Theor Appl Genet (2015) 128:41–53
DOI 10.1007/s00122-014-2411-y

ORIGINAL PAPER

Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.)

Akio Onogi · Osamu Ideta · Yuto Inoshita · Kaworu Ebana · Takuma Yoshioka · Masanori Yamasaki · Hiroyoshi Iwata

Received: 9 July 2014 / Accepted: 3 October 2014 / Published online: 24 October 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract

Key message **Our simulation results clarify the areas of applicability of nine prediction methods and suggest the factors that affect their accuracy at predicting empirical traits.**

Abstract Whole-genome prediction is used to predict genetic value from genome-wide markers. The choice of method is important for successful prediction. We compared nine methods using empirical data for eight phenological and morphological traits of Asian rice cultivars (*Oryza sativa* L.) and data simulated from real marker genotype data. The methods were genomic BLUP (GBLUP), reproducing kernel Hilbert spaces regression (RKHS), Lasso, elastic net, random forest (RForest), Bayesian lasso (Blasso), extended Bayesian lasso (EBlasso), weighted Bayesian shrinkage regression (wBSR), and the average of all methods (Ave). The objectives were to evaluate the predictive ability of these methods in a cultivar population, to characterize them by exploring the area of applicability of each method using simulation, and to investigate the causes of their different accuracies for empirical traits. GBLUP was the most accurate for one trait, RKHS and Ave for two, and RForest for three traits. In the simulation, Blasso, EBlasso, and Ave showed stable performance across the simulated scenarios, whereas the other methods, except wBSR, had specific areas of applicability; wBSR performed poorly in most scenarios. For each method, the accuracy ranking for the empirical traits was largely consistent with that in one of the simulated scenarios, suggesting that the simulation conditions reflected the factors that affected the method accuracy for the empirical results. This study will be useful for genomic prediction not only in Asian rice, but also in populations from other crops with relatively small training sets and strong linkage disequilibrium structures.

Communicated by Jose Crossa.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2411-y) contains supplementary material, which is available to authorized users.

A. Onogi · Y. Inoshita · H. Iwata (✉)
Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-Ku, Tokyo 113-8657, Japan
e-mail: aiwata@mail.ecc.u-tokyo.ac.jp

O. Ideta
Western Region Agricultural Research Center, National Agriculture and Food Research Organization, Fukuyama, Hiroshima 721-8514, Japan

K. Ebana
Genetic Resources Center, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan

T. Yoshioka · M. Yamasaki
Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University, Kasai, Hyogo 675-2103, Japan

Introduction

The recent development of molecular technologies has provided a new technique for improvement of quantitative traits in plant and animal breeding, called whole-genome prediction or genomic prediction. In this technique, genetic values of untested genotypes (lines or individuals) are predicted on the basis of genome-wide DNA markers such as single nucleotide polymorphisms (SNPs; Meuwissen et al. 2001). Selection based on whole-genome prediction, often called genomic selection, is drastically and rapidly

changing plant and animal breeding strategies (Hayes et al. 2009; Heffner et al. 2009; Jannink et al. 2010).

Because prediction accuracy directly influences genetic gain (Falconer 1981), the development of methods that make highly accurate predictions has been a major goal in studies that evaluate the potential of genomic selection. Methods include genomic BLUP (GBLUP) and its extension (VanRaden 2008; Aguilar et al. 2010; Christensen and Lund 2010); penalized regression methods such as ridge regression, Lasso, and elastic net (ENet) (Usai et al. 2009; Li and Sillanpaa 2012a; Ogutu et al. 2012); Bayesian regression methods such as BayesA and BayesB (Meuwissen et al. 2001; de los Campos et al. 2009; Hayashi and Iwata 2010; Habier et al. 2011); non-parametric regression methods to capture non-additive genetic effects (Gianola et al. 2006; Gianola and van Kaam 2008; Long et al. 2010; Ober et al. 2011); methods developed in the field of machine learning such as support vector machine and random forest (RForest) (Long et al. 2011a; Ogutu et al. 2011); and regression methods based on dimension reduction (Solberg et al. 2009; Long et al. 2011b). Ridge regression and its equivalent GBLUP, BayesA and BayesB, and Bayesian lasso (Blasso; Park and Casella 2008) are popular methods, and have been evaluated in many studies (reviewed in de los Campos et al. 2013). In contrast, such methods as Lasso (Tibshirani 1996), ENet (Zou and Hastie 2005), and RForest (Breiman 2001) have been evaluated only in a small number of studies (de los Campos et al. 2013), although these methods are commonly used in the field of pattern recognition and machine learning.

Comparison among prediction methods has revealed the factors that influence the relative performance of each method. For traits with fewer QTLs, the methods that have a variable selection feature, such as BayesB or Lasso, tend to outperform the methods that assume equal contributions of markers to genetic variance, such as ridge regression or GBLUP (Daetwyler et al. 2010, 2013). Because the performance of variable selection is affected by the magnitude of linkage disequilibrium (LD), the relative prediction accuracy of BayesB and ridge regression can differ among populations with different LD structures even when their genetic architecture is shared (Wimmer et al. 2013). Ridge regression and GBLUP tend to be inferior to BayesB when predicted individuals are genetically distant from the training set, because these methods depend on information on relatedness rather than LD between markers and QTLs (Habier et al. 2007; Zhong et al. 2009). Non-parametric regression such as RKHS tends to be superior to additive linear regression for non-additive traits (Long et al. 2010; Ober et al. 2011; Gonzalez-Camacho et al. 2012). These factors (genetic architecture, LD structure, and relationship between the training and validation sets) may differ among populations or traits, and complex interplay between them probably influences the relative performance of the prediction methods. Thus, empirical method evaluation for the data of interest is necessary to choose the appropriate methods.

Prediction methods have been compared empirically for crops; for example, wheat (Crossa et al. 2010; Heffner et al. 2011; Perez-Rodriguez et al. 2012), maize (Crossa et al. 2010; Albrecht et al. 2011; Riedelsheimer et al. 2012; Zhao et al. 2012; Crossa et al. 2013), and barley (Lorenz et al. 2012; Endelman et al. 2014). Lorenzana and Bernardo (2009) comprehensively compared multiple methods using populations of maize, barley, and Arabidopsis, whereas Heslot et al. (2012) used populations of maize, barley, wheat, and Arabidopsis; some of these populations were the same in both studies. These studies provided useful information on genomic prediction for traits and populations closely related to those tested. However, a drawback of empirical dataset comparison is that it is difficult to generalize the conclusions, because the genetic architecture of traits is generally ambiguous, and the causes of the differences in method accuracy are often unclear. For accurate characterization of prediction methods, empirical dataset comparison is insufficient.

This study compared whole-genome prediction methods using a population of Asian cultivated rice represented by both an empirical dataset and datasets simulated by using real marker genotype data. In simulations, the number of QTLs (N_{qtl}), the size of the training set (N_{train}), heritability, the presence of epistasis, and the extent of LD were considered as conditions. The prediction methods were GBLUP, RKHS, Lasso, ENet, RForest, Blasso, extended Bayesian Lasso (EBlasso; Mutshinda and Sillanpaa 2010), weighted Bayesian shrinkage regression (wBSR; Hayashi and Iwata 2010), which is equivalent to BayesA and BayesB, and the average of all the methods (Ave). EBlasso was evaluated for genomic prediction for the first time. The objectives were: (1) to evaluate the predictive ability of these methods for rice cultivars; (2) to characterize the methods by exploring the area of applicability of each in simulations; and (3) to investigate the causes of differences in method accuracy for empirical data using the simulation results. Because the simulation is based on real marker genotype data, the results will be informative for other traits not tested here. Moreover, they will be useful for prediction in other populations, particularly those that have small training sets and strong LD structures, similar to the population used in this study.

Materials and methods

Plant materials and phenotype evaluation

We used a dataset of 110 rice cultivars developed mainly in Japan (Supplementary Table 1). A comparative study

was conducted for eight traits: days to heading (DH), culm length (CL), panicle length (PL), panicle number (PN), grain length (GL), grain width (GW), brown rice length (BL), and brown rice width (BW). Phenotypes were evaluated at the National Agriculture and Food Research Organization, Western Region Agricultural Research Center, in Fukuyama, Hiroshima, Japan, for six consecutive years (2006–2011). DH was measured as the number of days from sowing to the time when more than half of inflorescences had emerged. CL was defined as the distance from the soil surface to the panicle node. PL was the distance from the panicle node to the head without the awns. PN was the number of normal panicles. GL, GW, BL, and BW were measured using a digital slide gauge. Phenotypic records were averaged over the 6 years. The average values were standardized when the prediction methods were trained.

Marker genotype data

DNA was extracted from one typical individual per cultivar using the CTAB method. Genotypes of 3,102 genome-wide bi-allelic markers were determined for the 110 cultivars. Among these, 3,071 were SNP markers developed from the genome sequence of Japanese cultivars (Yamamoto et al. 2010; Nagasaki et al. 2010), and the other 31 were SSR markers (Yamasaki and Ideta 2013). Genotyping was done on an Illumina BeadStation 500G genotyper (Illumina Inc., San Diego, CA, USA) according to the manual. Linkage and physical distances between adjacent markers were 1.0^{-6}–11.6 cM, and 0.079–2,504 kb, respectively, where 10 cM $\approx$ 2,500 kb. The averages (SDs) were 0.49 ($\pm$0.76) cM and 122.3 ($\pm$167.7) kb. The mean minor allele frequency was 0.309 ($\pm$0.124). LD between marker pairs within the same chromosome was measured as squared correlation (r^2) of marker genotypes coded as 0 or 2.

Genetic structure

Genetic structure of the 110 cultivars was investigated using hierarchical clustering. Clustering was based on Euclidian distances calculated from the marker genotypes using the R function hclust (R Development Core Team 2011).

Prediction methods

The prediction methods evaluated in this study are described in detail in Supplementary Methods. Briefly, we used the R packages, rrBLUP (ver. 4.2; Endelman 2011) for GBLUP and RKHS, glmnet (ver. 1.9-5; Friedman et al. 2010) for Lasso and ENet, and randomForest (ver. 4.6-7; Breiman 2001) for RForest. For Blasso, EBlasso, and wBSR, we used programs written in C that were based on variational Bayesian algorithms (Li and Sillanpaa 2012b; Hayashi and Iwata 2013). Ave is the average of all the eight methods compared.

Simulation analysis

To obtain further insights related to the differences in the performance of the prediction methods, we simulated the datasets on the basis of empirical rice genotype data. The conditions considered were N_{qtl}, heritability, N_{train}, epistasis, and the extent of LD. Throughout the simulations, QTLs were chosen from the markers and predictions were made on the basis of the markers that were not selected as QTLs. N_{qtl} was 6, 12, 36, or 120. When $N_{qtl} = 6$, QTLs were selected from different chromosomes. Otherwise, QTLs were selected randomly. On the assumption that the genetic variance explained by each QTL was equal, additive effects of QTLs were determined according to the allele frequencies of the selected markers. The signs of the QTL effects were determined randomly. The breeding value of each of the 110 cultivars was then calculated by summing up all the effects of the QTL alleles harbored by each cultivar. Phenotypic values were generated by adding random Gaussian noise to the breeding values. The noise variance was determined on the basis of the breeding value variance and narrow-sense heritability, which was assumed to be 0.1, 0.3, 0.5, 0.7, or 0.9. Because we used 11-fold cross-validation (CV) to evaluate the predictive ability as described in the next section, the sample size of 110 corresponds to $N_{train} = 100$.

We also generated datasets with 330 and 550 cultivars, which correspond to $N_{train} = 300$ and 500, respectively. To retain the LD structure observed in real marker genotype data, we used Cholesky decomposition of the correlation matrix among markers according to Wimmer et al. (2013). **C**, the correlation matrix of real marker data, was approximated by a positive definite matrix, **C***, by using the function nearPD of the R package MatrixR. By Cholesky decomposition of **C***, the upper diagonal matrix **U** was calculated so that $\mathbf{C}^* = \mathbf{U}'\mathbf{U}$. New genotypes **G** were generated as **WU**, where **W** is a binary matrix with the size of 330 (or 550) $\times$ 3,102 (i.e., the number of markers). The elements (-1 or 1) for the marker j in **W** were drawn randomly from a Bernoulli distribution with the parameter p_j, which was the allele frequency of the marker j in the original data. Because all cultivars were assumed to be inbred, the expected genotype frequency for each marker was equal to its allele frequency.

To simulate epistasis between two QTLs, $N_{qtl}/3$ pairs of non-overlapping QTLs were selected randomly. The epistatic effect of each pair was chosen randomly from the additive effects of the paired QTLs. The epistatic

components of the genetic value were generated as products of the epistatic effect and the genotypes of the paired QTLs. The total genetic value was calculated for each cultivar by summing up the additive and epistatic components. Phenotypic values were generated by adding random noise to the total genetic values; the broad-sense heritability was assumed to be 0.1, 0.3, 0.5, 0.7, or 0.9. Typically, 40 % of the total genetic variance was explained by epistatic variance. Note that, when we denote the conditions of simulation, we use the term "heritability" regardless of whether it is narrow-sense heritability, as in scenarios without epistasis, or broad-sense heritability, as in those with epistasis.

To investigate the influence of long-range LD on the predictive ability of the methods at $N_{qtl} = 6$, we also generated datasets in which the LD structure in the real marker genotype data was retained only around QTLs: genotypes of the markers located >10 cM from the QTLs were randomly permuted. On average, 325.3 (±87.9) markers were located within 10 cM of the QTLs and were retained without permutation.

In total, we generated 150 scenarios. Out of these, 120 resulted from the combinations of four N_{qtl} values (6, 12, 36, and 120), five heritability values (0.1, 0.3, 0.5, 0.7, and 0.9), three N_{train} values (100, 300, and 500), and two conditions (with or without epistasis). The remaining 30 scenarios had disturbed LD structures and resulted from combinations of one N_{qtl} (6) and different conditions of heritability, N_{train}, and epistasis. For each N_{qtl} value, 100 QTL sets were selected from the markers. The QTL positions and effects were shared among scenarios simulated under the same N_{qtl}. For each scenario, phenotypic values were generated once from each QTL set. Consequently, 100 replicates per scenario were tested.

Evaluation of prediction accuracy

Throughout this study, we performed 11-fold CV to evaluate the predictive ability. Because the sample sizes (numbers of cultivars) were 110, 330, and 550, the N_{train} values in CV were 100, 300, and 500, respectively. The same folds were used for each prediction method. We conducted CV 100 times for the empirical data and once for each replicate of the simulated data. For simulated data with $N_{train} = 300$ and 500, cultivars were randomly partitioned into each fold. When the empirical genotype data and simulated data with $N_{train} = 100$ were used, we partitioned cultivars into the folds in such a way that the original genetic structure (i.e., composition of genetic groups) was maintained in the training sets as much as possible. To make the comparison of prediction accuracy among studies that evaluate the methods meaningful, we measured the relationship between the training and validation sets using the mean squared relationship (rel^2) as suggested by Daetwyler et al. (2013). rel^2 was calculated from the realized relationship matrix generated by the A.mat function in the R package rrBLUP. For Lasso, ENet, Blasso, EBlasso, and wBSR, hyperparameters were determined using nested tenfold CV (Lasso and ENet) or fivefold CV (Blasso, EBlasso, and wBSR) in each cycle of 11-fold CV. The details are given in Supplementary Methods.

In empirical data analysis, prediction accuracy was measured using the Pearson correlation coefficient between the predicted and phenotypic values. In simulated data analysis, the correlation coefficient between the predicted and true genotypic values was used. The true genotypic values were obtained by summation of the additive and epistatic values when epistasis was simulated. The predicted values in CV were pooled across folds, and correlation coefficients were calculated for the pooled values. Accuracy among prediction methods was compared using Tukey's test. The R functions aov and TukeyHSD were used.

Coefficients of variation of prediction accuracy

To estimate the differences in accuracy among the prediction methods, we calculated the coefficients of variation of accuracy. Larger coefficients indicate greater relative differences in accuracy among the methods; in these cases, the choice of methods has a larger impact.

Searching simulation scenarios closest to the empirical traits

We searched the simulation scenarios with accuracy rankings among the prediction methods closest to those for the empirical traits. Spearman's correlation coefficient was used to measure the similarity of rankings. To compare narrow-sense heritability of the empirical traits with narrow- or broad-sense heritability in the scenarios closest to the empirical traits, we estimated the variance components (σ_u^2 and σ_e^2) for the empirical traits using GBLUP.

Results

Genetic and LD structures

The cultivar population used in this study consisted of two major genetic groups, which included 61 and 49 cultivars (Supplementary Fig. 1). In CV used for comparison of prediction methods applied to the empirical data, the average rel^2 between cultivars in the training and validation sets was 0.154 (±0.03).

As expected from the mode of reproduction and the breeding schemes of Japanese rice cultivars, LD extended over large distances: r^2 between marker pairs on the same

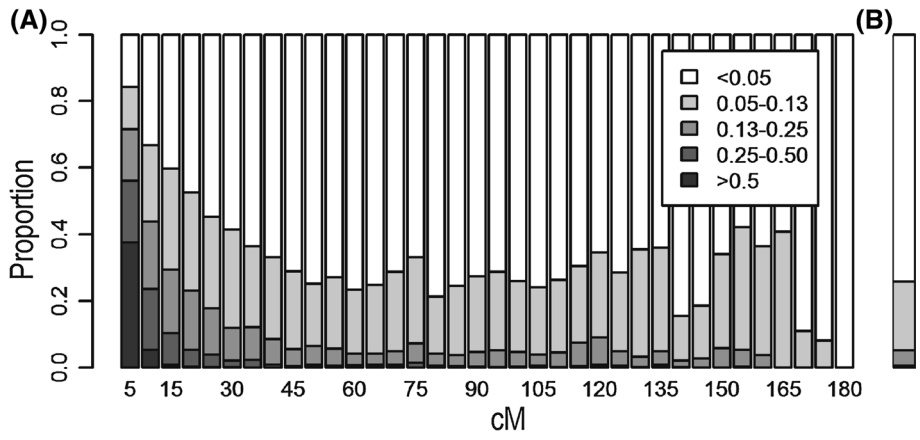

Fig. 1 Linkage disequilibrium (LD) measured as r^2. **a** Proportions of five fractions of r^2 values across the linkage distances between the marker pairs located on the same chromosomes. The proportions were calculated for each 5-cM window from 0 to 180 cM. **b** Proportions calculated for marker pairs located on different chromosomes (background LD)

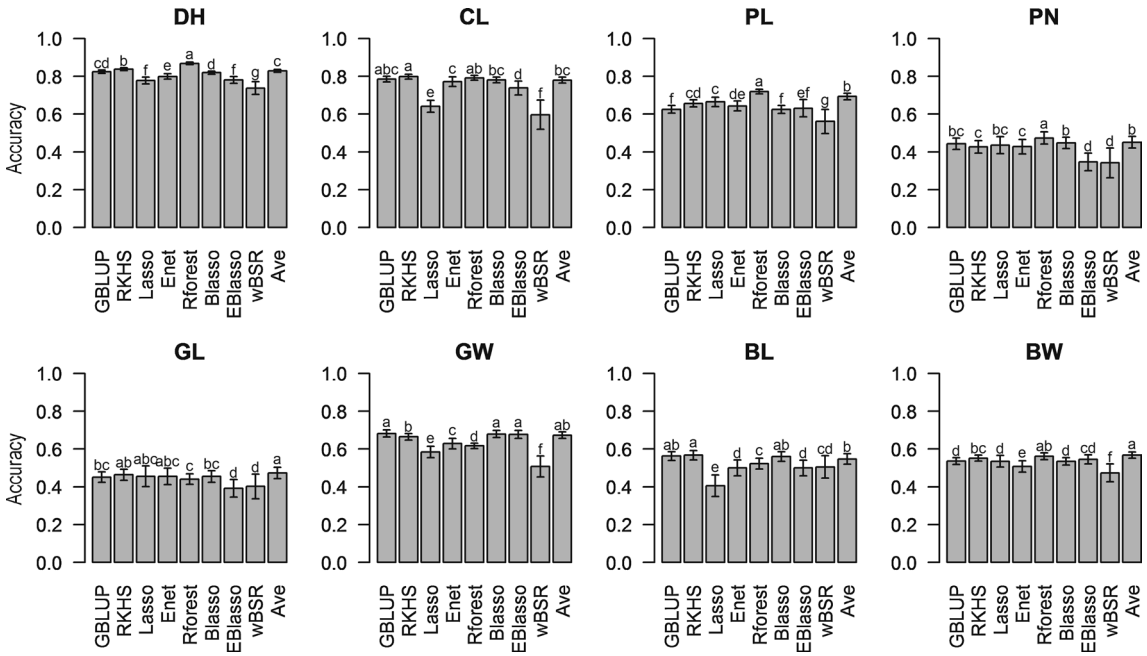

Fig. 2 Prediction accuracy of the nine methods for empirical traits. Accuracy was measured as the Pearson correlation coefficient between predicted and phenotypic values. *Different subscripts* indicate significant differences ($P < 0.05$). *DH* days to heading, *CL* culm length, *PL* panicle length, *PN* panicle number, *GL* grain length, *GW* grain width, *BL* brown rice length, *BW* brown rice width

chromosomes exceeded the background level at distances of ≤40 cM (Fig. 1). The mean values of r^2 calculated at 1-cM windows were >0.1 until 12 cM. The mean r^2 between adjacent markers was 0.52 (±0.37).

Comparison of prediction methods for empirical data

The phenotypic values of the eight traits are summarized in Supplementary Fig. 2. The accuracy of genomic prediction for these traits is shown in Fig. 2. In general, the accuracy was high for DH and CL and moderate for the other traits. The most accurate predictions were obtained with RForest for DH, PL, and PN; with RKHS for CL and BL; and with GBLUP for GW. The least accurate predictions were obtained with wBSR for DH, CL, PL, PN, GW, and BW; with Lasso for BL; and with EBlasso for GL. Ave generated the most accurate predictions for GL and BW. The coefficients of variation of prediction accuracy among the methods were 0.05 (DH), 0.10 (CL), 0.08 (PL), 0.13 (PN), 0.09 (GL), 0.10 (GW), 0.11 (BL), and 0.06 (BW). The metrics representing the behavior of Lasso (the number of markers fitted, i.e., markers with non-zero effects), ENet (hyperparameter α), and wBSR (hyperparameter π) are listed in Supplementary Table 2. The hyperparameters were chosen using nested CV (see Supplementary Methods for details).

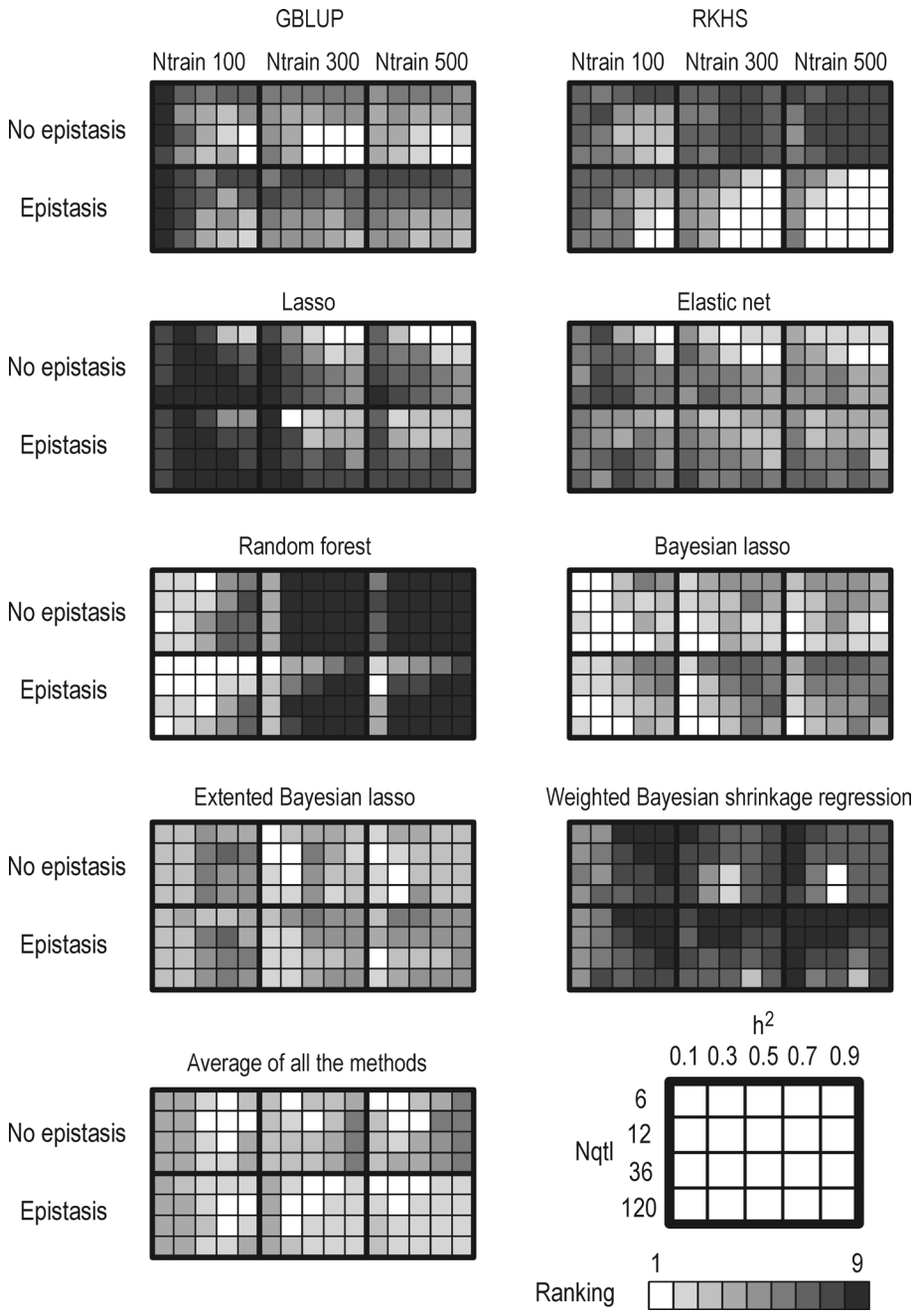

Fig. 3 Rankings of the prediction methods in the simulation analyses. Simulation scenarios (120 in total) varied in the number of QTLs (N_{qtl}, rows), heritability (h^2, columns), the size of the training set (N_{train}), and the presence or absence of epistasis. The *rankings* are shown as different shades from *white to black*

Comparison of prediction methods for simulated data

The prediction methods were compared using the datasets that were simulated from the empirical genotype data and had different genetic architectures. The simulated datasets (120 scenarios in total) varied in N_{qtl}, N_{train}, heritability, and the presence or absence of epistasis. The rankings of the prediction methods are summarized in Fig. 3. The accuracy for each scenario is presented in Supplementary Figs. 3, 4, 5. The metrics representing the behaviors of Lasso, ENet, and wBSR are described in Supplementary Figs. 6, 7, 8.

GBLUP tended to be ranked high for scenarios with high N_{qtl} and heritability values; this tendency became more noticeable in the absence of epistasis (Fig. 3). RKHS tended to outperform other methods when epistasis was simulated and the N_{qtl} and heritability values were high. Lasso and ENet tended to be ranked high when N_{qtl} was small and heritability was high. RForest performed well when $N_{train} = 100$. Notably, when $N_{train} = 100$, epistasis was simulated, and $N_{qtl} = 6$, RForest provided the most accurate prediction regardless of heritability. Blasso and EBlasso tended to be ranked higher when heritability was lower. wBSR performed poorly in most scenarios.

A notable feature of Blasso and EBlasso was their stability across scenarios. The average rankings for Blasso (3.4 ± 1.9) and EBlasso (3.7 ± 1.5) were better than those for GBLUP (5.2 ± 2.2), RKHS (5.1 ± 2.6), Lasso (6.5 ± 2.5), ENet (4.9 ± 1.9), RForest (6.1 ± 3.1), and wBSR (7.3 ± 1.7). Ave (2.8 ± 1.4) was also stable across scenarios.

The coefficient of variation of prediction accuracy decreased with increasing heritability or N_{train} (Table 1), suggesting that the choice of prediction methods becomes more important at low heritability or N_{train}. The coefficient of variation was not affected by N_{qtl} or epistasis (data not shown).

Table 1 Coefficients of variation of accuracy among the nine prediction methods in simulation scenarios where $N_{qtl} = 6$ and epistasis was absent

N_{train}	Heritability				
	0.1	0.3	0.5	0.7	0.9
100	3.02 (6.39)	0.85 (2.23)	0.20 (0.43)	0.11 (0.07)	0.09 (0.05)
300	0.54 (1.16)	0.05 (0.03)	0.04 (0.02)	0.05 (0.02)	0.06 (0.02)
500	0.20 (0.58)	0.04 (0.02)	0.04 (0.01)	0.04 (0.02)	0.05 (0.01)

Coefficients of variation were calculated for each replicate and averaged. SDs are shown in parentheses

Comparison of prediction methods for simulated data with disturbed long-range LD

To investigate the influence of long-range LD on the accuracy of the methods, the genotypes of the markers located >10 cM away from the simulated QTLs were randomly permuted. N_{qtl} was fixed at 6. The rankings of the methods are presented in Supplementary Fig. 9 and their accuracy is shown in Supplementary Fig. 10. Two major changes in the ranking were observed: (1) RKHS lost its superiority when epistasis was present, $N_{train} = 300$ or 500, and heritability was moderate or high; (2) EBlasso tended to become inferior to other methods when $N_{train} = 300$ or 500 and heritability was low. The metrics representing the behaviors of Lasso, ENet, and wBSR are described in Supplementary Figs. 6, 7, 8.

Scenarios closest to empirical traits

To investigate the causes of the differences in accuracy among the prediction methods in empirical analyses, we searched for the simulation scenarios closest to the empirical results for ranking of the prediction methods (Table 2; Fig. 4). Although moderate coefficients were observed across a broad range of scenarios for each trait (Fig. 4), the closest scenarios showed relatively high correlation coefficients (Table 2). For each trait, heritability in the closest scenarios was moderate or high (0.5–0.9). N_{qtl} was relatively small (6 or 12) except for GW and BL. Scenarios with epistasis were found for five traits. Narrow-sense heritability estimated via variance component analysis using GBLUP is also presented in Table 2.

Discussion

We compared nine whole-genome prediction methods for morphological and phenological traits of Asian rice cultivars, and for datasets simulated on the basis of real marker genotype data. GBLUP, RKHS, and RForest provided the best accuracy for one, two, and three, respectively, out of the eight traits. Ave improved the accuracy for two traits. The simulation results suggest that GBLUP, RKHS, RForest, Lasso, and ENet can be regarded as "specialist methods", whereas Blasso, EBlasso, and Ave showed stable performance across simulated scenarios and can be regarded as "generalist methods". Below, we discuss the properties of each method. Then we discuss the causes of the differences in prediction accuracy among the methods for real data. Finally, we provide recommendations on the choice of methods for Asian rice cultivars.

GBLUP

Because GBLUP assumes that every marker contributes equally to genetic variance, GBLUP was ranked higher with increasing N_{qtl}, as expected. The tendency for superiority of GBLUP or ridge regression in simulations with large N_{qtl} has been previously reported (e.g., Daetwyler et al. 2010; Jia and Jannink, 2012). The present study suggests that GBLUP loses its superiority when N_{train} is small and heritability is low, and is outperformed by Blasso, EBlasso, and RForest in such cases (Fig. 3). Estimation of parameters of GBLUP from data via REML is likely problematic when heritability is low and N_{train} is small, whereas prior distributions of Blasso and EBlasso probably compensate for reduced information from data. This interpretation implies that, under these conditions, GBLUP in a Bayesian framework (Legarra et al. 2008; Makowsky et al. 2011) or Bayesian ridge regression (Crossa et al. 2010; Perez-Rodriguez et al. 2012) can be more robust than REML-based GBLUP.

RKHS

The simulation results suggest that RKHS performs well in scenarios where heritability is high, N_{qtl} is large, and epistasis is present. The superiority of RKHS over additive regression methods in scenarios with epistasis is

Table 2 Simulation scenarios closest to the empirical traits in terms of the ranking of the prediction methods

Trait	Scenario			ρ (p value)[b]	$V_u/(V_u + V_e)$[c]
	N_{qtl}	h^{2a}	Epistasis		
DH	12	0.7	+	0.95 (3.5e − 4)	1.00
CL	12	0.7	+	0.82 (1.2e − 2)	1.00
PL	6	0.9	+	0.78 (1.7e − 2)	0.71
PN	6	0.5	−	0.78 (1.7e − 2)	0.51
GL	12	0.9	−	0.67 (5.9e − 2)	0.40
GW	36	0.9	−	0.93 (7.5e − 4)	0.82
BL	120	0.7	+	0.88 (3.1e − 3)	0.55
BW	12	0.9	+	0.85 (6.1e − 3)	0.52

[a] Broad-(narrow-) sense heritability when epistasis is (is not) simulated

[b] Spearman's rank correlation coefficient

[c] Estimated narrow-sense heritability. The additive genetic (V_u) and residual (V_e) variances were estimated using GBLUP. The realized genomic relationship matrix used in GBLUP was created using the A.mat function in the rrBLUP package (Endelman 2011)

consistent with the theory of this method and has been previously reported (e.g., Gonzalez-Camacho et al. 2012). The reason for the effect of heritability on the ranking of RKHS would be similar to that for GBLUP: the REML-based methods would be more sensitive to the reduction of information from data than Bayesian hierarchical methods. Bayesian inference of RKHS (Gonzalez-Recio et al. 2008; de los Campos et al. 2010) probably makes the method more robust, although it increases the computational time because of Markov chain Monte Carlo sampling. Because RKHS is based on a kernel matrix generated from whole-genome markers, it is intuitively understandable that RKHS is superior in scenarios where N_{qtl} is large. The results that RKHS lost its superiority when long-range LD was disturbed probably support this interpretation: when long-range LD is intact, the effects of QTLs will disperse across markers in a certain range, which would benefit RKHS. Conversely, the effects of QTLs will be limited to the markers adjacent to these QTLs when long-range LD was disturbed. RKHS would lose its superiority in this case.

Lasso

Lasso was ranked higher in scenarios with smaller N_{qtl}, higher heritability, and larger N_{train}. Lasso has the variable selection feature because of the L_1 penalty, whereas ridge regression shrinks all coefficients together toward zero because of the L_2 penalty (Hastie et al. 2009). Therefore, Lasso outperformed GBLUP and RKHS, the counterparts of ridge regression, as expected, in scenarios with smaller N_{qtl}. Similar results were reported in previous simulation studies (Usai et al. 2009; Ogutu et al. 2012). This superiority of Lasso was rapidly lost with increasing N_{qtl} and decreasing heritability and N_{train}. As shown by Donoho and Stodden (2006), the performance of Lasso in variable selection is sensitive to the ratio between the numbers of true non-zero variables (k), samples (n), and variables (p): at a given underdeterminedness level (n/p), variable selection fails to work once the sparsity level (k/n) is beyond a certain point (breakdown point). The breakdown point comes earlier with decreasing underdeterminedness level or increasing noise (residual variance). In genomic prediction, if we consider that the underdeterminedness level is usually much lower than one and QTL effects are dispersed across multiple markers because of LD (i.e., k is expected to be greater than N_{qtl}), Lasso expectedly has a narrow area of applicability.

When variables are highly correlated, Lasso performs poorly for variable selection (Buhlmann and van de Geer 2011). In fact, Wimmer et al. (2013) showed that variable selection of Lasso became harder with stronger LD structure. We disturbed long-range LD by permuting the genotypes of markers that were farther from the QTLs than 10 cM, and expected the accuracy of Lasso to increase.

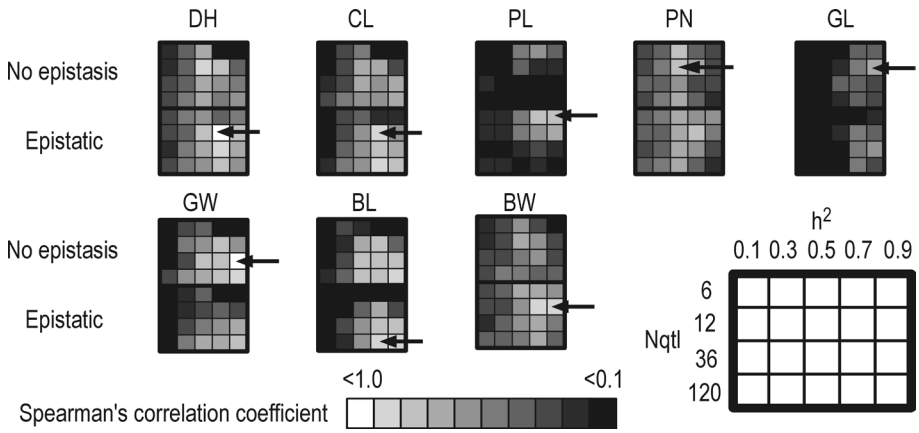

Fig. 4 Spearman's correlation coefficients between the rankings of the prediction methods for empirical traits and those in scenarios with $N_{train} = 100$. *Arrows* indicate the scenarios that showed the highest coefficients, i.e., scenarios closest to the empirical traits. *DH* days to heading, *CL* culm length, *PL* panicle length, *PN* panicle number, *GL* grain length, *GW* grain width, *BL* brown rice length, *BW* brown rice width

A slight decrease in accuracy (Supplementary Figs. 3, 4, 5, 10) indicates that, at least under this sparsity (>6/100, because k will be larger than N_{qtl}) and underdeterminedness (~100/3,000) levels, the magnitude of LD in this population does not greatly affect the predictive performance of Lasso.

ENet

ENet showed ranking tendencies similar to those of Lasso, probably because both methods share the variable selection feature. When epistasis was absent, ENet provided the most accurate prediction in scenarios where $N_{train} = 300$ or 500, heritability = 0.7 or 0.9, and $N_{qtl} = 12$. These conditions were similar to the optimal conditions for Lasso except that optimal N_{qtl} for Lasso was 6. This result suggests that the area of applicability of ENet is characterized by N_{qtl} slightly larger than that optimal for Lasso. Because of the L_1 and L_2 penalties, ENet has the variable selection feature as Lasso has, and also shrinks the coefficients of selected variables together as ridge regression does (Zou and Hastie 2005; Hastie et al. 2009). The advantages of ENet for genomic prediction in comparison with Lasso are that ENet can select correlated markers with nonzero effects as a group; and ENet can select more markers than N_{train}, whereas the maximum number of markers selected by Lasso is equal to N_{train} (Zou and Hastie 2005). Thus, more markers are selected by ENet than by Lasso, which is consistent with our observations (data not shown) and was reported by Li and Sillanpaa (2012a). This feature is probably involved in defining the area of applicability of ENet, which favors larger N_{qtl} than Lasso. However, in scenarios where $N_{qtl} = 36$ and 120, ENet tended to be inferior to GBLUP (ridge regression). Similar results have been reported by Wimmer et al. (2013): ENet tends to be intermediate between Lasso and ridge regression or inferior to both, but is rarely superior. Perhaps this result stems from the narrow area of applicability of ENet in genomic prediction.

RForest

The area of applicability of RForest included scenarios where heritability was low and N_{qtl} and N_{train} were small. The superiority of RForest in low-heritability scenarios may be attributable to bagging (Breiman 1996), which is known to work well with noisy data (Dietterich 2000). Regardless of heritability and N_{train}, the accuracy of RForest tended to decrease slightly with increasing N_{qtl}. We speculate that because large N_{qtl} and strong LD made the effects of many markers similar to each other, correlation among the residuals of trees was not decreased much by the random choice of markers at each node (for the principles of RForest, see Supplementary Methods). Alternatively, the reduced accuracy might result from the insufficient number of variables randomly selected at each node (mtry), which was set at one-third of the number of markers (i.e.,~1,000). As N_{train} increased, the ranking of RForest tended to decrease, although its accuracy increased. This indicates that other regression methods such as GBLUP or ENet benefited more from increased N_{train} than RForest did, possibly because RForest uses on average 63 % of the samples for the learning tree because of bootstrapping (Kohavi 1995).

RForest tended to be ranked higher in scenarios where epistasis was simulated. This feature was expected because of the recursive structure of the regression tree. Attempts to identify epistatic SNPs (QTLs) using RForest have been reported (Bureau et al. 2005; Jiang et al. 2009; Yao et al. 2013). The capability of RForest to capture epistasis is likely restricted by two factors: (1) at least one marker (QTL) among those involved in epistasis should have detectable additive effects so that it can be used for splitting; and (2) all the markers involved in epistasis should be selected as candidates at least once at different tandem nodes. The first factor has been pointed out by Breiman et al. (1984) in the context of determining the right size of trees. The second one emerges because RForest selects the mtry variables randomly as candidates at each node to reduce the correlation between residuals of trees (Breiman 2001). To detect a high-order interaction, mtry should be sufficiently large, but large mtry increases the correlation between trees and reduces the effectiveness of bagging (Hastie et al. 2009). In addition, the depths of the trees should be sufficiently large. In the present study, we simply simulated epistasis as an additive-by-additive interaction between two QTLs. If epistasis is simulated in a more complex manner, RForest might lose its superiority. When N_{train} was large (300 or 500), RForest (and RKHS in the low-heritability cases) was often inferior to additive regression methods (Blasso or ENet), although epistasis was simulated. The additive regression methods achieved higher accuracy than RForest via more accurate prediction for the additive components of genotypic effects (data not shown). Thus, if the proportion of epistatic variance increases more than that used in this study (typically 40 %), the relative rankings of RForest (and RKHS) and additive regression methods might differ.

Blasso

Blasso showed stable performance across scenarios. As shown by Park and Casella (2008), Blasso (and probably EBlasso) is a compromise between Lasso and ridge regression: as the regularization parameter λ_B^2 increases (for the explanation of λ_B^2, see Supplementary Methods), Blasso pulls small marker effects close to zero faster than ridge regression does; but Blasso does not compress them to zero

as Lasso does. This property, together with the prior information on parameters and a grid search of the hyperparameter of λ^2_B, probably allows stable performance of Blasso across genetic architectures. Although ENet can also be seen as a compromise between ridge regression and Lasso because it shares the variable selection feature with Lasso (Hastie et al. 2009), the tendencies in the rankings of ENet and Blasso differed considerably. With a benchmark dataset (which is oligogenic) from the XIIth QTLMAS workshop (Lund et al. 2009), Blasso was found to be less accurate than Lasso and ENet (Li and Sillanpaa 2012a) but more accurate than ridge regression (Usai et al. 2009). We also observed this tendency in our simulation scenarios when $N_{qtl} = 6$, $N_{train} = 300$ or 500, and heritability >0.3. Li and Sillanpaa (2012a) attributed the lower accuracy of Blasso to a stronger underestimation of regression coefficients, which probably results from a compromise between ridge regression and Lasso.

EBlasso

To overcome the above drawback of Blasso, Mutshinda and Sillanpaa (2010) proposed EBlasso, which has the marker-specific shrinkage factor η^2_j (for the explanation of η^2_j, see Supplementary Methods), and showed that EBlasso is more suitable for QTL mapping than Blasso. We evaluated the potential of EBlasso for genomic prediction. In simulation analyses, both Blasso and EBlasso showed stable performance across genetic architectures and N_{train} values. As expected, EBlasso ranked higher than Blasso when $N_{qtl} = 6$ and epistasis was absent. This suggests its superiority over Blasso in capturing the association signals of QTLs. However, when $N_{qtl} = 6$ and heritability was high, EBlasso was inferior to Lasso. This result suggests that EBlasso still tends to underestimate the regression coefficients in comparison with Lasso. Disturbing long-range LD led to a strong dependence of EBlasso performance on heritability when $N_{train} = 300$ and 500. This behavior resembled that of Lasso. However, the cause of this phenomenon is difficult to deduce. Further studies are required to understand this property of EBlasso in genomic prediction.

wBSR

We expected wBSR to perform better in scenarios with small N_{qtl}, but its performance was generally poor and no clear areas of applicability were found. This unexpected result probably stems from three causes. One is the low training set size. BayesA or BayesB, which are statistically equivalent to wBSR, outperform GBLUP and ridge regression for simulated oligogenic traits (e.g., Zhang et al. 2010; Clark et al. 2011; Sun et al. 2012; Daetwyler et al. 2013); BayesB may outperform GBLUP and ridge regression for relatively polygenic traits (Zhang et al. 2010; Sun et al. 2012). In the above studies, the training set sizes (≥1,000) were much larger than our largest N_{train}. The second cause is a strong LD structure. Similar to variable selection by Lasso, that by BayesB can be hampered by strong LD. Thus, the predictive ability of BayesB is also affected by strong LD (Wimmer et al. 2013). As shown in Supplementary Fig. 8, the π values chosen via CV were decreased by disturbing long-range LD, particularly when $N_{train} = 300$ or 500 and heritability >0.3. This indicates that a smaller number of markers had heavier weights when LD was disturbed, i.e., variable selection was stronger. However, the accuracy of wBSR was still inferior to that of Lasso or ENet, which suggests that long-range LD was not the only factor that affected wBSR performance. The third cause is the non-optimized hyperparameters, ν and S^2 (for the explanation of ν and S^2, see Supplementary Methods), which define the prior distribution of the marker effect variance. We tuned π via a grid search, but fixed ν at 4 and determined S^2 at each π value according to the assumption on the proportion of phenotypic variance that markers can explain. Because this proportion was unknown, we set it to 0.5. This choice might affect the performance. Moreover, the equation to determine S^2 that we adopted assumes linkage equilibrium (Habier et al. 2011). For populations with strong LD, this simplified assumption might not result in an optimal value of S^2. As pointed out by Gianola et al. (2009), the severe influence of the prior distribution of marker effects is unavoidable under the model structure of wBSR, where variance of the effect of each marker has to be learned on the basis of only one observation, i.e., the effect of the marker. This influence can be alleviated to some extent by estimating hyperparameters as unknown variables (Nadaf et al. 2012).

Ave

Ave is a naïve application of ensemble learning. A key for successful ensemble learning is the diversity of ensemble members. Krogh and Vedelsby (1995) showed that the ensemble generalization error can be reduced by increasing the variance of predictions among ensemble members, which is referred to as ensemble ambiguity. Ave applied to a simulated benchmark dataset of Hickey and Gorjanc (2012) did not improve the accuracy, especially for oligogenic traits (Daetwyler et al. 2013). A possible reason for the failure could be that the ensemble ambiguity was not sufficiently large, because the ensemble included similar methods such as linear regression with the variable selection feature (BayesB, BayesC, Lasso, and Bayesian stochastic search variable selection), and ridge regression and its equivalents (GBLUP and pedigree-based BLUP). Particularly for the additive oligogenic traits, the ensemble ambiguity will not be increased by adding such regression methods

as ensemble members, because most regression methods with the variable selection feature will provide similar predictions based on several major QTLs. Considering this, the averaging approach would be more effective for complex and non-additive traits. In our simulations, this seems to hold because Ave showed a slightly better performance when epistasis was present (Fig. 3). However, in a study by Heslot et al. (2012), the accuracy was improved only for 2 traits out of 18, although the superiority of RKHS for most traits suggested non-additive genetic effects controlled by a number of QTLs. Further studies are required to understand the properties of the averaging (combining) approach, in particular the appropriate member combinations.

Causes of the differences in prediction accuracy for empirical data

In comparative studies using empirical data, the causes of the differences in accuracy among methods are usually unclear. We attempted to infer the causes by searching the scenarios closest to the empirical results for ranking of the methods. This inference is challenging because inference of genetic architecture requires many more samples (cultivar numbers) than those used in the present study, as suggested by Agarwala et al. (2013). The results obtained were uncertain, as illustrated by a broad range of moderate correlation coefficients (Fig. 4). In addition, large discrepancies between the two kinds of estimates of heritability, those from the closest scenario and those from the variance component analysis, were observed, particularly for GL and BW (Table 2). These discrepancies might stem from too simplified simulation schemes, unreliable estimation of the variance components because of the small sample size, or both, in addition to the difference in the definition of heritability, i.e., narrow- or broad-sense. Nevertheless, for most traits, the rankings of the methods for the empirical traits did not deviate considerably from those for the closest scenarios (Table 2), suggesting that these scenarios may include factors causing the differences in accuracy in empirical analyses.

Two discrepancies were observed in this experiment. First, the observed accuracies for DH and CL ranged from 0.60 (wBSR for CL) to 0.87 (RForest for DH), whereas those for the closest scenarios ranged from 0.43 (wBSR) to 0.51 (RForest), although the accuracy was measured in simulations as the correlation between the predicted and *true* genotypic values. This suggests the factors that influence the accuracy without considerably affecting the rankings of the methods. Second, the difference in N_{qtl} of the closest scenarios between GL (12) and BL (120) seems too great, although the two traits are unlikely to have considerably different genetic architectures. The phenotypic correlation was high (0.87). We speculate that some non-genetic factors, such as measurement error, might affect the phenotypic values for GL because the prediction accuracy for GL was generally lower than for BL, the Spearman's correlation coefficient was the lowest, and narrow-sense heritability estimated from the variance components was also the lowest (Table 2).

Choice of the prediction methods in Asian rice breeding

When epistasis is expected to influence the traits, we recommend RKHS or RForest, which have complementary areas of applicability. When epistasis is absent, we recommend GBLUP or Blasso. The situations where Lasso or ENet are useful will be limited, unless the size of the training set is very large. Although Blasso is appealing in terms of its robustness, averaging predictions of several methods is superior to Blasso in this sense. The choice of the prediction method becomes less important as N_{train} or heritability increases, whereas the method should be chosen carefully when N_{train} is small or heritability is low. RKHS (or kernel regression) and ensemble learning, including methods based on bagging, such as RForest, have a good potential for further improvement.

Author contributions AO designed this study, performed all statistical analyses, and drafted the manuscript. YI collaborated on preliminary statistical analyses. KE, OI, TY, and MY prepared plant materials and performed phenotypic evaluation. KE, TY, and MY collected marker genotype data. MY, KE, and HI drafted the manuscript. HI conceived and supervised this study. All authors have read and approved the final manuscript.

Acknowledgments This work was supported by a Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows (26.10661), and also by the FIRST Program of JSPS. We would like to thank the technical staff at the National Agriculture and Food Research Organization Western Region Agricultural Research Center.

Conflict of interests The authors declare that they have no competing interests.

References

Agarwala V, Flannick J, Sunyaev S, GoT2D Consortium, Altshuler D (2013) Evaluating empirical bounds on complex disease genetic architecture. Nature Genet 45:1418–1429

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 93:743–752

Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, Ouzunova M, Simianer H, Schon CC (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet 123:339–350

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
Breiman L (2001) Random forests. Mach Learn 45:5–32
Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC Press, Boca Raton
Buhlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin
Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdeweg P (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28:171–182
Christensen OF, Lund MS (2010) Genomic prediction when some animals are not genotyped. Genet Sel Evol 42:2
Clark SA, Hickey JM, van der Werf JHJ (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18
Crossa J, Campos GL, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724
Crossa J, Beyene Y, Kassa S, Perez P, Hickey JM, Chen C, de loss Campos G, Burgueno J, Windhausen VS, Buckler E, Jannink JL, Lopez CMA, Babu R (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 (Bethesda) 3:1903–1926
Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031
Daetwyler HD, Calus MP, Pong-Wong R, de los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–365
de los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb) 92:295–308
de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345
de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385
Dietterich TG (2000) Ensemble methods in machine learning. In: Kittler J, Roli F (eds) Multiple classifier systems. Springer, Berlin, pp 1–15
Donoho D, Stodden V (2006) Breakdown point of model selection when the number of variables exceeds the number of observations. In: Proceedings of the international joint conference on neural networks, pp 1916–1921
Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255
Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, Jannink JL (2014) Optimal design of preliminary yield trials with genome-wide markers. Crop Sci 54:48–59
Falconer DS (1981) Introduction to quantitative genetics, 2nd edn. Longman Inc., New York
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22
Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363
Gianola D, van Kaam JB (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289–2303
Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776
Gonzalez-Camacho JM, de los Campos G, Perez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–771
Gonzalez-Recio O, Gianola D, Long N, Weigel KA, Rosa GJ, Avendano S (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics 178:2305–2313
Habier D, Fernando RL, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397
Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinfo 12:186
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
Hayashi T, Iwata H (2010) EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet 11:3
Hayashi T, Iwata H (2013) A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinfo 14:34
Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443
Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12
Heffner EL, Jannink JL, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65–75
Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160
Hickey JM, Gorjanc G (2012) Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3 (Bethesda) 2:425–427
Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomic Proteomic 9:166–177
Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192:1513–1522
Jiang R, Tang W, Wu X, Fu W (2009) A random forest approach to the detection of epistatic interactions in case-control studies. BMC Bioinfo 10(Suppl 1):S65
Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the international joint conference on artificial intelligence (IJCAI) vol 14, pp 1137–1145
Krogh A, Vedelsby J (1995) Neural network ensembles, cross validation, and active learning. In: Tesauro G, Touretzky DS, Leen TK (eds) Advances in neural information processing systems 7. MIT Press, Cambridge, pp 231–238
Legarra A, Robert-Granie C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. Genetics 180:611–618
Li Z, Sillanpaa MJ (2012a) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theor Appl Genet 125:419–435
Li Z, Sillanpaa MJ (2012b) Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190:231–249
Long N, Gianola D, Rosa GJ, Weigel KA, Kranis A, Gonzalez-Recio O (2010) Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetic Res (Camb) 92:209–225
Long N, Gianola D, Rosa GJ, Weigel KA (2011a) Application of support vector regression to genome-assisted prediction of quantitative traits. Theor Appl Genet 123:1065–1074

Long N, Gianola D, Rosa GJ, Weigel KA (2011b) Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. J Anim Breed Genet 128:247–257

Lorenz AJ, Smith KP, Jannink JL (2012) Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Sci 52:1609–1621

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161

Lund MS, Sahana G, de Koning DJ, Su G, Carlborg O (2009) Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC Proc 3 (Suppl 1):S1

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet 7:e1002051

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

Mutshinda CM, Sillanpaa MJ (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. Genetics 186:1067–1075

Nadaf J, Riggio V, Yu TP, Pong-Wong R (2012) Effect of the prior distribution of SNP effects on the estimation of total breeding value. BMC Proc 6(Suppl 2):S6

Nagasaki H, Ebana K, Shibaya T, Yonemaru JI, Yano M (2010) Core single-nucleotide polymorphisms—a tool for genetic analysis of the Japanese rice population. Breed Sci 60:648–655

Ober U, Erbe M, Long N, Porcu E, Schlather M, Simianer H (2011) Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 188:695–708

Ogutu JO, Piepho HP, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc 5(Suppl 3):S11

Ogutu JO, Schulz-Streeck T, Piepho HP (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proc 6(Suppl 2):S10

Park T, Casella G (2008) The Bayesian lasso. J Am Stat Assoc 103:681–686

R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. ISBN 3-900051-07-0. http://www.R-project.org/

Perez-Rodriguez P, Gianola D, Gonzalez-Camacho JM, Crossa J, Manes Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2:1595–1605

Riedelsheimer C, Technow F, Melchinger AE (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13:452

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH (2009) Reducing dimensionality for prediction of genome-wide breeding values. Genet Sel Evol 41:29

Sun X, Qu L, Garrick DJ, Dekkers JC, Fernando RL (2012) A fast EM algorithm for BayesA-like prediction of genomic breeding values. PLoS One 7:e49157

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58:267–288

Usai MG, Goddard ME, Hayes BJ (2009) LASSO with cross-validation for genomic selection. Genetic Res (Camb) 91:427–436

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, Schon CC (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195:573–587

Yamamoto T, Nagasaki H, Yonemaru J, Ebana K, Nakajima M, Shibaya T, Yano M (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics 11:267

Yamasaki M, Ideta O (2013) Population structure in Japanese rice population. Breed Sci 63:49–57

Yao C, Spurlock DM, Armentano LE, Page CDJ, Vandehaar MJ, Bickhart DM, Weigel KA (2013) Random forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. J Dairy Sci 96:6716–6729

Zhang Z, Liu J, Ding X, Bijma P, de Koning DJ, Zhang Q (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5:e12648

Zhao Y, Gowda M, Liu W, Wurschum T, Maurer HP, Longin FH, Ranc N, Reif JC (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet 124:769–776

Zhong S, Dekkers JC, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B 67:301–320