Applying gradient tree boosting to QTL mapping with Shapley additive explanations

Tomohiro Ishibashi, Akio Onogi*

Department of Life Sciences, Faculty of Agriculture, Ryukoku University, Otsu, Shiga, 520-2194, Japan

Corresponding author:

Akio Onogi

Postal address: 1-5, Yokotani, Oe-cho, Seta, Otsu, Shiga, 520-2194, Japan

Phone: +81-(0)77-599-5719

E-mail: onogiakio@gmail.com

ORCID: 0000-0003-1707-9539

Running head

Applying SHAP-assisted XGBoost to QTL mapping

Keywords

Epistasis, gene interaction, variable selection, association mapping, SHAP

Abstract

Motivation: Mapping quantitative trait loci (QTLs) is essential in plant and animal genetics, but identifying epistasis remains challenging. Machine learning methods, such as XGBoost with Shapley additive explanations (SHAP), can enhance QTL mapping by capturing complex relationships and evaluating feature importance.

Result: SHAP-assisted XGBoost (SHAP-XGB) was compared to composite interval mapping (CIM), multiple interval mapping (MIM), inclusive CIM (ICIM), and BayesC using simulations and rice heading date data. SHAP-XGB performed comparably for main QTL effects and surpassed other methods in detecting epistatic interactions. SHAP's ability to assess local importance improved the interpretation of marker interactions in plant and animal genetics.

Availability and Implementation: R scripts for SHAP-XGB are available at https://github.com/Onogi/SHAP-XGB.

Contact: onogiakio@gmail.com

Supplementary information: Supplementary information, figures, and tables are available at the journal's website.

Introduction

Detecting quantitative trait loci (QTLs) underpinning variations in quantitative traits is one of the major goals in analyzing quantitative traits of plants and animals. Detecting QTLs can lead to useful applications such as efficient selection based on QTL genotypes, genomic engineering to modify phenotypes, and advancing quantitative genetics theories by deciphering the genetic architecture of quantitative traits (Falconer and Mackay 1996). Interval mapping (IM) is a typical method to detect QTLs in inbred line crosses (Lander and Botstein 1989). IM tests associations of each marker and marker interval with the phenotypes independently. In practice, composite IM (CIM), an extension of IM, is the most popular method (Jansen and Stam 1994; Zeng 1994). CIM considers markers linked with major QTLs as cofactors to enhance detection power and precision when multiple QTLs exist.

Interactions between QTLs, known as epistasis, have also been of interest in quantitative genetics because genes (proteins) function in corporations. Detecting epistatic QTLs will broaden the applications of the QTLs described above and will be helpful to depict gene networks underlying complex quantitative traits. A successful example of QTL and epistatic QTL mapping in plant breeding is the rice heading date. To date, a number of QTL/genes and their networks under short- and long-day conditions have been elucidated (Hori et al. 2016; Wei et al. 2020). Methods for detecting epistatic QTLs include multiple interval mapping (MIM, Kao et al. 1999; Laurie et al. 2014), inclusive CIM (ICIM, Li et al. 2008), and Bayesian regression where whole genome-wide markers are used as explanatory variables simultaneously (Zhang and Xu 2005; Li and Sillanpää 2012). A method to detect interactions between QTLs and genetic background was also proposed (Jannink and Jansen 2001). Detection of QTL/gene interactions has been of interest in genome-wide association studies (GWAS), although detection is more difficult than in inbred line crosses because of the generally low allele frequencies. These methods include hierarchical Bayesian modeling for high-order gene interactions (Bayesian high-order interaction toolkit (BHIT) (Wang et al. 2015) and a marginal epistasis test (MAPIT) (Crawford et al. 2017).

Recent machine learning models, such as gradient boosting and deep neural networks, have shown high predictive abilities in various applications. Because these models are inherently capable of capturing complex interactions among features, they are expected to be useful for QTL mapping, particularly for complex QTL interactions, using line crosses. However, unlike linear regressions, the contributions of each feature and feature interaction are not explicitly outputted. Thus, such complex models should be used with efficient and reliable methods to extract the importance of features for selection. Several examples can be seen in GWAS. Johnsen et al. (2021) applied Shapley additive explanations (SHAP) to estimate the feature importance of extreme gradient boosting (XGBoost) (Chen and Guestrin 2016) in a GWAS using UK Biobank data. Cui et al. (2022) applied a Shapley-based interaction score (Sundararajan et al. 2020) to deep neural networks to identify gene-gene interactions in a GWAS. The Shapley value was originally proposed to evaluate players' contributions

to rewards in cooperative games (Shapley 1953). By considering players and rewards as features and prediction outcomes, respectively, the Shapley value can be used to evaluate the feature and feature combination importance of the prediction models. SHAP is based on the Shapley value and is characterized by the local importance of the feature; that is, the importance of the feature for each instance (i.e., sample) is estimated (Lundberg and Lee 2017). Because the summation of the SHAP values of the features for an instance is equal to the predicted values of the instance (a property called local accuracy), it is feasible to determine which features have a large impact on the prediction of the instance (Lundberg and Lee 2017; Lundberg et al. 2018). SHAP also has two additional properties–missingness and consistency–that are desirable as feature importance indices.

It may be possible to map QTLs and QTL interactions using SHAP or Shapley value-based scores and recently advanced machine learning more efficiently and effectively than conventional methods. However, no attempts have been reported for mapping populations, and comparisons with conventional methods such as CIM, MIM, ICIM, and Bayesian regressions have not been conducted. In this study, we evaluated the potential of XGBoost for QTL and epistatic QTL mapping using SHAP in inbred line crosses. XGBoost is an algorithm for gradient boosting, which is a class of ensemble learning (Chen and Guestrin 2016). XGBoost was selected because it is one of the most prevalent gradient boosting algorithms. We characterized this SHAP-assisted XGBoost (SHAP-XGB) by applying the method to simulated and real biparental populations and comparing it with conventional methods. The results illustrated the strength of SHAP-XGB in detecting and interpreting epistatic QTLs, as expected, and suggested the possibility that SHAP-XGB complements conventional methods.

Materials and Methods

Shapley values

In cooperative game theory, the Shapley value measures how much each player contributes to the game's reward. Assessing the contributions of individual players independently is insufficient in cooperative games because there are usually interactions between players. The core concept of

Shapley's value is the 'marginal contribution'. Marginal contribution represents the extent to which a player's participation in a cooperative game increases the game's rewards. This varies according to the combination of players and order of participation. Therefore, the Shapley value of a player is defined by calculating the marginal contribution of a player to all possible player combinations and participation orders and taking their average. In other words, the Shapley value is the average marginal contribution of the player. The Shapley value for player *i* is defined as follows:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f(S \cup \{i\}) - f(S)] \quad (1),$$

where *N* is a set of players, *S* is a subset of players, *M* is the number of players, and *f* is the function that outputs rewards for the given players.

SHAP

Players of Shapley values are replaced with features (marker genotypes in the case of QTL mapping), and rewards are replaced with the model output (fitted or predicted phenotypes in the case of QTL mapping) in SHAP. The model function to output the rewards given to players takes the form of the expected values, that is, $f(S)$ in Eq. (1) is expressed as $E[f(x)|x_S]$ in SHAP, where x and x_S indicate all features and the features in subset *S*, respectively. Thus, the SHAP value of feature *i* of an instance (i.e., genotype/line in the case of QTL mapping) is

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} \left[E\left[f(x)|x_{S \cup \{i\}}\right] - E[f(x)|x_S]\right]. \quad (2)$$

Here, XGBoost was used as the model f.

SHAP is an additive feature-contribution method with three properties: local accuracy, missingness, and consistency. Local accuracy indicates that the summation of the feature contributions defined by Eq. (2) and is equal to the output of the model. Missingness means that importance is not assigned to a missing feature. Consistency implies that if the model is changed such that a feature has a greater impact on the model, the contribution assigned to that feature does not decrease.

The interaction index was used to assess the impact of one feature of a model when combined

with another. The Shapley interaction index between features i and j is calculated as (Fujimoto et al. 2006),

$$\varphi_{ij} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(M-|S|-2)!}{2(M-1)!} \delta_{ij}(S), \quad (3)$$

where $\delta_{ij}(S) = f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S)$. The SHAP interaction value was calculated by replacing the function $f(S)$ with expectations, as shown in Eq. (2). The SHAP interaction value is split equally between each feature combination, that is, $\varphi_{ij} = \varphi_{ji}$. Thus, total interaction effects are calculated as $\varphi_{ij} + \varphi_{ij} = 2\varphi_{ij}$. The SHAP value ($\varphi_i$) of feature i for an instance is decomposed to the main effect of the feature (φ_{ii}) and the interaction effects between the feature and the others as

$$\varphi_i = \varphi_{ii} + \sum_{j \neq i} \varphi_{ij}.$$

That is, the main effect of feature i is obtained by subtracting the SHAP interaction value between feature i and j ($j \neq i$) from the SHAP value of feature i.

Lundberg et al. (2018) proposed an efficient algorithm for calculating SHAP values from tree-based machine learning models (referred to as Tree SHAP). The algorithm exploits the differences in outputs between models with and without a particular feature using a tree structure to evaluate $E[f(x)|x_S]$. The algorithm was implemented in R package xgboost (ver. 1.7.5.1).

Algorithm of SHAP-XGB

The algorithm of SHAP-XGB is described below.

Step 1: Create a bootstrap dataset.

Step 2: Fit an XGBoost model to the dataset, with early stopping determined from the remaining samples. The xgboost R package was used in this study.

Step 3: Calculate the SHAP values (φ_i) from the fitted samples.

Step 4: Decompose the SHAP values into the SHAP interaction values (φ_{ij}) and the remaining values (φ_{ii}).

Step 5: Repeat steps 1–4 100 times and calculate the average of φ_i, φ_{ij}, and φ_{ii}. Note that these values are measures of local importance (i.e., for each line).

Step 6: Average the absolute values of the values obtained in Step 5 across lines to obtain global importance. The averages of the absolute values of φ_i or φ_{ii} are referred to as the global SHAP scores and the averages of the absolute values of $2\varphi_{ij}$ are referred to as the global SHAP interaction scores.

Four notifications were added to these algorithms. First, although the absolute SHAP values are averaged across lines to obtain the global importance in Step 6, the consistency property still holds for the global importance scores (Lundberg et al. 2018). Second, bootstrapping was conducted to separate the data into training and validation sets, which were used for early stopping to prevent overfitting. Therefore, other resampling methods could be applied. The number of bootstrap replicates (100) was determined arbitrarily. Thus, fewer numbers—even one—may be acceptable, albeit with a loss of reliability of the SHAP values. Third, default values were used for the hyperparameters of xgboost except for the "early_stopping_rounds" argument, which was set to three. This argument specifies the number of rounds in which the boosting iterations stop if the prediction performance does not improve. These settings resulted in a moderate fit of the models to data. We compared settings that lead to underfitting and overfitting and found that the effects were negligible (see the Supplementary Information). Lastly, there would be two choices for detecting the main effects of QTLs: using SHAP values that contain interaction scores (i.e., φ_i) or using SHAP values from which interaction scores are removed (i.e., φ_{ii}). In this study, the former score is referred to as "compound" scores, and the latter is referred to as the "pure" scores for detecting QTL main effects. These two scores were compared in the Supplementary Information. Although slight differences in performance were observed, we used the compound effects for mapping the main effects for the reasons discussed in the Discussion section.

Comparing methods

The SHAP-XGB method was compared with CIM, MIM, ICIM, and BayesC (Habier et al. 2011) using simulated and real data. Note that BayesC and SHAP-XGB used only marker genotypes as covariates and did not use the genotype probabilities of the loci between markers. For the CIM, MIM, and ICIM, the intervals were set to 1 cM.

Composite interval mapping (CIM)

CIM was performed using the cim function of the R package qtl (ver. 1.60, Broman et al. 2003). The expectation-maximization algorithm was used. The number of covariates was three, and the window size was 10 cM. CIM was used to detect the main QTL effects. We assessed the influence of the number of covariates and window size on the simulated and real data and found that the influence was negligible (data not shown).

Multiple interval mapping (MIM)

MIM theory was first proposed by Kao et al. (1999). MIM assumes that the genotypic value is the sum of multiple QTLs and QTL interaction effects. QTL genotypes are probabilistically represented using neighboring marker genotypes and recombination frequencies; thus, the MIM model is a finite mixture model. The QTLs and QTL interactions included in the model were searched in a stepwise manner. MIM was conducted using the QTL cartographer (ver. 2.5_011, Wang et al. 2012). We followed the algorithm proposed by Laurie et al. (2014). The algorithm consisted of multiple steps, including a forward search of the main and interaction effects based on Bayesian information criteria, optimization of QTL positions, and testing of QTL effects with a certain significance level (5%).

Inclusive composite interval mapping (ICIM)

Li et al. (2007) and Li et al. (2008) proposed the ICIM theory. ICIM consists of two steps: in the first step, markers and marker interactions significantly associated with phenotypes are selected with stepwise regression and one- or two-dimensional interval mapping is conducted in the second step by adjusting the phenotypes with the selected markers that are not related to the target intervals. The ICIM was performed using IciMapping (ver. 4.2.53, Meng et al. 2015). The *p*-values of the markers for entering models in stepwise regression (parameter PIN) were set to 0.001 for the main effects (option

203 ICIM-ADD) and 0.0001 for mapping interaction effects (option ICIM-EPI). We attempted a higher
204 PIN value for ICIM-EPI (0.01) in the simulated data; however, the influence was negligible (data not
205 shown).

206 **BayesC**

207 We used BayesC as a representative Bayesian regression method because it circumvents a drawback
208 relevant to degrees of freedom in parameter estimation (Habier et al. 2011). BayesC analysis was
209 performed using the vigor function of the R package VIGoR (ver. 1.1.0, Onogi and Arakawa 2022).
210 This package uses fast variational Bayesian algorithms to estimate the posterior distributions of
211 Bayesian alphabets. The model included all markers and marker interactions (multiplications of marker
212 genotypes) and estimated their coefficients simultaneously. The prior distribution of a coefficient was
213 normal if the indicator of the marker (or marker combination) was 1 and 0 if the indicator was 0. The
214 prior distributions of indicators were Bernoulli distributions with hyperparameters of 0.05 for main
215 effects and 0.001 for interaction effects. Different prior normal distributions were assumed for main
216 and interaction effects. Bootstrapping was applied to make the inferences about the indicators more
217 robust. The model was fitted to bootstrap samples, and the posterior means of the indicators were
218 averaged across the bootstrap samples. We used 1000 bootstrap samples in this study.

219

220 **Simulations**

221 The usefulness of SHAP-XGB was first investigated using simulations that mimicked biparental F_2
222 and RIL populations. The conditions of the simulations were as follows: The number of chromosomes
223 was 8, the length of each chromosome was 120 cM, the number of markers (SNPs) per chromosome
224 was 20, and the number of QTLs was five. The QTLs were assigned so as to be one QTL per
225 chromosome, and chromosomes harboring the QTLs were randomly selected. Among the 10
226 combinations of QTLs, half were assumed to have interaction effects, and the combinations with
227 interaction effects were randomly selected. The main and interaction effects of the QTLs were
228 generated using the standard normal distribution. The broad-sense heritability was set at 0.8. Although

229 only additive effects are discussed here, the investigation of dominance effects is provided in the
230 Supplementary Information. The results for dominance effects were generally consistent with those
231 for additive effects.

232 We simulated four scenarios using these conditions: (1) F_2 population with $N = 200$ where N is the
233 number of genotypes (lines). (2) RIL population with $N = 200$. (3) F_2 population with $N = 400$. (4) RIL
234 population with $N = 400$. Interaction effects were simulated by multiplying the corresponding QTL
235 genotypes. A summary of the simulation scenarios and realized proportions of variances due to the
236 additive and interaction effects are presented in Table 1. The QTLs and marker genotypes were
237 generated using the R package qtl (ver. 1.60). QTL effects and phenotypes were generated using house-
238 developed R scripts. The number of replications per scenario was 10, which was determined by
239 considering the manual operations for MIM and ICIM.

240

241 **Table 1** Simulation scenarios[1].

Scenario	Pop.	Size	Proportion of variance[2]	
			V_A	V_I
1	F_2	200	0.476 ± 0.104	0.328 ± 0.112
2	RILs	200	0.408 ± 0.175	0.385 ± 0.176
3	F_2	400	0.630 ± 0.091	0.175 ± 0.087
4	RILs	400	0.383 ± 0.218	0.423 ± 0.216

242 [1]All populations were assumed to be derived from biparental crosses.
243 [2]V_A: additive variance, V_I: epistasis variance.
244

The accuracy of QTL detection was evaluated using receiver operating characteristic (ROC) curves and the area under the curve (AUC). When drawing ROC curves, markers located within 20 cM of the QTLs were regarded as positive markers. For interaction effects, combinations of markers located within 20 cM of QTLs relevant to the interactions were regarded as positive combinations. ROC curves and AUC were calculated using the R package pROC ver. 1.18.0 (Robin et al. 2011). Differences in the AUC values between the methods were tested using the Wilcoxon signed-rank test (the significance level was 5% with Bonferroni correction). The global SHAP and global SHAP interaction scores were regarded as QTL signals for SHAP-XGB. The LOD scores were regarded as the signals for CIM and ICIM, respectively, and the bootstrap means of the marker indicators were regarded as the signals for BayesC. Because MIM implemented in QTL cartographers returns a list of detected QTLs and QTL interactions, ROC curves could not be drawn, and consequently, AUC could not be calculated (note that ROC curves require scores of all variables). Instead, the true and false positive rates were calculated.

Real data

The real data were obtained from Onogi et al. (2016). Briefly, days to heading of rice were evaluated for 174 Koshihikari/Kasalath//Koshihikari backcross inbred lines (BILs) and their parental lines. The experiments were conducted in six experimental fields in Japan and Vietnam over three years, resulting in nine environments. The nine environments are summarized in Table 2, and details were provided in Table S1 in Onogi et al. (2016). The lines were genotyped for 162 biallelic markers.

Dependency plots

SHAP interaction values were calculated for each instance (line) because SHAP was originally proposed as a local importance measure. Thus, the interactions between QTLs (markers) can be visualized by plotting the SHAP interaction values of the marker combinations of interest for all lines. These plots are referred to as dependency plots. Dependency plots were constructed using the R

271 package SHAPforxgBoost (ver. 0.1.1) with some modifications.

272

273 **Table 2** Nine environments in the rice heading date data[1].

Environment name	Location	Year	Transplanting date
Tsukuba2007	Tsukuba (N36°01'/E140°06')	2007	8 Jun.
Fukuoka2008	Fukuoka (N33°37' /E130°28')	2008	24 Jun.
HaNoi2008	HaNoi (N21°01' /E105°80')	2008	3 Jul.
Ishigaki2008	Ishigaki (N24°23' /E124°11')	2008	17 Jul.
Ishikawa2008	Ishikawa (N36°30' /E136°36')	2008	16 May
ThaiNguyn2008	ThaiNguyn (N21°35' /E105°49')	2008	15 Jul.
Tsukuba2008E	Tsukuba (N36°01' /E140°06')	2008 (Early sowing)	25 Apr.
Tsukuba2008L	Tsukuba (N36°01' /E140°06')	2008 (late sowing)	9 Jul.
Tsukuba2009	Tsukuba (N36°01' /E140°06')	2009	21 May

274 [1]Details of these environments are provided in Table S1 of Onogi et al. (2016).

275

276 Results

277 In comparative analyses mapping main QTL effects, SHAP-XGB generally showed higher or
278 comparable AUC values than CIM, ICIM, and BayesC (Figure 1). When compared with MIM, SHAP-
279 XGB showed true positive rates comparable to MIM at the false positive rates that MIM showed
280 (Figure 1). In comparative analyses mapping QTL interaction effects, SHAP-XGB showed
281 significantly higher AUC values than ICIM and BayesC (Figure 2). When compared with MIM,
282 SHAP-XGB showed higher true positive rates than MIM under false positive rates comparable to those
283 of MIM (Figure 2), suggesting that SHAP-XGB can detect more QTL interactions than MIM under
284 comparable false positive rates. Interactions between markers (QTLs) can be interpreted using
285 dependency plots. The SHAP interaction values (local importance metrics) for a replication of Scenario
286 2 are shown in Figure 3. The effects of two QTLs, *QTL2* and *QTL3* tagged with *D3M19* and *D4M11*

markers, respectively, were inverted depending on the genotype of *QTL4* tagged with *D6M1*. These figures illustrate that dependency plots using local importance scores can facilitate the interpretation of interactions.

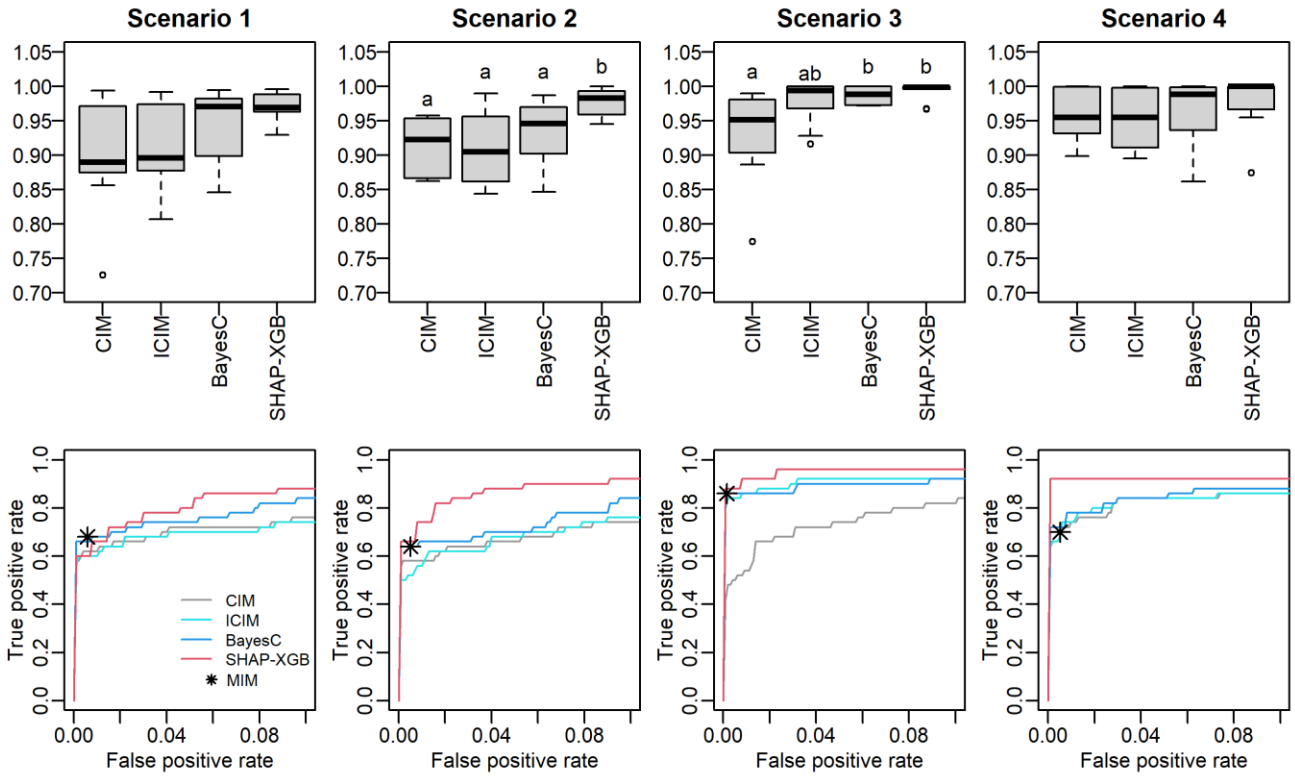

Figure 1 Area under the curve (AUC) values and receiver operation characteristics (ROC) curves in mapping main QTL effects. The top and bottom panels are the AUC values and ROC curves, respectively. The ROC curves are the mean curves of 10 replications. Scenarios 1 and 3 simulated F_2 populations, and Scenarios 2 and 4 simulated recombinant inbred line populations. The population sizes were 200 for Scenarios 1 and 2, and 400 for Scenarios 3 and 4. Different letters in the box plots denote significant differences (P <0.05). Because the results of MIM are the lists of detected QTLs, true and false positives are shown with the asterisk in the bottom plots. CIM: composite interval mapping, ICIM: inclusive CIM, SHAP-XGB: Shapley additive explanations-assisted XGBoost, MIM: multiple interval mapping.

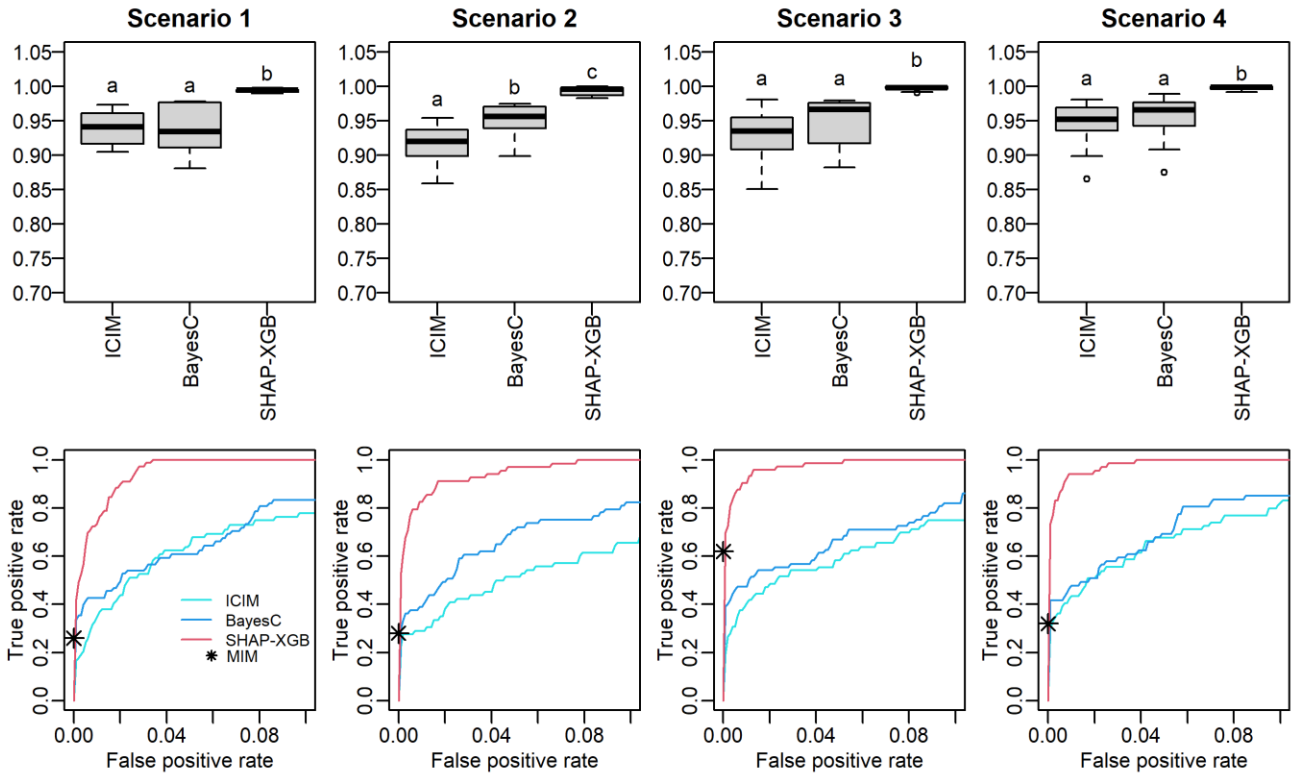

Figure 2 Area under the curve (AUC) values and receiver operation characteristics (ROC) curves in mapping QTL interaction effects. The top and bottom panels are the AUC values and ROC curves, respectively. The ROC curves are the mean curves of 10 replications. Scenarios 1 and 3 simulated F_2 populations and Scenarios 2 and 4 simulated recombinant inbred line populations. The population sizes were 200 for Scenarios 1 and 2, and 400 for Scenarios 3 and 4. Different letters in the box plots denote significant differences (P <0.05). Because the results of MIM are the lists of detected interactions, true and false positives are shown with the asterisk in the bottom plots. ICIM: inclusive composite interval mapping SHAP-XGB: Shapley additive explanations-assisted XGBoost, MIM: multiple interval mapping.

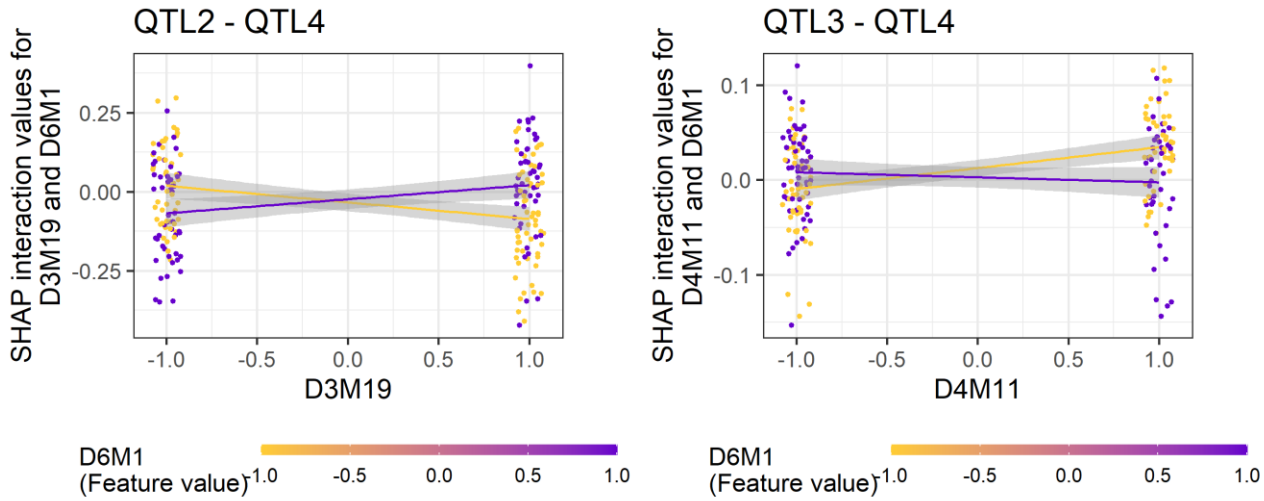

Figure 3 Dependencies between QTLs in a simulated dataset. The Shapley additive explanations (SHAP) interaction values of each line are shown as dots. The left panel shows the interaction between the 2nd and 4th QTLs, and the right panel shows that between the 3rd and 4th QTLs. *D3M19*, *D4M11*, and *D6M1* are the markers linked to the 2nd, 3rd, and 4th QTLs, respectively. Genotypes of *D6M1* are shown using different colors. Dots are scattered towards the x-axis. The regression lines were added for each genotype of *D6M1*. The shades denote the standard errors. The left panel shows that, when the genotype of *D6M1* is −1, SHAP interaction values (i.e., phenotypes) tend to decrease as the genotype of *D3M19* changes from −1 to 1, and when the genotype of *D6M1* is 1, SHAP interaction values tend to increase. The right panel shows inverted tendencies between *D4M11* and *D6M1*. The gray areas indicate the 95% confidence intervals of regression lines.

The XGBoost models explained 94.3% (± 1.9%) of the phenotypic variance with the setting of moderate fitting in the real rice data analyses. SHAP-XGB and the compared methods showed similar results in mapping the main QTL effects (Figure S1, Figure 4, and Table S1), and markers linked to *Hd6* (chr. 3, Takahashi et al. 2001), *Hd1* (chr. 6, Yano et al. 1997), and *DTH8* (chr. 8, Wei et al. 2010) showed higher signals for each method. The ICIM results were generally lenient when mapping epistatic QTLs (Figure S2 and Figure 5), whereas the BayesC results were stringent (or conservative). SHAP-XGB showed intermediate results between ICIM and BayesC, and could pinpoint interacting combinations between known QTLs within reasonably narrow margins. The results of the MIM in the mapping interactions were relatively stringent; nine interactions were detected in nine environments

(Table S1). The global SHAP interaction scores in each environment showed interesting trends across latitudes (Figure 6). Interactions between *Hd1* and *Hd6*, *Hd1* and *DTH8*, and *Hd1* and *Hd2*/*OsPRR37* (Yano et al. 1997; Koo et al. 2013) showed higher scores at higher latitudes (environments Tsukuba2007, Tsukuba2008E, Tsukuba2009, Ishikawa2008, and Fukuoka2008), but lower scores at lower latitudes (environments Ishigaki2008, HaNoi2008, and ThaiNguyn2008). In contrast, the interaction between *DTH8* and *Hd2* showed larger scores at lower latitudes but lower scores at higher latitudes.

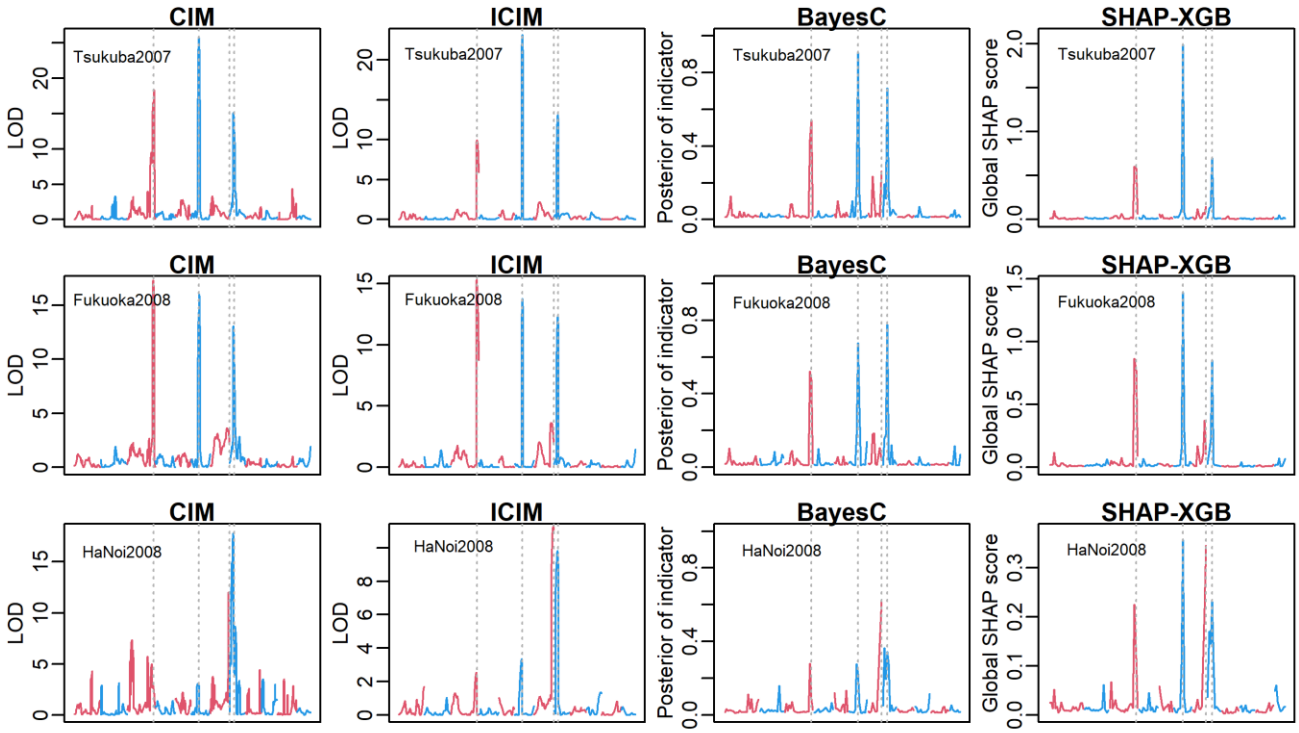

Figure 4 Comparison of QTL mapping methods to detect main effects in the rice heading date data. The y-axis denotes the QTL signals (CIM and ICIM, LOD score; BayesC, posterior indicator variable; SHAP-XGB, global SHAP score). The x-axis denotes the chromosome positions. The neighboring chromosomes are distinguished using different colors. The vertical bars indicate the four major QTLs (left to right, *Hd6*, *Hd1*, *Hd2*, and *DTH8*). Note that, the x axes differed among methods because CIM and ICIM conducted interval mapping whereas BayesC and SHAP-XGB used only marker genotypes as covariates. Three representative environments (Tsukuba2007, Fukuoka2008, and HaNoi2008) are presented. The full results are shown in Figure S1. CIM: composite interval mapping, ICIM: inclusive CIM, SHAP-XGB: Shapley additive explanations-assisted XGBoost.

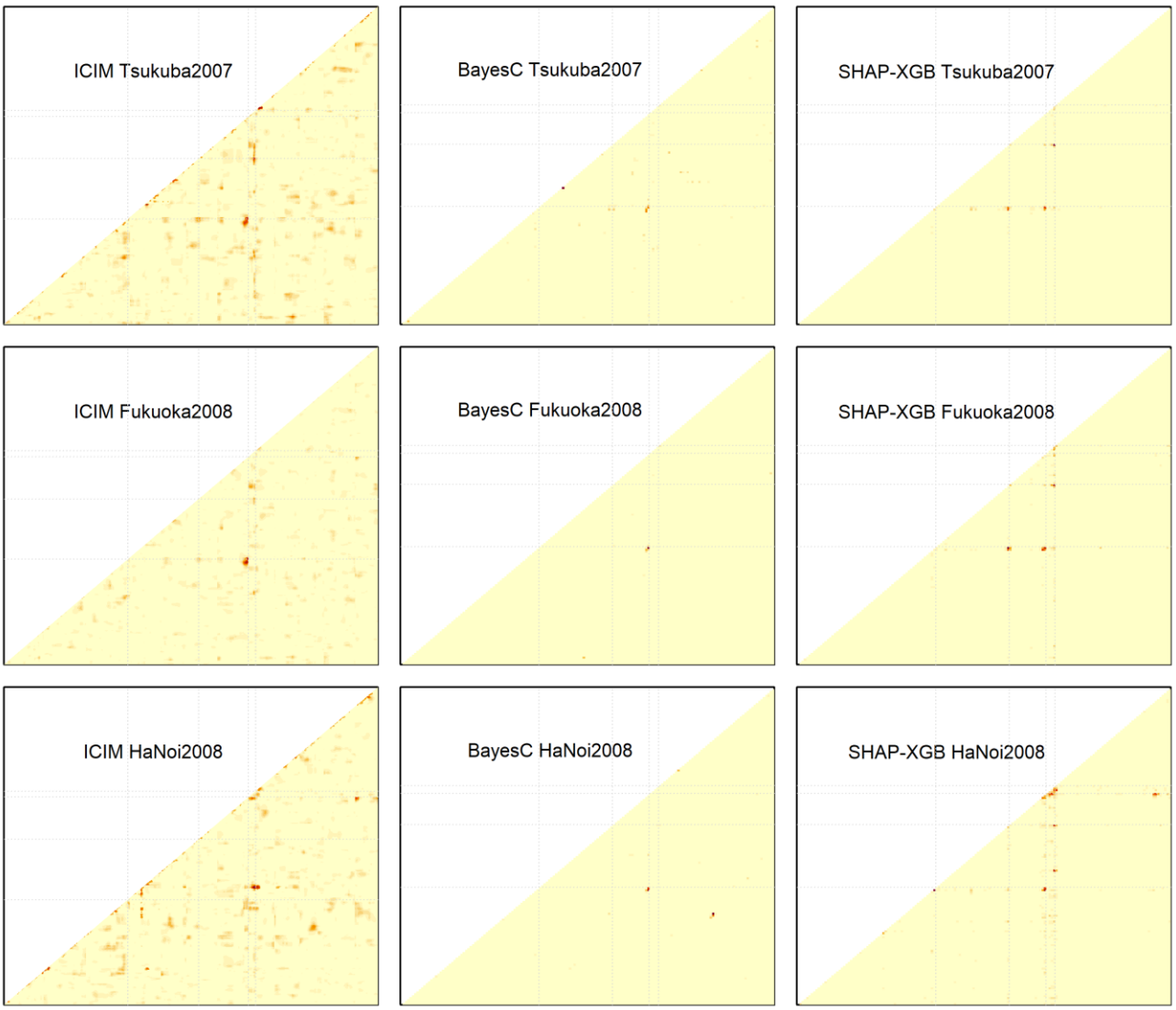

Figure 5 Comparison of QTL mapping methods to detect interaction effects in the rice heading date data. The y- and x-axes denote the relative positions of markers. The signal levels obtained with each QTL mapping method (ICIM, LOD score; BayesC, posterior indicator variable; SHAP-XGB, global SHAP interaction score) are shown using warm colors. The gray broken lines indicate the four major QTLs (left to right or bottom to top, *Hd6*, *Hd1*, *Hd2*, and *DTH8*). Note that positions of major QTLs in the x axes slightly differed among methods because CIM and ICIM used interval mapping whereas BayesC and SHAP-XGB used only marker genotypes as covariates. Three representative environments (Tsukuba2007, Fukuoka2008, and HaNoi2008) are presented. The full results are shown in Figure S2. ICIM: inclusive composite interval mapping, SHAP-XGB: Shapley additive explanations-assisted XGBoost.

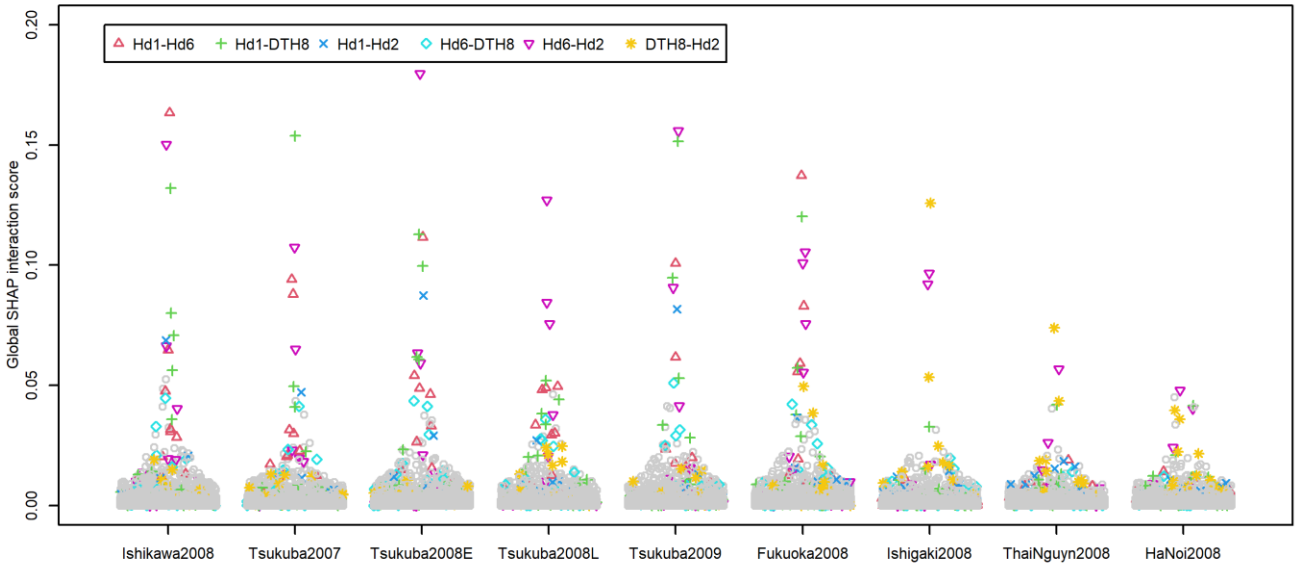

Figure 6 Summary of the global Shapley additive explanations (SHAP) interaction scores obtained in the rice heading date data. The y-axis denotes the global SHAP interaction scores, which are the averages of the absolute SHAP interaction values across all lines. The global SHAP interaction scores were horizontally jittered according to the densities of scores. The interaction values for the combinations of markers that were located within 20 cM from the known QTLs (*Hd6*, *Hd1*, *Hd2*, and *DTH8*) were regarded as the signals for the corresponding QTL interactions and highlighted using different colors and point shapes. The interactions between neighboring markers (< 20 cM apart) were removed for ease of visualization.

The interaction between *Hd1* and three QTLs (*DTH8*, *Hd6*, and *Hd2*) in Tsukuba2009 and the interaction between *DTH8* and *Hd2* in Ishigaki2008 were visualized using dependence plots (Figure 7). The Koshihikari allele (1) of *DTH8* and *Hd6* tended to shorten days to heading at Tsukuba2009 when the Koshihikari alleles were harbored at *Hd1*, whereas the Kasalath allele (−1) tended to prolong days to heading when the Kasalath alleles were harbored at *Hd1*. Contrasting tendencies were observed for the interactions between *Hd1* and *Hd2* in Tsukuba2009 and *Hd2* and *DTH8* in Ishigaki2008.

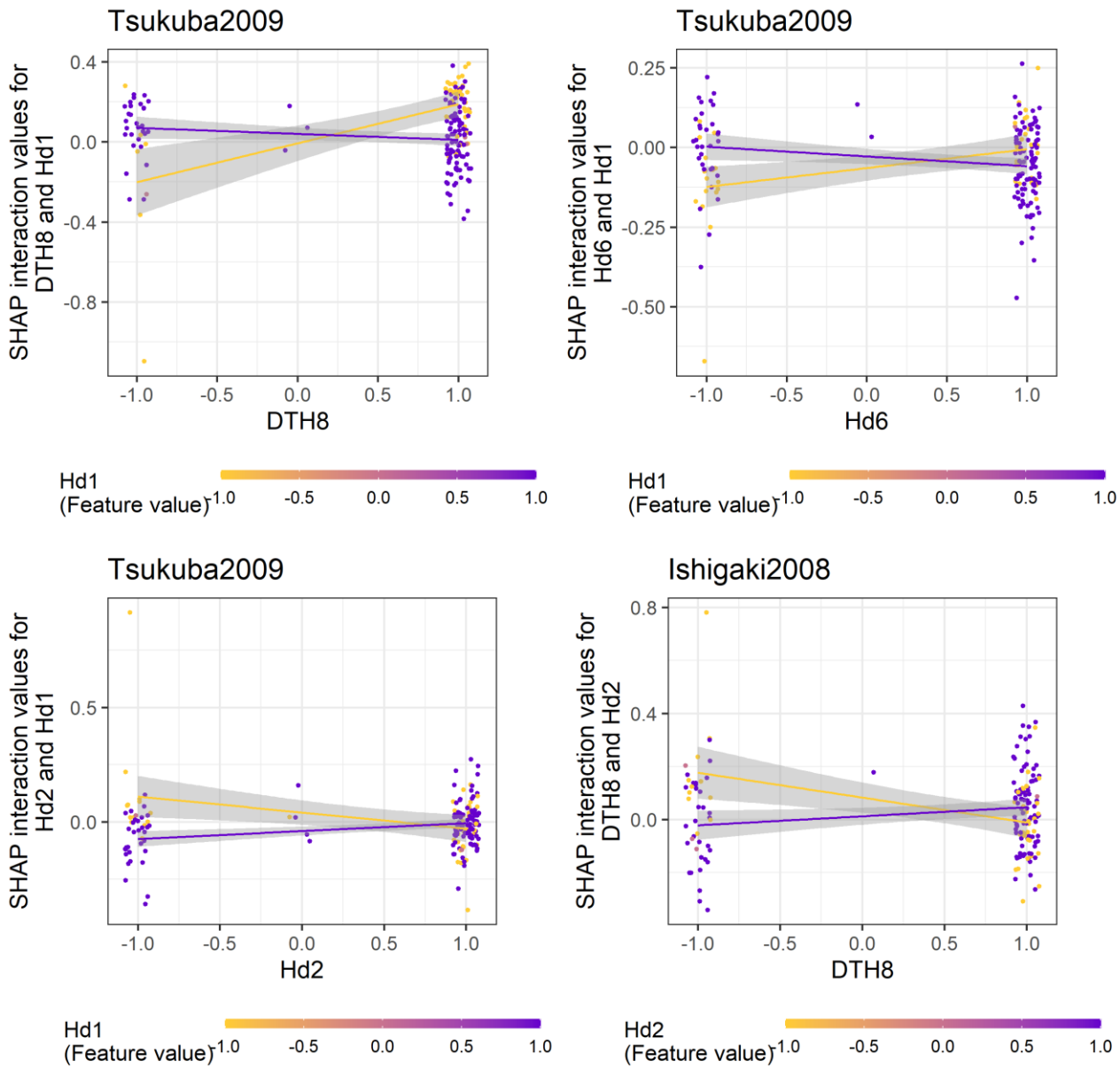

Figure 7 Dependencies between major QTLs shown with the Shapley additive explanations (SHAP) interaction values. The left top, right top, and left bottom panels show the dependency between *Hd1* and the three QTLs (*DTH8*, *Hd6*, and *Hd2*) at Tsukuba2009, respectively. The right bottom panel shows the dependency between *DTH8* and *Hd2* at Ishigaki2008. Each dot represents the SHAP interaction values for each line. Marker genotypes were regarded as the QTL genotypes. These markers were selected when the corresponding marker combinations showed the highest global SHAP interaction scores for the QTL interactions. For each panel and each QTL (marker), 1 and −1 indicate the Koshihikari and Kasalath alleles, respectively. The gray areas indicate the 95% confidence intervals of regression lines.

Discussion

We applied SHAP-assisted XGBoost (SHAP-XGB) to QTL mapping in biparental populations and compared it with conventional mapping methods in this study. SHAP-XGB generally showed comparable accuracy in finding QTL main effects with conventional methods and higher accuracy in finding epistatic QTL effects. The dependency plots enabled easy interpretation of the interactions. These results suggested that SHAP-XGB complements existing methods, particularly for identifying novel interactions between QTLs.

There are ways to measure feature importance for gradient boosting other than SHAP, including "gain" (the average increase of the objective function when the feature is added to trees) and "split count" (the number of counts that trees split using the feature). Although these indices are provided by the R package of XGBoost, we did not adopt them because they do not hold the consistency property (Lundberg et al. 2018) and the package did not support them as importance of feature interactions. Nevertheless, considering that there are various indices to measure feature importance, such as local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016) and the Shapley–Taylor interaction index (Sundararajan et al. 2020), it would still be worth investigating which indices are suitable for QTL mapping and whether the properties revealed here are applicable when SHAP is replaced with other indices. Such investigations are required to mitigate the negative effects of LD, as will be discussed later.

Although the simulated and real data analyses consistently suggest the usefulness of SHAP-XGB in QTL mapping, the marker genotypes have an undesirable property, i.e., LD between markers, for evaluating feature importance. Because XGBoost randomly samples markers to build a tree f, QTL effects are diluted with the markers linked to the QTLs, consequently decreasing the Shapley values of each linking marker. In fact, the global SHAP scores and SHAP interaction values in Figures 4 and 6 shrank to 0 (2.0 at most) considering that the units of the scores/values are the same as the phenotypic values (i.e., days). These decreased values were probably due to the correlated marker genotypes. The LD between the markers affected another aspect. In rice heading date analyses, interactions between

421 neighboring markers were occasionally assigned large SHAP interaction values, resulting in a
422 reduction in the main "pure" scores (these interactions were removed from Figure 6 for ease of
423 visualization). However, such interactions are caused by LD between neighboring markers and thus
424 should be interpreted as part of the main QTL effects. Although this decrease in pure main scores can
425 be circumvented using "compound" scores for mapping, caution is required when interpreting pure
426 scores. It has also been pointed out that approximations of Shapley values with SHAP become
427 inaccurate when features are correlated (Aas et al. 2021) because the evaluations of the expectation in
428 Eq. (2) assume independence between the features. Thus, taken together, LD would affect the detection
429 power of SHAP-XGB. One solution is to replace SHAP with other indices that consider the
430 correlations among features (Aas et al. 2021).

431 Interactions between the major heading date QTLs weakened at lower latitudes (Figure 6 and
432 Table 2). These results were reasonable because *Hd6*, *Hd2*, and *DTH8* are involved in heading date
433 variation under long-day conditions (Hori et al. 2016; Wei et al. 2020). Interactions between *Hd1* and
434 three QTLs (*Hd6*, *Hd2*, and *DTH8*) are plausible under natural long-day conditions. For example,
435 interactions between functional *Hd6* alleles (e.g., Kasalath allele) and *Hd1* alleles (e.g., Nipponbare
436 allele) were reported under long-day conditions (Ogiso et al. 2010). Interactions between *Hd1* and *Hd2*
437 under long-day conditions (Lin et al. 2000), and interactions among *Hd1*, *Hd2*, *DTH8*, and *Ghd7* under
438 natural long-day conditions were also reported (Zhang et al. 2019). The interactions between *Hd2* and
439 *DTH8*, which is prominent in lower latitudes, are interesting but also plausible because interactions
440 between the functional/non-functional alleles of *Hd2* and *DTH8* alleles under natural short-day
441 conditions have been reported (Zhang et al. 2019). Thus, these results indicated the usefulness of
442 epistatic QTL mapping using SHAP-XGB for identifying novel interactions.

443 SHAP-XGB may be applicable to studies with denser DNA markers, e.g., interval mapping or
444 GWAS; however, the computational time and required memory size will increase. The increase in
445 computational time depends on the XGBoost fitting and SHAP value calculations. Fitting XGBoost to
446 the rice real data (n = 176 and p = 162) took 0.111 ± 0.055 seconds using a machine with Intel Corei7-

13700k (3.40 GHz), 24 logical cores, and 128GB RAM, and calculation of SHAP and SHAP interaction scores took 0.007 ± 0.014 and 0.316 ± 0.144 seconds, respectively. We generated a feature matrix with n = 1,760 and p = 162,000 from the standard normal distribution and fitted the SHAP-XGB to investigate the scalability of SHAP-XGB. XGBoost took 51.34 seconds, and the calculation of SHAP took 430.89 seconds using the same machine, which is sufficiently practical. However, the calculation of the SHAP interaction scores failed because of a memory allocation error, and this error could not be fixed until p decreased to 2,000. SHAP interaction scores are calculated for all marker combinations in all samples (i.e., an $n \times p \times p$ array is created), and the memory requirement may be a major bottleneck. Thus, it may be necessary to prune markers using, for example, t-tests or sure independence screening (Fan and Lv 2008). In GWAS, population stratification will be another issue. One solution would be to adjust the phenotypic values with population stratification before application.

The lack of thresholds for determining significant markers is a practical drawback for each application of SHAP-XGB (QTL mapping and GWAS). Permutation tests are one possible solution. Permutation tests are generally used to determine the thresholds for IM and CIM (e.g., Broman et al. 2003) and have been applied to detect gene-gene interactions based on the Shapley-based interaction score and neural networks (Cui et al. 2022). However, further investigation is required to determine whether these tests work as expected.

Data availability

Real data are shared at https://github.com/Onogi/HeadingDatePrediction. R scripts for SHAP-XGB are available at https://github.com/Onogi/SHAP-XGB.

Acknowledgements

The authors would like to thank Ryukoku University for providing financial support.

Funding

This study was supported by Ryukoku University.

Conflict of Interest

The authors declare no competing interests.

Author contributions

TI and AO collaborated in analyzing the data and drafting the manuscript. AO conceived the study.

References

Aas K, Jullum M, Løland A (2021) Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell 298:103502. https://doi.org/10.1016/j.artint.2021.103502

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889–890. https://doi.org/10.1093/bioinformatics/btg112

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 785–794. https://doi.org/10.1145/2939672.2939785

Crawford L, Zeng P, Mukherjee S, Zhou X (2017) Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLOS Genet 13:e1006869. https://doi.org/10.1371/journal.pgen.1006869

Cui T, El Mekkaoui K, Reinvall J, Havulinna AS, Marttinen P, Kaski S (2022) Gene-gene interaction detection with deep learning. Commun Biol 5:1238. https://doi.org/10.1038/s42003-022-04186-y

Falconer DS, Mackay TFC (1996). Introduction to quantitative genetics (fourth edition). Pearson Education Ltd, London.

Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc B 70:849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Fujimoto K, Kojadinovic I, Marichal J (2006) Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. Games Econ Behav 55:72–99. https://doi.org/10.1016/j.geb.2005.03.002

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinform 12:186. https://doi.org/10.1186/1471-2105-12-186

Hori K, Matsubara K, Yano M (2016) Genetic control of flowering time in rice: Integration of Mendelian genetics and genomics. Theor Appl Genet 129:2241–2252. https://doi.org/10.1007/s00122-016-2773-4

Jannink JL, Jansen R (2001) Mapping epistatic quantitative trait loci with one-dimensionalgenome searches. Genetics 157:445–454. https://doi.org/10.1093/genetics/157.1.445

Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455. https://doi.org/10.1093/genetics/136.4.1447

Johnsen PV, Riemer-Sørensen S, DeWan AT, Cahill ME, Langaas M (2021) A new method for exploring gene-gene and gene-environment interactions in GWAS with tree ensemble methods and SHAP values. BMC Bioinform 22:230. https://doi.org/10.1186/s12859-021-04041-7

Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152: 1203–1216. https://doi.org/10.1093/genetics/152.3.1203

Koo BH, Yoo SC, Park JW, Kwon CT, Lee BD, An G, Zhang Z, Li J, Li Z, Paek NC (2013) Natural variation in *OsPRR37* regulates heading date and contributes to rice cultivation at a wide range of latitudes. Mol Plant 6:1877–1888. https://doi.org/10.1093/mp/sst088

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199. https://doi.org/10.1093/genetics/121.1.185

Laurie C, Wang S, Carlini-Garcia LA, Zeng ZB (2014) Mapping epistatic quantitative trait loci. BMC Genet 15:112. https://doi.org/10.1186/s12863-014-0112-9

Li H, Ye G, Wang J (2007) A modified algorithm for the improvement of composite interval mapping. Genetics 175:361-374. https://doi.org/10.1534/genetics.106.066811

Li H, Ribaut JM, Li Z, Wang J (2008) Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. Theor Appl Genet 116:243–260. https://doi.org/10.1007/s00122-007-0663-5

Li Z, Sillanpää MJ (2012) Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190:231–249. https://doi.org/10.1534/genetics.111.134866

Lin HX, Yamamoto T, Sasaki T, Yano M (2000) Characterization and detection of epistatic interactions of 3 QTLs, *Hd1*, *Hd2*, and *Hd3*, controlling heading date in rice using nearly isogenic lines. Theor Appl Genet 101:1021–1028. https://doi.org/10.1007/s001220051576

Shapley LS (1953) A value for n-person games. Contributions to the Theory of Games 2.28, pp. 307–317

Lundberg SM, Lee S (2017) A unified approach to interpreting model predictions. Nips'17. Proceedings of the 31st international conference on neural information processing systems 4768–4777. Available from https://dl.acm.org/doi/10.5555/3295222.3295230

Lundberg SM, Erion GG, Lee S (2018) Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888. https://doi.org/10.48550/arXiv.1802.03888

Meng L, Li H, Zhang L, Wang J (2015) QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. Crop J 3:269–283. https://doi.org/10.1016/j.cj.2015.01.001

Ogiso E, Takahashi Y, Sasaki T, Yano M, Izawa T (2010) The role of casein kinase II in flowering time regulation has diversified during evolution. Plant Physiol 152:808–820. https://doi.org/10.1104/pp.109.148908

Onogi A, Watanabe M, Mochizuki T, Hayashi T, Nakagawa H, Hasegawa T, Iwata H (2016) Toward integration of genomic selection with crop modelling: The development of an integrated approach to predicting rice heading dates. Theor Appl Genet 129:805–817. https://doi.org/10.1007/s00122-016-2667-5

Onogi A, Arakawa A (2022) An R package VIGoR for joint estimation of multiple linear learners with variational Bayesian inference. Bioinformatics 38:3306–3309. https://doi.org/10.1093/bioinformatics/btac328

Ribeiro MT, Singh S, Guestrin C (2016) 'Why should i trust you?' Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 1135–1144. https://doi.org/10.1145/2939672.2939778

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 12:77. https://doi.org/10.1186/1471-2105-12-77

Sundararajan M, Dhamdhere K, Agarwal A (2020) The Shapley Taylor interaction index. Proceedings of the machine learning research 119:9259–9268. Available from https://proceedings.mlr.press/v119/sundararajan20a.html

Takahashi Y, Shomura A, Sasaki T, Yano M (2001) *Hd6*, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the alpha subunit of protein kinase CK2. Proc Natl Acad Sci U S A 98:7922–7927. https://doi.org/10.1073/pnas.111136798

Wang J, Joshi T, Valliyodan B, Shi H, Liang Y, Nguyen HT, Zhang J, Xu D (2015) A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. BMC Genom 16:1011. https://doi.org/10.1186/s12864-015-2217-6

Wang S, Basten CJ, Zeng ZB (2012) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC. (http://statgen.ncsu.edu/qtlcart/WQTLCart.htm)

Wei H, Wang X, Xu H, Wang L (2020) Molecular basis of heading date control in rice. aBIOTECH 1:219–232. https://doi.org/10.1007/s42994-020-00019-w

Wei X, Xu J, Guo H, Jiang L, Chen S, Yu C, Zhou Z, Hu P, Zhai H, Wan J (2010) *DTH8* suppresses flowering in rice, influencing plant height and yield potential simultaneously. Plant Physiol 153:1747–1758. https://doi.org/10.1104/pp.110.156943

Yano M, Harushima Y, Nagamura Y, Kurata N, Minobe Y, Sasaki T (1997) Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. Theor Appl Genet 95:1025–1032. https://doi.org/10.1007/s001220050658

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468. https://doi.org/10.1093/genetics/136.4.1457

Zhang B, Liu H, Qi F, Zhang Z, Li Q, Han Z, Xing Y (2019) Genetic interactions among *Ghd7*,

Ghd8, *OsPRR37* and *Hd1* contribute to large variation in heading date in rice. Rice (N Y) 12:48. https://doi.org/10.1186/s12284-019-0314-x

Zhang YM, Xu S (2005) A penalized maximum likelihood method for estimating epistatic effects of QTL. Heredity 95:96–104. https://doi.org/10.1038/sj.hdy.6800702