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

Abstract

Mapping quantitative trait loci (QTLs) is one of the major goals of quantitative genetics; however, identifying the interactions between QTLs (i.e., epistasis) remains challenging. Recently developed machine learning methods, such as deep learning and gradient boosting, are transforming the real world. These methods could advance QTL mapping methodologies because of their high capability for capturing complex relationships among features. One problem with applying such complex models to QTL mapping is evaluation of feature importance. In this study, XGBoost, a popular gradient tree boosting algorithm, was applied for QTL mapping in biparental populations with Shapley additive explanations (SHAPs). SHAP is a local (i.e., instance-wise) importance index with the desired properties as feature importance indices. The SHAP-assisted XGBoost (SHAP-XGB) was compared with conventional methods, including likelihood ratio tests (LRT), composite interval mapping (CIM), multiple interval mapping (MIM), and BayesC, using simulations and rice heading date data. SHAP-XGB performed comparable to CIM, MIM, and BayesC in mapping main QTL effects and was superior to conventional methods in mapping QTL interaction effects. As SHAP can evaluate local importance, interactions between markers can be visualized by plotting SHAP interaction values for each instance (plant/line). These results illustrated the strength of SHAP-XGB in detecting and interpreting epistatic QTLs and suggest the possibility that SHAP-XGB complements conventional methods.

Keywords

Association mapping, gene interaction, variable selection, SHAP, explainability

Key message

SHAP-assisted XGBoost is useful for mapping and interpreting QTL interactions and can complement conventional methods, such as multiple interval mapping and Bayesian regressions.

Introduction

Detecting quantitative trait loci (QTLs) underpinning variations in quantitative traits is one of the major goals in analyzing quantitative traits. Detecting QTLs can lead to useful applications such as efficient selection based on QTL genotypes, genomic engineering to modify phenotypes, development of medical technologies by understanding molecular mechanisms, 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) and inclusive CIM (Li et al. 2008). Bayesian regression, where whole genome-wide markers are used as explanatory variables simultaneously, were also applied in detecting epistatic QTLs (Zhang and Xu 2005; Li and Sillanpää 2012). 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. 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. For example, 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. The GWASs mentioned above partially prove this possibility; however, evaluation studies are limited. In particular, the performance and properties in mapping populations, such as bi-parental populations, have not been elucidated, and comparisons with conventional methods, such as CIM, MIM, and Bayesian regressions, have not been conducted. In this study, we evaluated the potential of XGBoost for QTL and epistatic QTL mapping using SHAP. 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

Extreme gradient boosting

XGBoost is based on decision trees and represents the response of instance *i* (the phenotype of plant/line *i* in the case of QTL mapping), y_i, with the linear combination of *K* decision trees as

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),$$

where f is a tree with structure *q* and leaf weight *w* and x_i is the vector of feature (marker genotypes in the case of QTL mapping). In other words, y_i is represented as the sum of the outputs of multiple decision trees. The decision trees are sequentially added at each iteration. Thus, the objective function of XGBoost at iteration *t*, that is, when adding *t*th tree, can be written as

$$L^t = \sum_{i=1}^{n} l\left(y_i,\ \hat{y}_i^{t-1} + f_t(x_i)\right) + \Omega(f_t),$$

where *n* is the number of instances (plants/lines), l is a loss function, $\hat{y}_i^{t-1}$ is $\hat{y}_i$ at iteration $t-1$, f_t is the tree added at iteration *t*, and Ω is a penalty function on the tree structure (i.e., the number of leaves and weights). A penalty function is added to avoid overfitting the model. By approximating f with Taylor expansions and dropping the constant terms, the objective function becomes

$$\tilde{L}^t = \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T,$$

where g_i and h_i are the first and second order gradients of the loss function, I_j is the set of instances assigned to leaf *j*, *T* is the number of leaves, and λ and γ are the regularization parameters. The optimal weight of leaf *j* (w_j^*) and the corresponding optimal value of $\tilde{L}^{t*}$ are

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \text{ and } \tilde{L}^{t*} = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T,$$

respectively. Using these results, whether a leaf I should be split into I_L and I_R can be evaluated with the following equation:

$$\frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma.$$

This equation was derived by subtracting $\tilde{L}^{t*}$ when leaf I was split into I_L and I_R from $\tilde{L}^{t*}$ when leaf I was a single leaf. Thus, when this equation is positive, $\tilde{L}^{t*}$ decreases because of splitting. XGBoost uses this formula to search for the tree structures to be added, making the method fast and scalable. The authors proposed exact greedy and approximate algorithms for this type of search.

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, and rewards are replaced with the model output

(e.g., fitter/predicted phenotypes) 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 is

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

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 and illustrated in Fig. 1.

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.

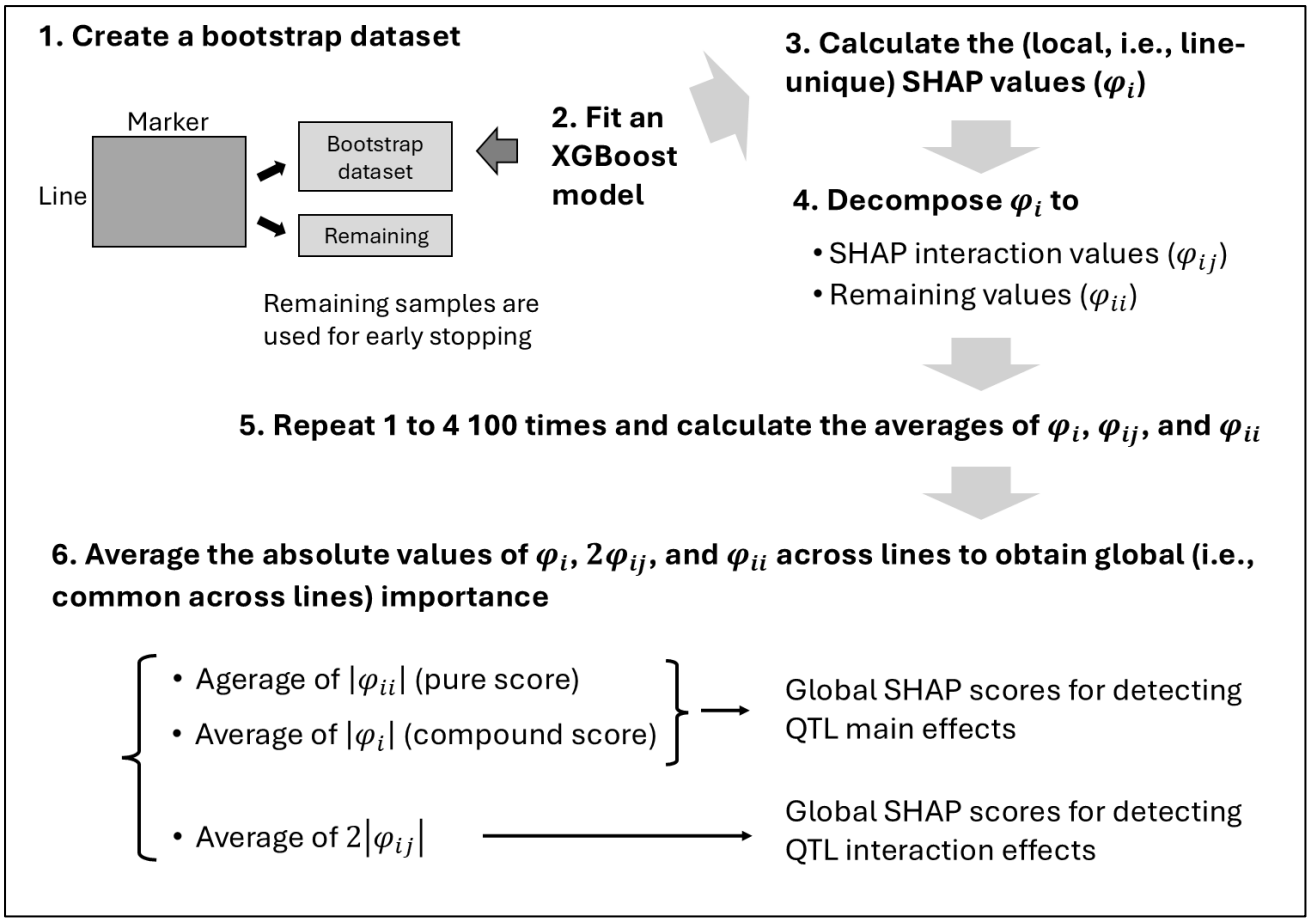

Fig. 1 Illustrations of the Shapley additive explanations-assisted XGBoost (SHAP-XGB) algorithm.

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, as shown later. We also attempted settings that lead to underfitting and overfitting and compared them (see the "Algorithm assessment"

202 section). Lastly, there would be two choices for detecting the main effects of QTLs: using SHAP values
203 that contain interaction scores (i.e., φ_i) or using SHAP values from which interaction scores are
204 removed (i.e., φ_{ii}). In this study, the former score is referred to as "compound" scores, and the latter
205 is referred to as the "pure" scores for detecting QTL main effects. These two scores were compared as
206 described in the "Algorithm assessment" section.
207
208 **Comparing methods**
209 The SHAP-XGB method was compared with other methods using simulated and real data. These
210 methods were selected based on their prevalence in the QTL analyses of the inbred line crosses. Note
211 that likelihood ratio tests (LRT), BayesC, and SHAP-XGB used only marker genotypes as covariates
212 and did not use the genotype probabilities of the loci between markers. For the CIM and MIM, the
213 intervals were set to 1 cM.
214 **Likelihood ratio test (LRT)**
215 LRT was conducted using the interactionPval function of the R package SNPassoc (ver. 2.1.0,
216 González et al. 2007). The function is based on linear regression and simultaneously performs several
217 tests, including LRTs, for the main QTL and interaction effects. The likelihood ratios of these tests can
218 be written as $\frac{L_m(i)}{L_0}$, and $\frac{L_f(i+j+i*j)}{L_m(i+j)}$, respectively, where L_m indicates the likelihood of the models that
219 include the main effects of markers i and j, L_0 indicates the likelihood of the null model (i.e., model
220 without marker covariates), and L_f indicates the likelihood of the full model that include the
221 interaction effects between markers i and j.
222 **Composite interval mapping (CIM)**
223 CIM was performed using the cim function of the R package qtl (ver. 1.60, Broman et al. 2003). The
224 default parameter values were used. Briefly, the expectation-maximization algorithm was used, three
225 covariates were used, and the window size was 10 cM. CIM was used to detect the main QTL effects.
226 We assessed the influence of the number of covariates and window size on the simulated and real data

(Figs. S1 and S2) and found that the influence was negligible.

Multiple interval mapping

MIM was conducted using QTL cartographer (ver. 2.5_011, Wang et al. 2012). The MIM theory was first proposed by Kao et al. (1999). We followed the algorithm of Laurie et al. (2014) which was illustrated in the section "Instruction for using the procedures in Windows QTL Cartographer" of the article. The algorithm consists 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%). These steps require manual program operation, which hampers massive simulation analysis. The program abnormally ended in total nine replications of simulations. However, the causes were unknown because it seemed to happen randomly. These replications were excluded when summarizing the performance of the MIM.

BayesC

BayesC (Habier et al. 2011) was performed using the vigor function of R package VIGoR (ver. 1.1.0, Onogi and Arakawa 2022). This package adopts fast variational Bayesian algorithms to estimate the posterior distributions of Bayesian alphabets. For genotype *i*, the model can be written as:

$$y_i = \mu + \sum_p \beta_p x_{p,i} + \sum_p \sum_{k \neq p} \beta_{pk} x_{p,i} x_{k,i} + e_i$$

where y_i is the phenotypic values, μ is the intercept, β_p is the effect of marker *p*, $x_{p,i}$ is the genotype of marker *p*, β_{pk} is the interaction effect between marker *p* and *k*, and e_i is the residual. The prior distributions of β_p and β_{pk} were

$$\beta_p \sim N(0,\ \sigma_M^2) \text{ if } \gamma_p = 1, \text{ and } \beta_p = 0 \text{ if } \gamma_p = 0, \text{ and}$$

$$\beta_{pk} \sim N(0,\ \sigma_I^2) \text{ if } \gamma_{pk} = 1, \text{ and } \beta_{pk} = 0 \text{ if } \gamma_{pk} = 0,$$

where γ_p and γ_{pk} are the marker and marker combination inclusion probabilities, respectively. The prior distributions of γ_p and γ_{pk} were the Bernoulli distributions with hyperparameters $\kappa_M = 0.05$ and $\kappa_I = 0.001$, respectively. Default values were used for the other hyperparameters. Bootstrapping was applied to make the inferences on the inclusion probability more robust. That is, the model was

252 fitted to bootstrap samples, and the posterior means of the inclusion probabilities were averaged across
253 the bootstrap samples. In this study, 1000 bootstrap samples were used.
254
255 **Simulations**
256 The usefulness of SHAP-XGB was first investigated using simulations that mimicked biparental F_2
257 and RIL populations. The conditions of the simulations were as follows: The number of chromosomes
258 was 8, the length of each chromosome was 120 cM, the number of markers (SNPs) per chromosome
259 was 20, the number of QTLs was five, and the number of individuals was 200. The QTLs were assigned
260 so as to be one QTL per chromosome, and chromosomes harboring the QTLs were randomly selected.
261 Among the 10 combinations of QTLs, half were assumed to have interaction effects, and the
262 combinations with interaction effects were randomly selected. The main and interaction effects of the
263 QTLs were generated using the standard normal distribution. The broad-sense heritability was set at
264 0.8.
265 We simulated six scenarios using these conditions: (1) F_2 population with only additive effects
266 (QTL genotypes were coded as −1, 0, and 1). (2) RIL population with additive effects only. (3) F_2
267 population with only complete dominance effects (QTL genotypes were coded as 0, 1, and 1). (4) F_2
268 population with only overdominance effects (QTL genotypes were coded as 0, 1, and 0). (5) F_2
269 population with only additive effects and with epistatic variances increased by setting the variance of
270 interaction effects to 16. (6) F_2 population had only additive effects, but the number of QTLs with
271 nonzero additive effects was limited to two (i.e., three QTLs only had interaction effects without main
272 effects). (7) F_2 population with only additive effects, of which the size was increased to 400. (8) RIL
273 population with only additive effects, of which the size was increased to 400. When only additive
274 (dominance) effects were simulated, interaction effects were simulated as additive-by-additive
275 (dominance-by-dominance) effects by multiplying the corresponding QTL genotypes. A summary of
276 the simulation scenarios and realized proportions of variances due to the additive and interaction
277 effects are presented in Table 1. The QTLs and marker genotypes were generated using the R package

qtl (ver. 1.60). QTL effects and phenotypes were generated using house-developed R scripts. The number of replications per scenario was 10, which was determined by considering the MIM manual operations.

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 p values and LOD scores were regarded as the signals for LRT and CIM, respectively, and the bootstrap means of the inclusion probabilities (γ_p and γ_{pk}) 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.

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

Scenario	Pop.	QTL effects[2]		Size	Proportion of variance[3]		Purpose	Results
		Main	Interaction		V_A	V_I		
1	F_2	A	A by A	200	0.476 ± 0.104	0.328 ± 0.112	Evaluation of QTL detection power	Figs. 2, 3, and 4
							Comparison of pure and compound SHAP values	Fig. S3
							Comparison of XGB settings	Figs. S4 and S5
							Comparison of CIM parameters	Fig. S1
2	RILs	A	A by A	200	0.408 ± 0.175	0.385 ± 0.176	Evaluation of QTL detection power	Figs. 2, 3, and 4
							Comparison of XGB settings	Figs. S4 and S5
							Comparison of CIM parameters	Fig. S1
3	F_2	CD	CD by CD	200	0.417 ± 0.222	0.407 ± 0.212	Evaluation of QTL detection power (dominance)	Fig. S6
4	F_2	OD	OD by OD	200	0.462 ± 0.134	0.322 ± 0.130	Evaluation of QTL detection power (dominance)	Fig. S6
5	F_2	A	A by A	200	0.115 ± 0.037	0.679 ± 0.032	Comparison of pure and compound SHAP values	Fig. S3
6[4]	F_2	A	A by A	200	0.313 ± 0.178	0.476 ± 0.176	Comparison of pure and compound SHAP values	Fig. S3
							Evaluation of false positive inflation in main effect mapping due to epistasis	Fig. S7
7	F_2	A	A by A	400	0.630 ± 0.091	0.175 ± 0.087	Evaluation of QTL detection power in larger populations	Figs. 2, 3, and 4
							Comparison of CIM parameters	Fig. S1
8	RILs	A	A by A	400	0.383 ± 0.218	0.423 ± 0.216	Evaluation of QTL detection power in larger populations	Figs. 2, 3, and 4
							Comparison of CIM parameters	Fig. S1

296 [1]All populations were assumed to be derived from biparental crosses. SHAP: Shapley additive
297 explanations, XGB: Xgboost.
298 [2]A: additive, CD: complete dominance, OD: overdominance.
299 [3]V_A: additive variance, V_I: epistasis variance.
300 [4]Scenario 6 simulated five QTLs, of which three only had interaction effects.
301

Mapping dominance effects

Whereas the cim function of the qtl package and MIM by the QTL cartographer test the additive and dominance effects automatically, LRT by the interactionPval function, BayesC by vigor, and SHAP-XGB need a specification which effect is targeted. For LRT, the argument "model" of the function was set as "co" (codominant) to test the additive effects, "do" to test the dominance effects, and "ov" to test the overdominance effects. For BayesC and SHAP-XGB, genotype codes were modified depending on the target effects. The genotypes were coded as [−1, 0, 1], [0, 1, 1], and [0, 1, 0] to test the additive, complete dominance, and overdominance effects, respectively.

MIM tested multiple combinations of interaction effects, that is, additive-by-additive, additive-by-dominance, and dominance-by-dominance. Thus, even when only dominance-by-dominance interactions were simulated, the results occasionally included additive-by-additive or additive-by-dominance interactions for the correct marker combinations. Because it was too strict to regard these results as false positives, they were considered true positives.

Algorithm assessment

The performance of the compound and pure global SHAP scores in detecting QTL main effects was compared using simulation scenarios (1) (F_2 population with only additive effects), (5) (F_2 population with only additive effects and with increased epistatic variances), and (6) (F_2 population with only additive effects, but the number of QTLs that have non-zero additive effects was limited to two).

Two additional settings of XGBoost were compared to investigate the effects of the degree of fit of the XGBoost models to the data on the QTL detection power. In the underfitting setting, the learning rate "eta" was set to 0.1 (default is 0.3) and the "max_depth," which specifies the maximum depth of trees, was set to 3 (the default is 6). The "early_stopping_rounds" was set to 3. In the overfitting setting, the hyperparameters were the default values and the number of rounds (boosting iterations) was set to 50 without early stopping. Simulation Scenarios 1 and 2 were used for comparison.

328 **Real data**

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

334

335 **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

336 [1]Details of these environments are provided in Table S1 of Onogi et al. (2016).

337

338 **Dependency plots**

339 SHAP interaction values were calculated for each instance (line) because SHAP was originally
340 proposed as a local importance measure. Thus, the interactions between QTLs (markers) can be
341 visualized by plotting the SHAP interaction values of the marker combinations of interest for all lines.
342 These plots are referred to as dependency plots. Dependency plots were constructed using the R
343 package SHAPforxgBoost (ver. 0.1.1) with some modifications.

Results

There are two choices when applying SHAP values for mapping QTL main effects: using compound global SHAP scores that contain interaction scores or using pure global SHAP scores from which interaction scores are removed. We first compared these two types of values for mapping the QTL main effects in simulation Scenarios 1, 5, and 6. Scenario 1 simulated the F_2 populations with additive and additive-by-additive interactions. Scenarios 1 and 5 differed in the proportion of phenotypic variance explained by epistatic variance (0.328 vs. 0.679 on average; Table 1), and Scenarios 1 and 6 differed in the number of QTLs that had main effects (5 vs. 2; Table 1). In Scenario 6, three QTLs only had interaction effects. The AUC values and ROC curves were comparable between the compound and pure scores (Fig. S3). A slight increase in the detection power was observed for the compound scores in Scenario 5. In contrast, the false positive rates increased slightly for the compound scores in Scenario 6. This is reasonable because the three QTLs lacked main effects and only had interaction effects in Scenario 6. Thus, we concluded that the compound SHAP scores were comparable with or even better than the pure SHAP scores in QTL mapping as long as the QTLs that have non-zero main effects are involved in the interactions. Hereafter, we used compound scores to map the main QTL effects.

Next, we investigated the effect of the fitting degree of the XGBoost models on QTL mapping accuracy. We compared the three settings that led to underfitting, moderate fitting, and overfitting in simulation Scenarios 1 and 2. The SHAP values were either underestimated or overestimated when underfitting or overfitting was induced (Fig. S4). Nevertheless, as the relative values between the markers did not vary significantly, the AUC values and ROC curves were hardly changed by the fitting status (Fig. S5). We used a setting that induced a moderate fit.

In comparative analyses mapping main QTL effects, SHAP-XGB generally showed higher or comparable AUC values than LRT, CIM, and BayesC (Fig. 2). When compared with MIM, SHAP-XGB showed true positive rates comparable to MIM at the false positive rates that MIM showed (Fig. 2). In comparative analyses mapping QTL interaction effects, SHAP-XGB showed significantly higher AUC values than LRT and BayesC (Fig. 3). When compared with MIM, SHAP-XGB showed higher

true positive rates than MIM under false positive rates comparable to those of MIM (Fig. 3), suggesting that SHAP-XGB can detect more QTL interactions than MIM under comparable false positive rates. Interactions between markers (QTLs) can be interpreted using dependency plots. The SHAP interaction values (local importance metrics) for a replication of Scenario 2 are shown in Fig. 4. 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.

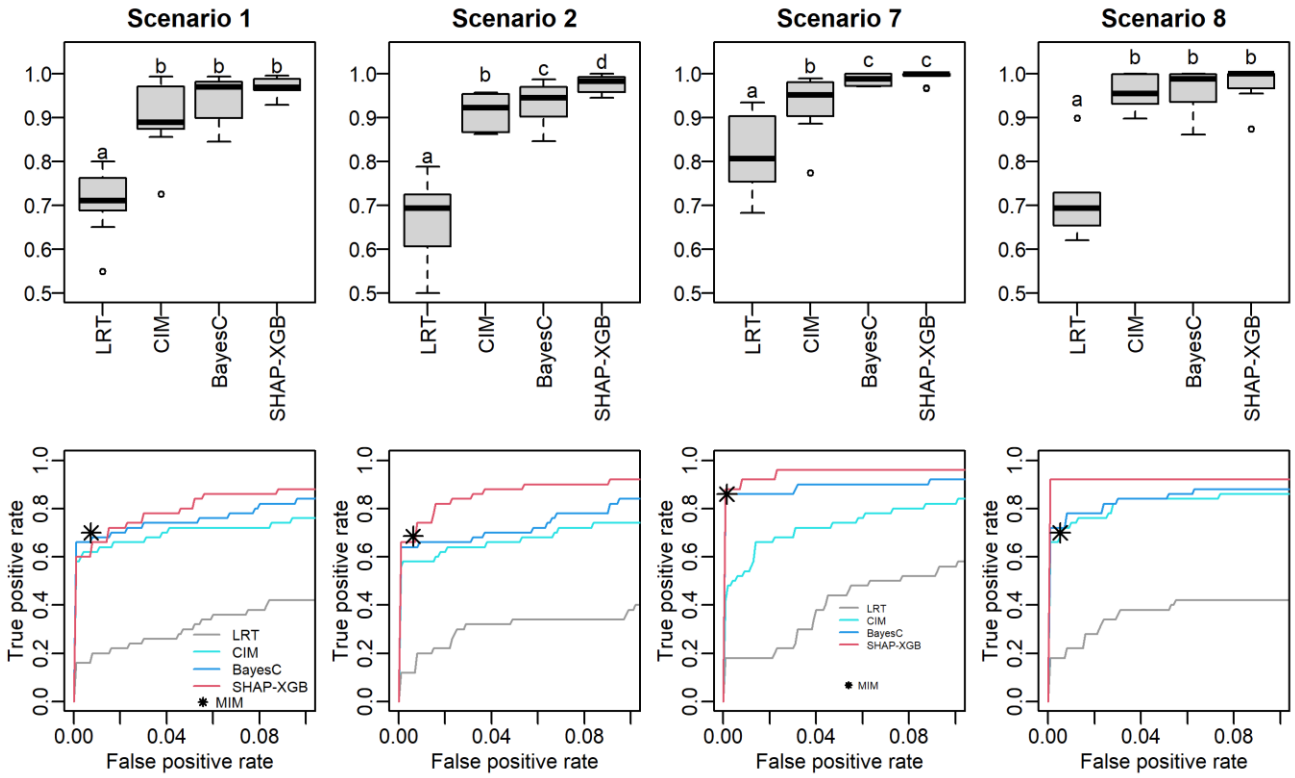

Fig. 2 Area under the curve (AUC) values and receiver operation characteristics (ROC) curves in mapping main QTL effects. The top and bottom panels are the AUC values and ROC curves, respectively. The ROC curves are the mean curves of 10 replications. Scenarios 1 and 7 simulated F_2 populations, and Scenarios 2 and 8 simulated recombinant inbred line populations. The population sizes were 200 for Scenarios 1 and 2, and 400 for Scenarios 7 and 8. 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. LRT: likelihood ratio test, CIM: composite interval mapping, SHAP-XGB: Shapley additive explanations-assisted XGBoost, MIM: multiple interval mapping.

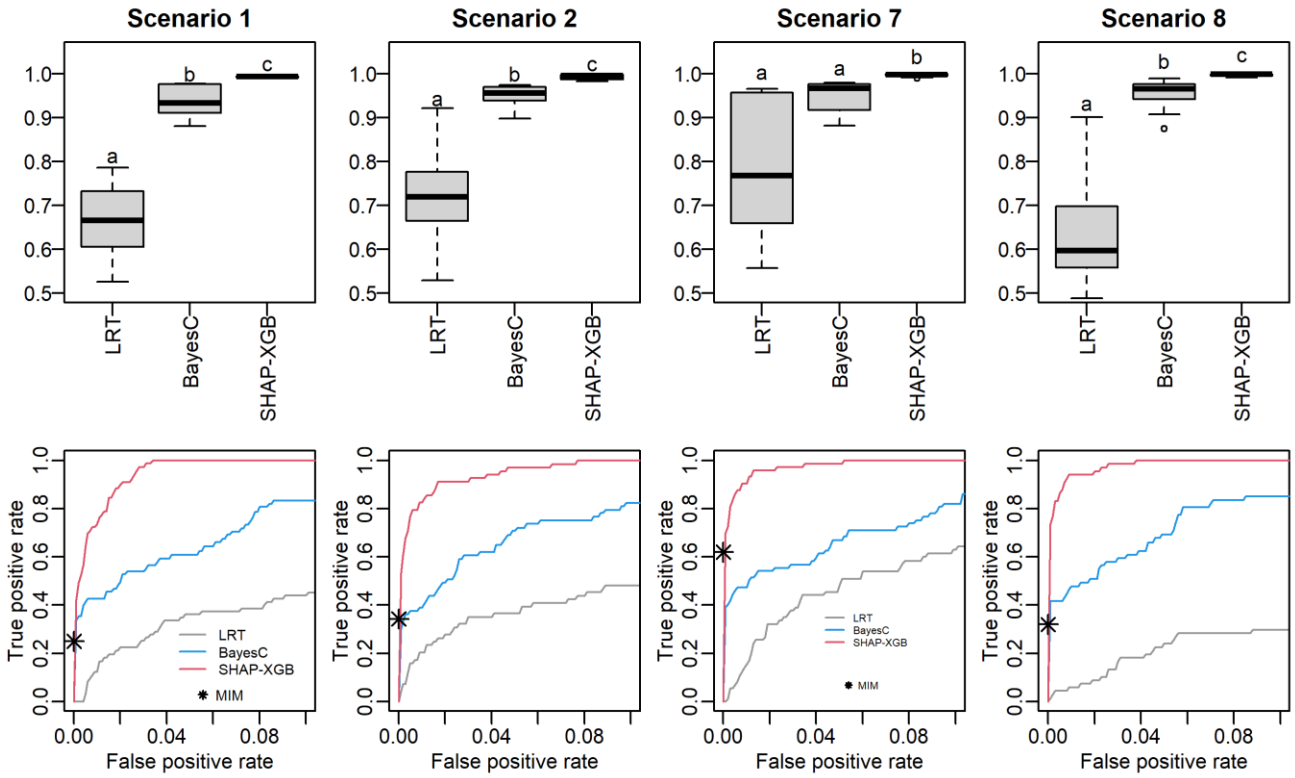

Fig. 3 Area under the curve (AUC) values and receiver operation characteristics (ROC) curves in mapping QTL interaction effects. The top and bottom panels are the AUC values and ROC curves, respectively. The ROC curves are the mean curves of 10 replications. Scenarios 1 and 7 simulated F_2 populations and Scenarios 2 and 8 simulated recombinant inbred line populations. The population sizes were 200 for Scenarios 1 and 2, and 400 for Scenarios 7 and 8. 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. LRT: likelihood ratio test, SHAP-XGB: Shapley additive explanations-assisted XGBoost, MIM: multiple interval mapping.

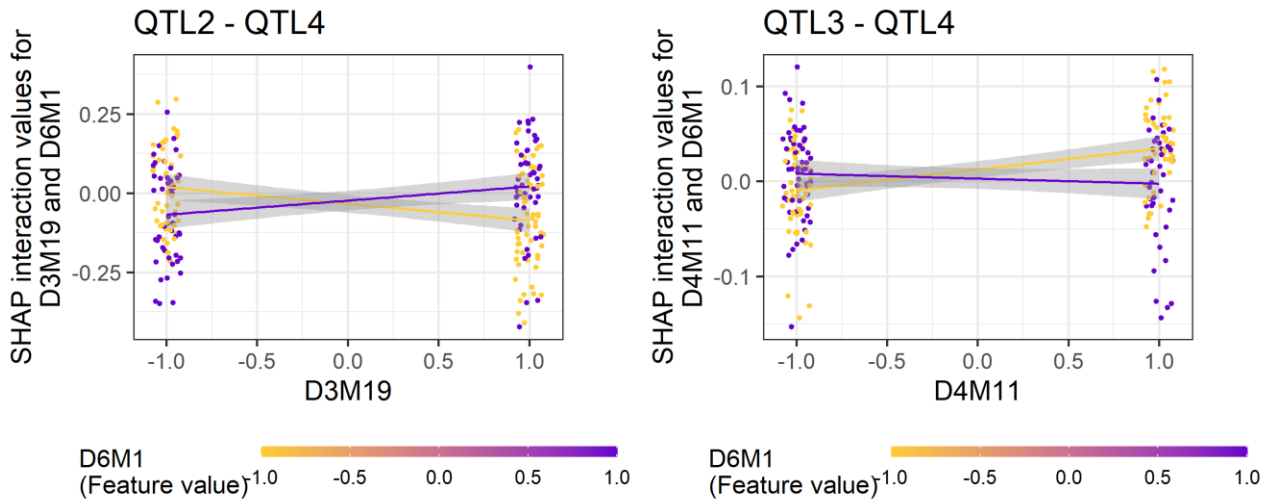

Fig. 4 Dependencies between QTLs in a simulated dataset. The Shapley additive explanations (SHAP) interaction values of each line are shown as dots. The left panel shows the interaction between the 2nd and 4th QTLs, and the right panel shows that between the 3rd and 4th QTLs. *D3M19*, *D4M11*, and *D6M1* 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.

In mapping dominance effects (Scenarios 3 and 4), SHAP-XGB showed comparable or only slightly decreased AUC values when additive effects were targeted compared to when dominance effects were targeted (Fig. S6). These results are in contrast to those of BayesC, where misspecification of the target effects resulted in a more prominent decrease in accuracy. These results suggest that SHAP-XGB was insensitive to the effect type. This characteristic was also observed when dominance-by-dominance interactions were targeted. In other words, dominance-by-dominance effects can be detected even when additive-by-additive effects were targeted.

Even when QTLs show only interaction effects, it may be likely that they are detected as having additive (main) effects. Thus, false positives in the detection of QTLs with additive effects may increase in the presence of epistasis. This possibility can be assessed by comparing Scenarios 1 and 6

422 because these scenarios differed in the number of QTLs that had main (additive) effects; in Scenario
423 1, five QTLs had both main and interaction effects, and in Scenario 6, although two QTLs had both
424 main and interaction effects, the remaining three QTLs had only interaction effects. However, the
425 results showed no evidence of false positives due to epistasis for each method in detecting QTLs with
426 additive effects (Fig. S7).

427 The XGBoost models explained 94.3% (± 1.9%) of the phenotypic variance with the setting of
428 moderate fitting in the real rice data analyses. SHAP-XGB and the compared methods showed similar
429 results in mapping the main QTL effects (Fig. S8, Fig. 5, and Table S1), and markers linked to *Hd6*
430 (chr. 3, Takahashi et al. 2001), *Hd1* (chr. 6, Yano et al. 1997), and *DTH8* (chr. 8, Wei et al. 2010)
431 showed higher signals for each method. The BayesC results were generally stringent (or conservative)
432 when mapping epistatic QTLs (Fig. S9 and Fig. 6). LRT showed wide positive regions for each
433 interacting combination, probably because of linkage disequilibrium (LD) between the markers (Fig.
434 S9 and Fig. 6). In contrast, SHAP-XGB could pinpoint interacting combinations within reasonably
435 narrow margins. The global SHAP interaction scores in each environment showed interesting trends
436 across latitudes (Fig. 7). Interactions between *Hd1* and *Hd6*, *Hd1* and *DTH8*, and *Hd1* and
437 *Hd2*/*OsPRR37* (Yano et al. 1997; Koo et al. 2013) showed higher scores at higher latitudes
438 (environments Tsukuba2007, Tsukuba2008E, Tsukuba2009, Ishikawa2008, and Fukuoka2008), but
439 lower scores at lower latitudes (environments Ishigaki2008, HaNoi2008, and ThaiNguyn2008). In
440 contrast, the interaction between *DTH8* and *Hd2* showed larger scores at lower latitudes but lower
441 scores at higher latitudes. As suggested by the simulation analyses (Fig. 3), the results of MIM in the
442 mapping interactions were relatively stringent; nine interactions were detected in nine environments
443 (Table S1). The detected interactions were generally consistent with those detected using SHAP-XGB;
444 for example, the interaction between *Hd6* and *Hd2* in Tsukuba2008L, the interactions between *Hd1*
445 and *Hd6*, *Hd6* and *Hd2*, and *DTH8* and *Hd2* in Fukuoka2008, and the interactions between *HD6* and
446 *Hd2* and *DTH8* and *Hd2* in Ishigaki2008 (Table S1 and Fig. 7). However, trends across latitudes were
447 missing from the MIM results.

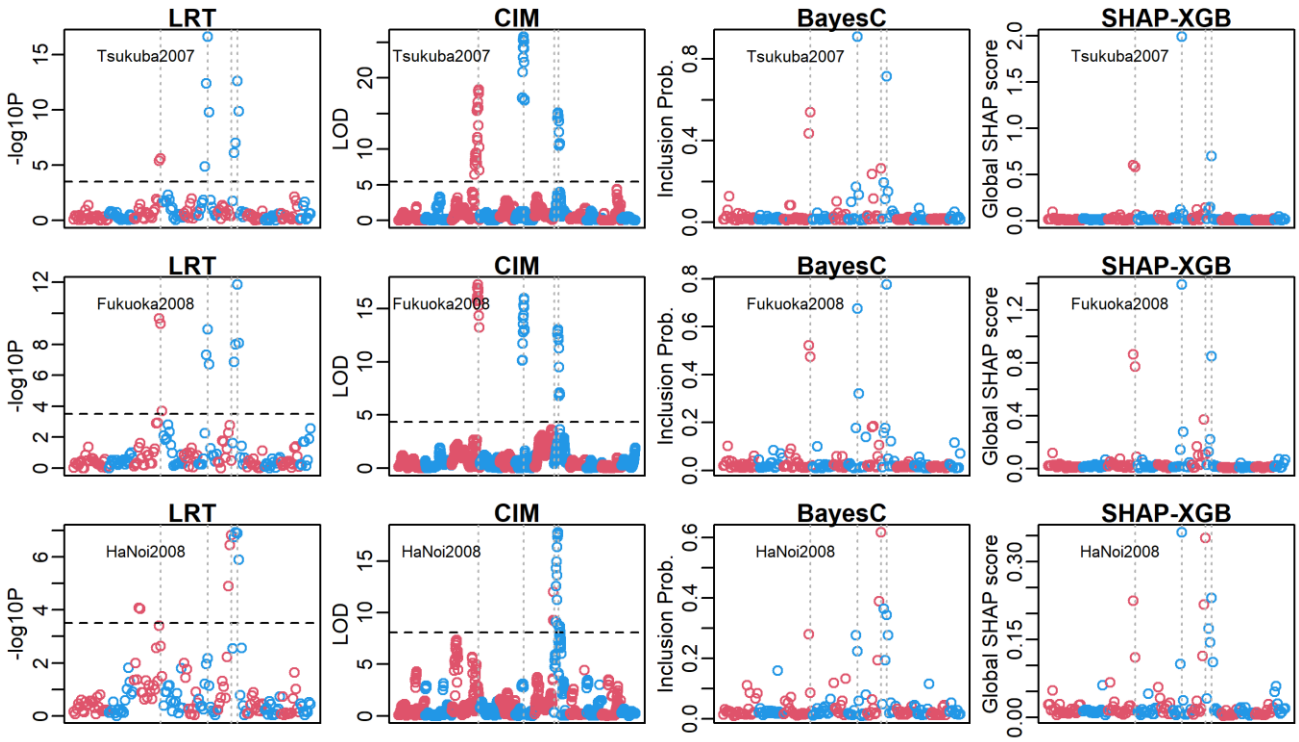

Fig. 5 Comparison of QTL mapping methods to detect main effects in the rice heading date data. The y-axis denotes the QTL signals (LRT, −log10p; CIM, LOD score; BayesC, posterior inclusion probability; SHAP-XGB, global SHAP score). The x-axis denotes the marker positions. The neighboring chromosomes are distinguished using different colors. The vertical bars indicate the four major QTLs (left to right, *Hd6*, *Hd1*, *Hd2*, and *DTH8*). The horizontal bars in LRT and CIM are the 5% thresholds after the Bonferroni correction and the 95% thresholds obtained by permutation tests, respectively. Three representative environments (Tsukuba2007, Fukuoka2008, and HaNoi2008) are presented. The full results are shown in Fig. S8. LRT: likelihood ratio test, CIM: composite interval mapping, SHAP-XGB: Shapley additive explanations-assisted XGBoost.

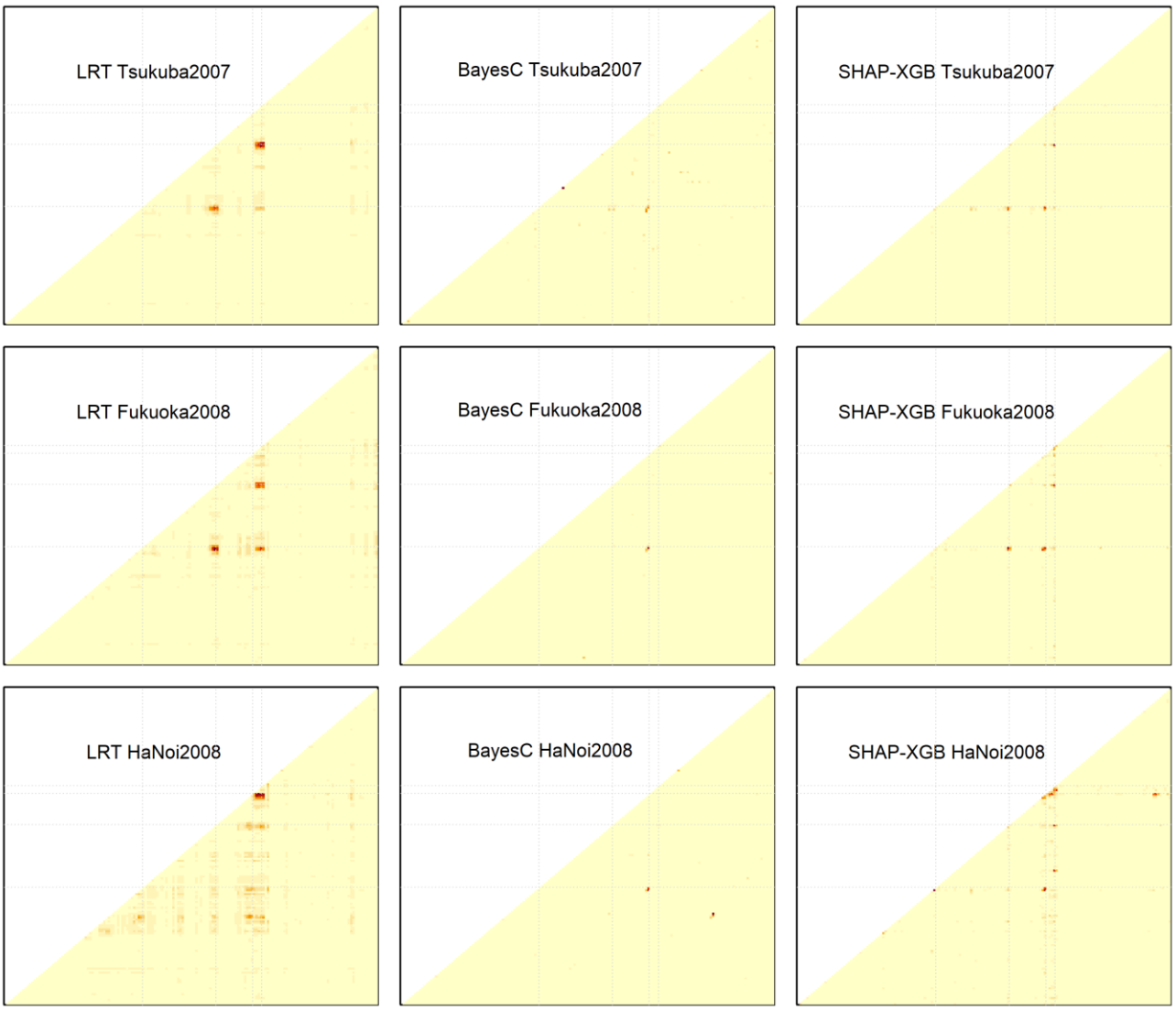

Fig. 6 Comparison of QTL mapping methods to detect interaction effects in the rice heading date data. The y- and x-axes denote the relative positions of markers. The signal levels obtained with each QTL mapping method (LRT, −log10p; BayesC, posterior inclusion probability; SHAP-XGB, global SHAP interaction 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*). Three representative environments (Tsukuba2007, Fukuoka2008, and HaNoi2008) are presented. The full results are shown in Fig. S9. LRT: likelihood ratio test, SHAP-XGB: Shapley additive explanations-assisted XGBoost.

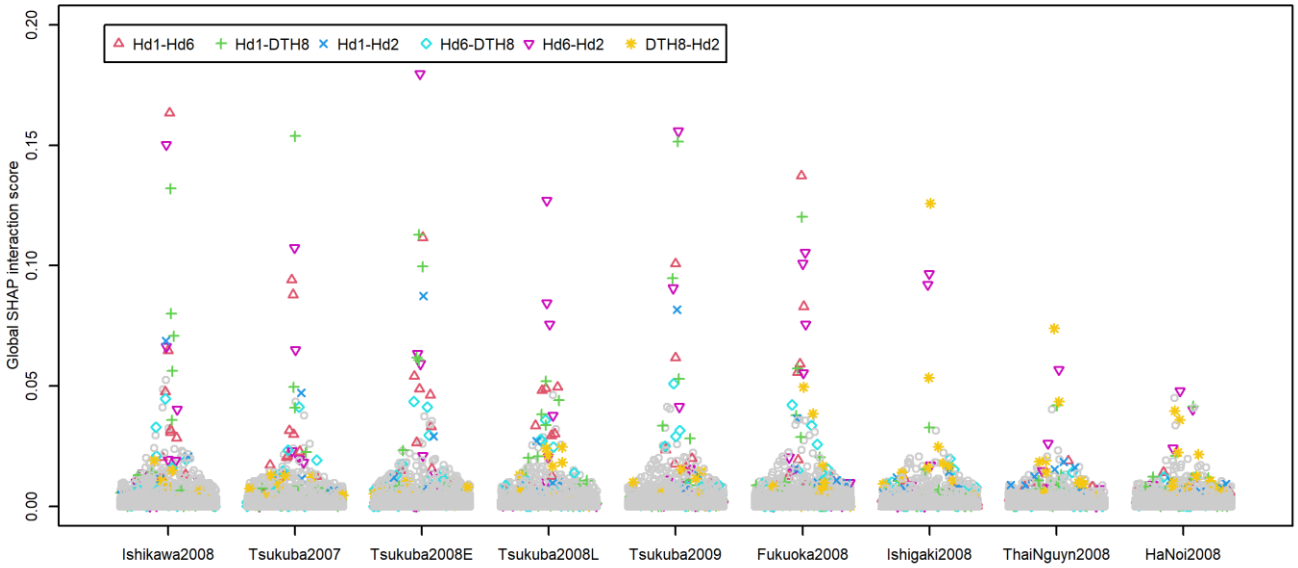

Fig. 7 Summary of the global Shapley additive explanations (SHAP) interaction scores obtained in the rice heading date data. The y-axis denotes the global SHAP interaction scores, which are the averages of the absolute SHAP interaction values across all lines. The global SHAP interaction scores were 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 (Fig. 8). 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.

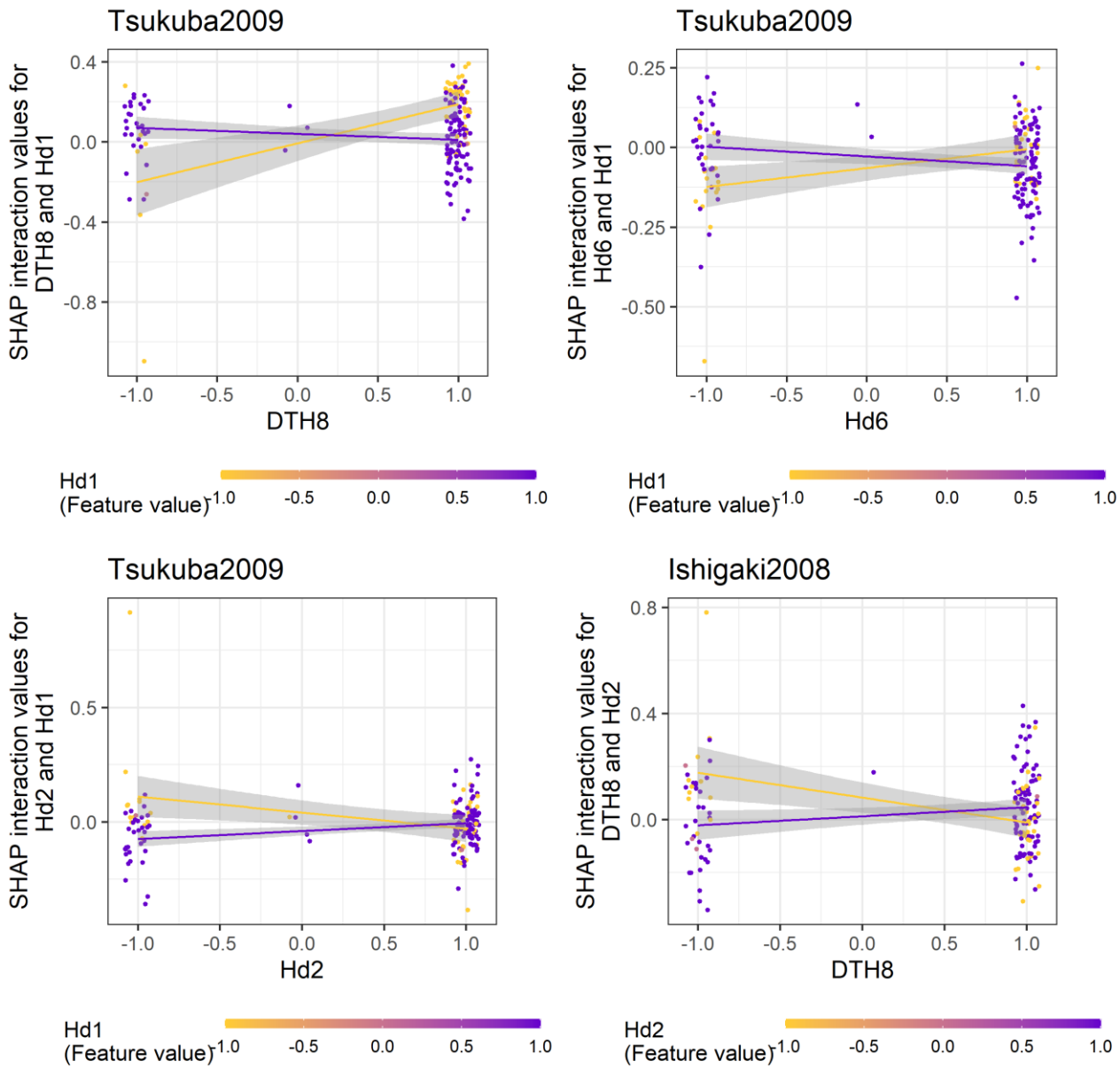

Fig. 8 Dependencies between major QTLs shown with the Shapley additive explanations (SHAP) interaction values. The left top, right top, and left bottom panels show the dependency between *Hd1* and the three QTLs (*DTH8*, *Hd6*, and *Hd2*) at Tsukuba2009, respectively. The right bottom panel shows the 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.

496 **Discussion**

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

503 The simulation results revealed several properties of the SHAP-XGB. First, compound and pure
504 global SHAP scores worked similarly to map the main effects of QTLs. While the compound scores
505 showed slightly increased detection power when epistasis variances were larger, they also showed a
506 slight increase in false positives when QTLs with interaction effects lacked the main effects. Thus,
507 using compound scores may be advantageous as long as QTLs with non-zero main effects are involved
508 in the interactions. Using compound scores may circumvent the reduction in pure scores due to LD, as
509 described later. Second, the degree of fit (i.e., the extent to which the model fits the data) of the
510 XGBoost models did not affect the QTL detection power. This is because the ROC curves and AUC
511 values were calculated based on the relative rankings of markers that did not differ significantly
512 depending on the fitting status. However, the result of over- or underfitting is the inflation or deflation
513 of SHAP values, which in turn may make interpretations of SHAP values difficult. Thus, careful tuning
514 of the fitting degree is recommended to improve interpretability. Finally, SHAP-XGB is robust to the
515 misspecification of target effects, i.e., additive or dominance. That is, even when SHAP-XGB targets
516 additive effects, it can map dominance effects or dominance-by-dominance interaction effects as
517 accurately as the existing methods with correct target specifications. This property may be
518 advantageous; however, it may be disadvantageous in some cases because the gene actions of the
519 detected QTLs cannot be identified.

520 There are ways to measure feature importance for gradient boosting other than SHAP, including
521 "gain" (the average increase of the objective function when the feature is added to trees) and "split

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

531 Although the simulated and real data analyses consistently suggest the usefulness of SHAP-XGB
532 in QTL mapping, the marker genotypes have an undesirable property, i.e., LD between markers, for
533 evaluating feature importance. Because XGBoost randomly samples markers to build a tree f, QTL
534 effects are diluted with the markers linked to the QTLs, consequently decreasing the Shapley values
535 of each linking marker. In fact, the global SHAP scores and SHAP interaction values in Fig. 5 and 7
536 shrank to 0 (2.0 at most) considering that the units of the scores/values are the same as the phenotypic
537 values (i.e., days). These decreased values were probably due to the correlated marker genotypes. The
538 LD between the markers affected another aspect. In rice heading date analyses, interactions between
539 neighboring markers were occasionally assigned large SHAP interaction values, resulting in a
540 reduction in the main pure scores (these interactions were removed from Fig. 7 for ease of
541 visualization). However, such interactions are caused by LD between neighboring markers and thus
542 should be interpreted as part of the main QTL effects. Although this decrease in pure main effects can
543 be circumvented using compound scores for mapping, caution is required when interpreting pure
544 scores. It has also been pointed out that approximations of Shapley values with SHAP become
545 inaccurate when features are correlated (Aas et al. 2021) because the evaluations of the expectation in
546 Eq. (2) assume independence between the features. Thus, taken together, LD would affect the detection
547 power of SHAP-XGB. One solution is to replace SHAP with other indices that consider the

correlations among features (Aas et al. 2021).

Interactions between the major heading date QTLs weakened at lower latitudes (Fig. 7 and Table 2). These results were reasonable because *Hd6*, *Hd2*, and *DTH8* are involved in heading date variation under long-day conditions (Hori et al. 2016; Wei et al. 2020). Interactions between *Hd1* and three QTLs (*Hd6*, *Hd2*, and *DTH8*) are plausible under natural long-day conditions. For example, interactions between functional *Hd6* alleles (e.g., Kasalath allele) and *Hd1* alleles (e.g., Nipponbare allele) were reported under long-day conditions (Ogiso et al. 2010). Interactions between *Hd1* and *Hd2* under long-day conditions (Lin et al. 2000), and interactions among *Hd1*, *Hd2*, *DTH8*, and *Ghd7* under natural long-day conditions were also reported (Zhang et al. 2019). The interactions between *Hd2* and *DTH8*, which is prominent in lower latitudes, are interesting but also plausible because interactions between the functional/non-functional alleles of *Hd2* and *DTH8* alleles under natural short-day conditions have been reported (Zhang et al. 2019). Thus, these results indicated the usefulness of epistatic QTL mapping using SHAP-XGB for identifying novel interactions.

SHAP-XGB may be applicable to GWAS; however, the computational time and required memory size will increase. The increase in computational time depends on the XGBoost fitting and SHAP value calculations. Fitting XGBoost to 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 (SNPs) using, for example, t-tests or sure independence screening (Fan and Lv 2008) in practical

GWAS applications. Population stratification is another issue in GWAS. 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.

Acknowledgements

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

Declarations

Funding

This study was supported by Ryukoku University.

Competing interests

The authors declare no competing interests.

Data availability

Real data are shared at https://github.com/Onogi/HeadingDatePrediction. Simulation data are available upon request.

Code availability

Examples of R scripts for SHAP-XGB are available at https://github.com/Onogi/SHAP-XGB.

Author contributions

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

Ethics approval

Not applicable

Consent to participate

Not applicable

Consent to publication

Not applicable

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

González JR, Armengol L, Solé X, Guinó E, Mercader JM, Estivill X, Moreno V (2007) SNPassoc: An R package to perform whole genome association studies. Bioinformatics 23:644–645. https://doi.org/10.1093/bioinformatics/btm025

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

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, 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

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

Short Legends of Supporting Information

Fig. S1: Comparison of composite interval mapping (CIM) with different hyperparameters in simulations.

Fig. S2: Comparison of composite interval mapping (CIM) with different hyperparameters in the real rice data.

Fig. S3: Comparison of area under the curve (AUC) values and receiver operating characteristics (ROC) curves between compound and pure global SHAP scores.

Fig. S4: Comparison of the compound global SHAP scores obtained using the different XGBoost settings.

Fig. S5: Comparison of the XGBoost settings in the simulated data sets.

Fig. S6: Area under the curve (AUC) values and receiver operation characteristics (ROC) curves in mapping main (A) and interaction (B) effects when dominance effects were simulated (Scenarios 3 and 4).

Fig. S7: Comparison of QTL mapping methods to detect main effects in Scenarios 1 and 6.

Fig. S8: Comparison of QTL mapping methods to detect main effects in the rice heading date data.

Fig. S9: Comparison of QTL mapping methods to detect interaction effects in the rice heading date data.

Table S1: QTLs and QTL interactions detected with multiple interval mapping in the real rice data.

Abbreviations

AUC: Area under the curve

CIM: Composite interval mapping

IM: Interval mapping

LD: Linkage disequilibrium

LRT: Likelihood ratio test

MIM: Multiple interval mapping

QTL: Quantitative trait locus

ROC: Receiver operating characteristic

SHAP: Shapley additive explanations

SNP: Single nucleotide polymorphism

XGB: XGboost