially improves the convergence of the MCMC estimations (Cowles 1996). Setting the initial values of the parameters as $\sigma_e^2 = 1$, $\sigma_{\beta_k}^2 = 1$, $\alpha_j = 0$, $\beta_{kl} = 0$, and $\gamma_k = 0$, the MCMC sampling proceeds as follows:

1. Update $\mathbf{W}$ and $\boldsymbol{\Sigma}$ with Eqs. A4 and A5, respectively.
2. Sample $\boldsymbol{\kappa}$ from the full conditional posterior distribution described in Eq. A3. (This step is only necessary for ordinal data.)
3. Sample $\mathbf{y}^*$. The full conditional posterior distribution of $\mathbf{y}^*$ is described in Eq. A1 for censored data and Eq. A2 for ordinal data.
4. Update $\mathbf{C}$ and $\mathbf{r}$ with Eqs. A6 and A7, respectively.
5. Sample $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (i.e., $\boldsymbol{\theta}$) from the full conditional posterior distribution described in Eq. A8.
6. Sample $\sigma_{\beta_k}^2$ for each k from the full conditional posterior distribution described in Eq. A9.
7. Sample σ_e^2 from the full conditional posterior distribution described in Eq. A10. (This step is not necessary for ordinal data.)
8. Sample $\boldsymbol{\gamma}$ from the full conditional posterior distribution described in Eq. A11.

The above process was repeated many times (see "Data analysis procedure" in "Materials and methods") to obtain the MCMC samples.

References

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679

Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. Mol Breed 19:341–356

Cowles MK (1996) Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. Stat Comput 6:101–111

Cui KH, Peng SB, Xing YZ, Xu CG, Yu SB, Zhang Q (2002) Molecular dissection of seedling-vigor and associated physiological traits in rice. Theor Appl Genet 105:745–753

Donnelly PJ, Tavaré S (1995) Goalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421

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

Godsill SJ (2001) On the relationship between Markov chain Monte Carlo methods for model uncertainty. J Comput Graph Stat 10:230–248

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732

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

Hawks JG (1983) The diversity of crop plants. Harvard University Press, Cambridge

He P, Li SG, Qian Q, Ma YQ, Li JZ, Wang WM, Chen Y, Zhu LH (1999) Genetic analysis of rice grain quality. Theor Appl Genet 98:502–508

Hobert JP, Casella G (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. J Am Stat Assoc 91:1461–1473

Huang N, Parco A, Mew T, Magpantay G, McCouh 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

Huang H, Eversley CD, Threadgill DW, Zou F (2007) Bayesian multiple quantitative trait loci mapping for complex traits using markers of the entire genome. Genetics 176:2529–2540

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Iwata H, Uga Y, Yoshioka Y, Ebana K, Hayashi T (2007) Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. Theor Appl Genet 114:1437–1449

Jannink JL (2007) Identifying quantitative trait locus by genetic background interaction in association studies. Genetics 176:553–561

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K et al (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. Science 301:376–379

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

Kingman JFC (1982) The coalescent. Stochastic Process Appl 13:235–248

Kojima S, Takahashi Y, Kobayashi Y, Monna L, Sasaki T, Araki T, Yano M (2002) *Hd3a*, a rice ortholog of the *Arabidopsis FT* gene, promotes transition to flowering downstream of *Hd1* under short-day conditions. Plant Cell Physiol 43:1096–1105

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

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. Nature Genet 8:365–372

Lanceras JC, Huang HL, Naivikul O, Vanavichit A, Ruanjaichon V, Tragoonrung S (2000) Mapping of genes for cooking and eating qualities in Thai Jasmine rice (KDML105). DNA Res 7:93–101

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2049

Laval G, Excoffier L (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20:2485–2487

Li ZK, Yu SB, Lafitte HR, Huang N, Courtois B, Hittalmani S, Vijayakumar CH, Liu GF, Wang GC, Shashidhar HE, Zhuang JY, Zheng KL, Singh VP, Sidhu JS, Srivantaneeyakul S, Khush GS (2003) QTL × environment interactions in rice. I. Heading date and plant height. Theor Appl Genet 108:141–153

Li J, Xiao J, Grandillo S, Jiang L, Wan Y, Qiyun D, 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. Theor Appl Genet 47:697–704

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175:879–889

Mei HW, Luo LJ, Ying CS, Wang YP, Yu XQ, Guo LB, Paterson AH, Li ZK (2003) Gene actions of QTLs affecting several agronomic traits resolved in a recombinant inbred rice population and two testcross populations. Theor Appl Genet 107:89–101

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Nat Acad Sci USA 70:3321–3323

Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (2007) Association mapping in plants. Springer, New York

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

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

Sillanpää MJ, Bhattacharjee M (2006) Association mapping of complex trait loci with context-dependent effects and unknown context variable. Genetics 174:1597–1611

Sillanpää MJ, Hoti F (2007) Mapping quantitative trait loci from a single-tail sample of the phenotype distribution including survival data. Genetics 177:2361–2377

Sillanpää MJ, Kilpikari R, Ripatti S, Onkamo P, Uimari P (2001) Bayesian association mapping for quantitative traits in a mixture of two populations. Genet Epidemiol 21(Suppl 1):S692–S699

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

Sorensen DA, Gianola D, Korsgaard I (1998) Bayesian mixed-effect model analysis of censored normal distribution with animal breeding applications. Acta Agric Scand 48:222–229

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) *Dwarf8* polymorphisms associate with variation with flowering time. Nature 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

Weber A, Clark RM, Vaughn L, Sanchez-Gonzalez JD, Yu JM, Yandell BS, Bradbury P, Doebley J (2007) Major regulatory genes in maize contribute to standing variation in teosinte (*Zea mays* ssp *parviglumis*). Genetics 177:2349–2359

Yamamoto T, Lin H, Sasaki T, Yano M (2000) Identification of heading date quantitative trait locus Hd6 and characterization of its epistatic interaction with Hd2 in rice using advanced backcross progeny. Genetics 154:885–891

Yamanaka S, Nakamura I, Watanabe KN, Sato YI (2004) Identification of SNPs in the *waxy* gene among glutinous rice cultivars and their evolutionary significance during the domestication process in rice. Theor Appl Genet 108:1200–1204

Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, Fuse T, Baba T, Yamamoto K, Umehara Y, Nagamura Y, Sasaki T (2000) Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering time gene *CONSTANS*. Plant Cell 12:2473–2484

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

Yi N, Xu S (2000) Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155:1391–1403

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

Yi N, Xu S, George V, Allison DB (2004) Mapping multiple quantitative trait loci for complex ordinal traits. Behav Genet 34:3–15

Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005) Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170:1333–1344

Yi N, Banerjee S, Pomp D, Yandell BS (2007) Bayesian mapping of genome-wide interacting QTL for ordinal traits. Genetics 176:1855–1864

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotech 17:155–160

Yu J, Pressoir G, Briggs W, Vroh Bi I, 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. Nature Genet 38:203–208

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An *Arabidopsis* example of association mapping in structured samples. PLoS Genet 3:e4

Zweig MH, Campbell G (1993) Receiver-operating characteristic (ROC) plots—a fundamental evaluation tool in clinical medicine. Clin Chem 39:561–577