Genome-wide association study and genomic prediction for yield and grain quality traits of hybrid rice

Peiyi Yu · Changrong Ye · Le Li · Hexing Yin · Jian Zhao · Yongka Wang · Zhe Zhang · Weiguo Li · Yu Long · Xueyi Hu · Jinhua Xiao · Gaofeng Jia · Bingchuan Tian

Received: 14 February 2022 / Accepted: 8 March 2022 / Published online: 18 March 2022 © The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract Genomic selection is an efficient tool for breeding selection, especially for quantitative traits controlled by multiple genes with low heritability. To validate the application of genomic selection in hybrid rice breeding, the yield and grain quality traits of 404 hybrid rice breeding lines were investigated, and the same accessions were genotyped by using a 56 K SNP chip. There were wide variances among the tested accessions for all the measured traits, and most of the traits were correlated. A total of 67 significant loci were identified for the yield-related traits, and 123 significant loci were identified for the grain quality traits by GWAS. Two of these loci associated with increasing grain yield but decreasing grain quality. The GEBVs of all the yield and grain quality traits were calculated by using 15 different prediction algorithms. The plant height, panicle length, thousand grain weight, grain length and width ratio, amyllose content, and alkali value have higher predictability than other traits. However, the predictive accuracy of different GS models is different for different traits. This study provided useful information for genomic selection of specific trait using proper markers and prediction models.

Keywords Genomic selection · Molecular breeding · Yield · Grain quality · Hybrid rice

Introduction

The use of heterosis in hybrid rice has become increasingly important since the beginning of hybrid rice extension in China (Ma and Yuan 2015). Hybrid rice has contributed greatly to food security in China and the world. In recent years, the average yield of rice in China has increased from 3.5 ton/ha in 1975 to 6 ton/ha in 1995, and to 7 ton/ha in 2018 (FAOSTAT), and the grain quality of rice has been improved. About 50% of newly registered rice varieties in China (national approval) have grain quality of grade one (Lu et al. 2019). However, it is time- and labor-consuming for developing a hybrid variety by conventional breeding even though marker-assisted selection (MAS) has been used. Future breeding of hybrid rice will benefit from the use of new breeding
technology integrated with genetics, genomics, computational science, and artificial intelligence.

Rice is a model species for genomic study of monocotyledonous plant. The genome of rice was fully sequenced in 2005 (International Rice Genome Sequencing Project and Sasaki 2005), and more than 3000 genes have been cloned and analyzed (Yao et al. 2018). Large number of molecular markers have been developed for MAS of important traits such as plant height, blast resistance, leaf blight resistance, submergence tolerance, and fragrance (Jena and Mackill 2008). However, the success of MAS heavily depended on level of heritability and genetic architectures of the selected traits. MAS is not effective for traits controlled by large number of genes/QTLs with small contribution. With the development of high-throughput sequencing and chip technology, genome-wide association study (GWAS) has been used for identification of useful genes/QTLs of complex traits (Li et al. 2018). A total of 130 associated loci were identified for 38 agronomic traits of 1,495 elite hybrid rice varieties and their inbred parental lines (Huang et al. 2015). QTLs related to heterosis and yield traits of hybrid rice were also identified (Zhen et al. 2017; Chen et al. 2019; Su et al. 2021). However, rice yield is mainly controlled by minor effect loci and hardly to be selected in breeding process. Thus, genomic selection (GS) or genome-wide selection (GWS) has been proposed as a promising tool and applied for animal and plant genetic improvement (Meuwissen et al. 2001). GS has higher genetic gain than marker assisted selection for complex traits controlled by large number of QTLs (Crossa et al. 2017b). However, GS has not been successfully used in hybrid rice breeding yet.

Genomic selection uses genotypes and phenotypes of target traits from individuals in a training population to establish prediction models, and uses the models to predict genomic estimated breeding values (GEBVs) of individuals based on their genotypes in a test population (Crossa et al. 2017a). The hypothesis is based on the assumption that with high-density SNP markers distributed throughout the whole genome, at least one SNP can be found in a linkage disequilibrium state with the quantitative genetic loci affecting the target trait, so that the effect of each QTL can be reflected by SNP markers (Meuwissen 2007). The statistical models of genome selection can be roughly divided into two categories. The first is the direct method, which takes the individual as the random effect, the genetic relationship matrix constructed by the genetic information of the reference population and the predicted population as the variance covariance matrix, to estimate the variance components through the iterative method, and obtain the predicted breeding value of the individual. The second is the indirect method, which first estimates the marker effect in the reference group, and then accumulates the marker effect combined with the genotype information from the prediction group to obtain the individual estimated breeding value of the prediction group (Zhang et al. 2011; Misztal and Legarra 2017). Different prediction modes use different statistical methods; thus, the efficiency of the models needs to be compared and validated before using for breeding selection.

Genomic selection has been successfully used in animal breeding programs to increase the rate of genetic gain of dairy cattle, pig, dairy goat, layer chicken, and fish (García-Ruiz et al. 2016; Samore and Fontanesi 2015; Mucha et al. 2015; Wolc et al. 2015; López et al. 2015). In recent year, simulations and experimental studies have been conducted to validate the efficiency of this method in plant breeding. Being specific to rice, the predictive accuracy of heading date, culm length, panicle length, panicle number, grain length, and grain width varied from 0.4 to 0.8 in a population of 110 rice cultivars using nine prediction methods (Onogi et al. 2015). The highest predictive abilities for spikelets per panicle, heading date, plant height, and protein content were 0.44–0.7 in a diverse population of 413 rice inbred lines from 82 countries genotyped with a 44 K SNP chip (Isidro et al. 2015). The GEBVs of other traits such as grain shape, grain yield, nitrogen balance index, panicle weight, grain weight, and blast resistance have been predicted using inbred lines or cultivars (Spindel et al. 2015; Yabe et al. 2018; Iwata et al. 2015; Grenier et al. 2015; Hassen et al. 2018; Huang et al. 2019). Genomic prediction has also been conducted for grain yield, thousand grain weight, and index of different traits of hybrid rice (Wang et al. 2017, 2018; Xu et al. 2014, 2018; Cui et al. 2020). The predicted GEBVs from different populations were similar; thus, genomic selection is a reliable method for rice breeding.

In this study, we investigated the yield-related traits and grain quality-related traits of 404 hybrid
Materials and methods

Plant materials

A total of 404 hybrid rice accessions were planted in the field in Changsha (N28.31, E113.31, A80m). The number of accessions investigated each year was 37, 46, 50, 102, 72, and 97 from 2014 to 2019. Hybrid rice varieties Fengliangyou 4 (FLY4) and Fyou498 (FY498) were used as common check variety and planted with the tested hybrids every year.

Field experiments and phenotyping data collection

A randomized complete block design (RCBD) with three replications was used for field experiments. For each hybrid, 250 seedlings were transplanted into a 13.3 m² plot (5 m × 2.66 m) with a density of 0.2 m × 0.266 M. At maturity, growth period (GP, days from seeding to harvest), number of tillers (TN), plant height (PH, cm), panicle length (PL, cm), number of grains per panicle (GN), spikelet fertility (SF, %), and thousand grain weight (TGW, g) were measured. Panicles from 1 m² area were harvested and dried for calculating the grain yield (YLD, kg/ha).

The grain quality was evaluated following the industrial standard for quality of rice variety (NY/T 593–2013, Ministry of Agriculture, China). The evaluated traits include BRR (brown rice rate, %), WRR (white rice rate, %), WWRR (whole white rice rate or head rice rate, %), GL (grain length, mm), GLWR (grain length/width ratio), CP (chalk percentage, %), CD (chalk degree), AC (amylose content, %), GC (gel consistency, mm), ALK (alkali value), TRANS (transparency), and OG (overall grade of grain quality, grade 1 is the best, grade 3 is the worst). The phenotypic data of all the traits were collected each year for all the accessions.

GWAS and GS analysis

Basic statistical analysis and Pearson correlations among traits were calculated by using Minitab 17 (Minitab LLC). GWAS was conducted by using TASSEL 5.0 (Bradbury et al. 2007). Sites with more than 10% missing data and minimum frequency less than 0.05 were removed, and the filtered dataset containing 34,832 high-quality SNPs was used for the following analysis. Cladogram was created by using neighbor joining algorithm in TASSEL, and phylogenetic tree was showed in archaeopteryx tree viewer. The kinship matrix with scaled IBS was generated using genotyping data. A united data file with genotyping and phenotyping data (mean values of each accession) of the hybrids was created by using union join. The united file along with kinship matrix was analyzed for marker-trait association using mixed linear model (MLM). The compression level was set to optimum level, and variance component estimation was set to P3D. A criteria for claiming a QTL was p < 1 × 10⁻⁴ (-log10 p-value > 4.0). The identified QTLs were named using the CGSNL nomenclature (McCouch and Committee on Gene Symbolization 2008).

Genomic selection models were built by using big scale ridge regression (bigRR), best line unbiased prediction (GBLUP), least absolute shrinkage and selection operation (LASSO), ridge regression
BLUP (rrBLUP), sparse partial least square regression GBLUP(SPLS), reproducing Kernel Hilbert Space (RKHS), random forest classifier (RFC), random forest regression (RFR), support vector regression (SVR), support vector linear classifier (SVC), and Bayesian regularized neural network (BRNN) in G2P (https://github.com/cma2015/G2P), an integrated genomic selection package for predicting phenotypes from genotypes in R. The main feature of each model was listed in Table S1. For example, the grain yield from multi-year and multi-site was calculated by using mixed linear model of IME4 program in R (Bates et al. 2015): 

\[ y_{ij} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \epsilon_{ij} \]

- \( y_{ij} \) is the yield of \( i \)th variety in \( j \)th environment.
- \( \mu \) is the overall average yield.
- \( \alpha_i \) is the varietal effect of \( i \)th variety.
- \( \beta_j \) is the environmental effect of \( j \)th environment.
- \( (\alpha \beta)_{ij} \) is the interaction effect of \( i \)th variety and \( j \)th environment.
- \( \epsilon_{ij} \) is the residual error.

Variety is fixed effect, while environment and variety and environment interaction are random effects.

GEBVs of different traits were calculated by using SOMMER for GBLUP model and G2P for other models in R program (Covarrubias-Pazaran 2016).

The predictive accuracy of the models was validated by using 5×cross validation method; all data were randomly divided into 5 subsets, with one subset being used as the validation set, and the other 4 subsets being used as the training set, until the complete prediction of all data. The calculation was done once for each subset of the data, and the average predictive accuracy was used as the predictive accuracy of the model.

To evaluate the effect of significant loci from GWAS analysis, the predictive accuracy of a model for each trait was compared between all SNP markers used and significant markers from GWAS excluded.

**Results**

Statistics of the phenotypic data

Eight yield-related traits and 12 grain quality-related traits were investigated. There was significant variation for each trait. Most of the traits showed normal distribution, except TN, GLWR, and CD with higher kurtosis values than other traits (Table 1).

Based on Pearson correlation, spikelet fertility was not correlated with number of tillers (TN) and thousand grain weight (TGW). Other yield-related traits were correlated (Table 2). For the grain quality-related traits, whole white rice rate (WWRR), grain length (GL), chalk degree (CD), and overall grade (OG) were not correlated with amylose content (AC) and gel consistency (GC). And whole white rice rate (WWRR), gel consistency (GC), alkali value (ALK), and transparency (TRANS) were not correlated with grain length/width ratio (GLWR), while most of other grain quality traits were correlated (Table 3).

Diversity and classification of the hybrids tested

All 404 hybrids were genotyped by using a 56 K SNP chip which includes 56,897 SNPs. After filtering, 34,832 high-quality SNPs remained. There are 1992–4736 SNP markers on each chromosome (Fig. 1). Phylogenetic tree from these high-quality SNPs showed that all the hybrids were indica rice, with the exception of three hybrids which have larger genetic distance from others, possibly due to introgression from japonica rice (Figure S1).

GWAS of yield-related traits and grain quality traits

A total of 67 significant loci were identified for the yield-related traits, and clusters of loci for different traits were identified on chromosomes 1, 3, 4, 5, 6, 9, 11, and 12 (Fig. 2, Table S2). QTLs for grain yield were identified on chromosomes 3, 4, 5, 6, 9, 10, and 12, and most of them were collocated with QTLs for plant height (PH), panicle length (PL), number of grains per panicle (GN), spikelet fertility (SF), and thousand grain weight (TGW).

A total of 123 significant loci were identified for the grain quality traits, and clusters of loci for different traits were identified on chromosomes 1, 2, 5, 6, 7, 9, 11, and 12 (Fig. 3, Table S3). QTLs for overall grade of grain quality were identified on chromosomes 2, 5, 6, and 12, and most of them were collocated with QTLs for alkali value (ALK), chalk percentage (CP), chalk degree (CD), amylose content (AC), gel consistency (GC), and transparency (TRANS).
We found that two SNP markers on chromosome 5 (AX-155748928) and chromosome 12 (AX-154698806) were significantly associated with different yield and grain quality traits (Table 4). There are few accessions with GG genotype for both markers AX-155748928 and AX-154698806, and the means of phenotypic traits were not different from AG genotype. When compared with the homozygote AA genotype, the heterozygotes (AG) of both markers AX-155748928 and AX-154698806 were higher in plant height, number of grains per panicle, yield, chalkiness, transparency, and overall grade (low quality), but lower in white rice rate and alkali value (Figure S2).

Genomic selection models for yield and grain quality traits

The GEBVs of all yield and grain quality traits were calculated by using 15 different prediction algorithms, and 5× cross validation was used to evaluate the prediction accuracy. For the same trait, the prediction abilities of different GS models were different. Also, for the same GS prediction model, the prediction abilities varied for different traits (Fig. 4). The plant height, panicle length, thousand grain weight, grain length, grain length and width ratio, chalk percentage, amylase content, alkali value, and gel consistency had higher predictive accuracy than other

### Table 1 Statistics of the yield and grain quality-related traits

| Traits       | Count | Mean | StDev | Minimum | Maximum | Skewness | Kurtosis |
|--------------|-------|------|-------|---------|---------|----------|----------|
| GP (days)    | 298   | 134.0| 13.5  | 112.8   | 157.7   | 0.1      | -1.3     |
| TN           | 298   | 16.4 | 2.1   | 11.5    | 31.8    | 1.5      | 9.5      |
| PH (cm)      | 393   | 114.4| 9.5   | 92.7    | 135.7   | 0.1      | -0.9     |
| PL (cm)      | 393   | 24.3 | 1.6   | 19.4    | 29.3    | 0.1      | -0.1     |
| GN           | 393   | 187.2| 23.5  | 129.7   | 258.9   | -0.2     | -0.1     |
| SF (%)       | 393   | 82.9 | 3.7   | 65.4    | 90.2    | -0.7     | 0.6      |
| TGW (g)      | 393   | 25.2 | 2.4   | 17.5    | 32.8    | 0.5      | 0.0      |
| YLD (kg/ha)  | 393   | 9185.8|1163.6|5910.0   |11,256.0|1.0      |0.3       |
| BRR (%)      | 392   | 79.3 | 1.7   | 72.3    | 83.1    | -0.9     | 1.3      |
| WRR (%)      | 392   | 70.1 | 1.8   | 64.5    | 75.6    | -0.4     | -0.2     |
| WWRR (%)     | 392   | 58.6 | 5.9   | 37.1    | 70.3    | -0.8     | 0.7      |
| GL (%)       | 392   | 6.6  | 0.3   | 5.7     | 7.7     | 0.3      | 0.5      |
| GLWR         | 392   | 3.2  | 0.2   | 2.3     | 4.2     | 0.9      | 3.1      |
| CP (%)       | 392   | 23.4 | 10.7  | 5.0     | 70.0    | 0.9      | 1.5      |
| CD           | 392   | 6.2  | 3.5   | 0.7     | 23.4    | 1.4      | 3.4      |
| AC (%)       | 392   | 15.6 | 2.5   | 12.0    | 24.9    | 1.3      | 1.0      |
| GC           | 392   | 67.9 | 14.6  | 30.0    | 90.0    | -0.9     | 0.1      |
| ALK          | 392   | 5.3  | 1.2   | 3.0     | 7.0     | -0.2     | -1.1     |
| TRANS        | 392   | 1.3  | 0.5   | 1.0     | 3.0     | 1.1      | 0.2      |
| OG           | 392   | 4.3  | 1.1   | 2.0     | 5.0     | -1.0     | -0.8     |

### Table 2 Pearson correlation among yield-related traits. Upper number is Pearson correlation, lower number is p value

*Significance at p < 0.05, **significance at p < 0.01, ***significance at p < 0.0005

|       | GP     | TN     | PH     | PL     | GN     | SF     | TGW    |
|-------|--------|--------|--------|--------|--------|--------|--------|
| TN    | -0.456***|        |        |        |        |        |        |
| PH    | 0.339***| -0.513***|        |        |        |        |        |
| PL    | 0.417***| -0.547***| 0.662***|        |        |        |        |
| GN    | 0.703***| -0.532***| 0.545***| 0.392***|        |        |        |
| SF    | 0.317***| -0.090 | 0.289***| 0.119* | 0.269***|        |        |
| TGW   | 0.337***| -0.550***| 0.324***| 0.458***| 0.135* | 0.007  |
| YLD   | 0.676***| -0.218***| 0.407***| 0.291***| 0.660***| 0.602***| 0.291***|
traits (Table 5). Thousand grain weight could be well predicted by all the models, while the transparency of the grain had very low predictive accuracy. The predictabilities for grain length and width ratio, chalk percentage, amylose content, alkali value, and gel consistency significantly varied among models. The predictive accuracy for grain yield ranged from 0.18 to 0.35 (average 0.31) (Table S4).

By comparing the GS models without or with the significant SNP markers from the GWAS analysis, most of the models had higher predictive accuracy when the significant SNP markers from GWAS were considered (Fig. 5, Table S5).

Discussions

QTL and genes affecting rice yield and grain quality

It is generally accepted that rice grain yield and quality are two negatively related traits. The high yield varieties usually have low grain quality. Rice breeders have been trying to balance these traits in the breeding process (Xiao et al. 2021). However, the genetic linkage between grain yield and grain quality have not been dissected. In this study, a total of 67 QTLs for grain yield and 123 QTLs for grain quality were identified by comprehensive evaluation of related traits. Among these loci, we found two SNP markers that were significantly associated with various yield and grain quality traits. The heterozygotes (AG) of markers AX-155748928 on chromosome 5 (chr5:5,417,532) and AX-154698806 on chromosome 12 (chr12:13,961,623) had high yield but low grain quality. The SNP marker AX-155748928 was located in the exon region of gene LOC_Os05g09590 with unknown function (putative uncharacterized protein). Based on the online database analysis (Proost and Mutwil 2017), this gene was highly expressed at seed development stage S4–S5 (11–29 days after pollination) (Figure S3). Gene LOC_Os05g09590 was significantly associated with grain chalkiness (Misra et al. 2019). The SNP marker AX-154698806 was located between genes LOC_Os12g24450 and LOC_Os12g24460 (5849 bp upstream of LOC_Os12g24460). As a putative unclassified retrotransposon protein, LOC_Os12g24450 was highly expressed at seed

| Table 3 Pearson correlation among grain quality traits. Upper number is Pearson correlation, lower number is p value |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| BR              | WRRR            | GL              | WR              | WWRR            | GLWR            | CP              | CD              | AC              | GC              |
| 0.502***        | 0.251***        | 0.161**         | 0.212***        | 0.371***        | 0.451***        | 0.451***        | 0.617***        | 0.604***        | 0.261***        |
| 0.502***        | 0.251***        | 0.161**         | 0.212***        | 0.371***        | 0.451***        | 0.451***        | 0.617***        | 0.604***        | 0.261***        |
| 0.502***        | 0.251***        | 0.161**         | 0.212***        | 0.371***        | 0.451***        | 0.451***        | 0.617***        | 0.604***        | 0.261***        |
| 0.502***        | 0.251***        | 0.161**         | 0.212***        | 0.371***        | 0.451***        | 0.451***        | 0.617***        | 0.604***        | 0.261***        |

*Significance at p<0.05, **significance at p<0.01, ***significance at p<0.001

P value

*Significance at p<0.05, **significance at p<0.01, ***significance at p<0.001
development stage S5 (21–29 days after pollination). And LOC_Os12g24460, a putative uncharacterized protein, was highly expressed at seed development stage S4 (11–20 days after pollination) (Figure S3). Further validation of the effects of these genes on rice yield and grain quality should find ways to improve both grain yield and quality.

Predictive accuracy of GS models

High prediction accuracy is a prerequisite for successful application of genomic selection. The prediction accuracy is often measured by the correlation between observed phenotypes and the predicted GEBVs or predicted phenotypes of cross-validation (Xu 2017).
The predictive accuracy is influenced by several factors such as population size, variation within the training population and between the training and the test populations, heritability of a trait, marker density, and statistical method (Crossa et al. 2017a; Robertsen et al. 2019). In this study, the predictive abilities of a number of models were measured by Pearson and Spearman correlations between predicted GEBVs and actual values of the hybrid rice accessions. The predictive accuracy of the same trait varied among models, and the predictive accuracy of the same model also showed varying performance on different traits. No single model can be used for a good prediction of all the traits since the genetic structure of a trait is

![Diagram of QTLs for grain quality traits identified by GWAS](image)

**Table 4** Summary of significant association between two SNP markers and yield and grain quality-related traits

| Trait | Locus   | Marker   | Chr | Position (bp) | F    | p         | R²   |
|-------|---------|----------|-----|--------------|------|-----------|------|
| GN    | qGN5.1  | AX-155748928 | 5   | 5,417,532    | 29.22| 1.13E-07  | 0.0749|
| YLD   | qYLD5.1 | AX-155748928 | 5   | 5,417,532    | 37.23| 2.56E-09  | 0.0954|
| ALK   | qALK5.1 | AX-155748928 | 5   | 5,417,532    | 15.54| 9.61E-05  | 0.0400|
| CD    | qCD5.1  | AX-155748928 | 5   | 5,417,532    | 23.29| 2.01E-06  | 0.0601|
| CP    | qCP5.1  | AX-155748928 | 5   | 5,417,532    | 24.02| 1.41E-06  | 0.0617|
| OG    | qOG5.1  | AX-155748928 | 5   | 5,417,532    | 18.44| 2.23E-05  | 0.0474|
| WR    | qWR5.1  | AX-155748928 | 5   | 5,417,532    | 16.76| 5.17E-05  | 0.0438|
| GN    | qGN12.1 | AX-154698806 | 12  | 13,961,623   | 11.87| 1.00E-05  | 0.0626|
| PH    | qPH12.1 | AX-154698806 | 12  | 13,961,623   | 24.40| 1.09E-10  | 0.1265|
| TGW   | qTGW12.1| AX-154698806 | 12  | 13,961,623   | 13.07| 3.25E-06  | 0.0742|
| YLD   | qYLD12.1| AX-154698806 | 12  | 13,961,623   | 19.72| 7.24E-09  | 0.1068|
| ALK   | qALK12.1| AX-154698806 | 12  | 13,961,623   | 12.15| 7.72E-06  | 0.0641|
| CD    | qCD12.1 | AX-154698806 | 12  | 13,961,623   | 20.64| 3.14E-09  | 0.0991|
| CP    | qCP12.1 | AX-154698806 | 12  | 13,961,623   | 22.83| 4.44E-10  | 0.1209|
| OG    | qOG12.1 | AX-154698806 | 12  | 13,961,623   | 16.20| 1.80E-07  | 0.0846|
| TRANS | qTRANS12.1| AX-154698806 | 12  | 13,961,623   | 13.70| 1.81E-06  | 0.0726|
| WR    | qWR12.1 | AX-154698806 | 12  | 13,961,623   | 14.67| 7.36E-07  | 0.0785|
Fig. 4 Comparison of predictive accuracy of different GS models for different traits

Table 5 Statistics of predictive accuracy of 15 models for each trait

| Trait                          | Average | Minimum | Maximum | Best modela |
|-------------------------------|---------|---------|---------|-------------|
| Plant height                  | 0.626   | 0.545   | 0.655   | RKHS        |
| Panicle length                | 0.631   | 0.546   | 0.671   | LASSO       |
| Number of grains per panicle  | 0.491   | 0.349   | 0.538   | RKHS        |
| Spikelet fertility            | 0.320   | 0.183   | 0.398   | BayesA      |
| Thousand grain weight         | 0.810   | 0.722   | 0.836   | BayesB      |
| Yield                         | 0.304   | 0.178   | 0.349   | BayesA      |
| Brown rice rate               | 0.427   | 0.306   | 0.482   | RKHS        |
| White rice rate               | 0.424   | 0.295   | 0.469   | RKHS        |
| Grain length                  | 0.569   | 0.460   | 0.620   | LASSO       |
| Grain length/width ratio      | 0.631   | 0.448   | 0.707   | BayesB      |
| Chalk percentage              | 0.510   | 0.299   | 0.556   | RKHS        |
| Transparency                  | 0.299   | 0.259   | 0.349   | RFR         |
| Amylose content               | 0.630   | 0.462   | 0.778   | BayesB      |
| Alkali value                  | 0.665   | 0.584   | 0.771   | LASSO       |
| Gel consistency               | 0.544   | 0.360   | 0.724   | RFR         |

The best model is the model with the highest predictive accuracy among the 15 models.

Fig. 5 Comparison of predictive accuracy of different GS models with and without the significant SNP markers from the GWAS analysis. a Thousand grain weight; b amylose content
complex and diverse. In practical breeding, multiple GS models should be compared and the best model should be used for prediction of each trait.

Previous studies showed that the significant markers for GWAS have obvious effect on genomic prediction and can be used to assist in deciding what model strategies should be considered (Wilson et al. 2021). In this study, when comparing the predictabilities of models with or without considering the SNP markers from GWAS, almost all the models have higher predictive accuracy with the consideration of markers from GWAS. Thus, markers associated with the trait should be considered in the genomic selection models for better predict accuracy.

Predictive accuracy and potential application of GS models

At present, there is no model that can be widely applied to all traits. Though the stability and accuracy of GS models are continuously improved over time, there are still two main challenges, namely, computational accuracy and computational efficiency. The direct method (represented by GBLUP) had the higher calculation efficiency (i.e., The time and resources required for the computation), but lower calculation accuracy when compared with the indirect method (represented by BayesB). The other factors which perplexed the direct method were the setup of parameters which highly depend on researchers’ experiences, as various parameter setups have profound effects on the final results. Similarly, though the indirect method has high accuracy, it is difficult to effectively guide breeding practice because of the large amount of calculation in the process of parameter solution and the inability to realize parallel operation. In this study, BayesB and RKHS had high predictive accuracy for plant height, panicle length, thousand grain weight, grain length and width ratio, amylose content, and alkali value; while BRNN, GBLUP, rrBLUP, RFC, SPLS, SVC, and SVR had low predictive accuracy for other traits (Table S4). The predictive accuracy was low (0.22–0.35) for grain yield, but high for yield component traits such as number of grains per panicle, spikelet fertility, and thousand grain weight. Thus, it will be useful to predict the yield-related traits such as grain weight rather than the yield itself. The predictive accuracies for grain quality traits were higher, with the only exception of transparency.

The goal of GS for hybrid breeding is using the genotypes of the parents to predict the performance of the hybrids, which will significantly reduce the number of crosses for field test (Xu et al. 2014, 2018; Labroo et al. 2021). Although only F1 hybrids were investigated in this study, in practical breeding, only the parents (sterile lines and restore lines) need to be genotyped, and then genotypes of the F1 hybrids can be simulated and used for prediction of yield and grain quality traits by the GS models. This will significantly reduce the number of crosses to be made and the hybrids to be tested; thus, genomic selection is more efficient for hybrid rice breeding.

Conclusions

In this study, a total of 404 hybrid rice accessions were genotyped by 56 K SNP chip, and 20 traits related to yield and grain quality were investigated. Sixty-seven significant loci were identified for the yield-related traits, and 123 significant loci were identified for the grain quality traits by genome-wide association study. Two of these loci associated with increasing grain yield but decreasing grain quality. Genomic selection models of 15 different prediction algorithms were established using the genotypic and phenotypic data. The GS models are useful for plant height, panicle length, thousand grain weight, grain length and width ratio, amylose content, and alkali value, but the predictive accuracy for other traits are low. Use of proper model for specific trait is important for successful genomic selection. The GS models could be used for prediction of some important traits of hybrid rice through the genotypes of the parental varieties.

Author contribution P Yu and C Ye wrote the original draft; L Li, H Yin, J Zhao, and Y Wang performed the experiments; Z Zhang, W Li, and Y Long analyzed the data; X Hu, J Xiao, G Jia, and B Tian designed and supervised this study.

Funding This study was financially supported by the National Key Research and Development Project of Ministry of Science and Technology (2017YFD0102002-4).

Data availability The data supporting the findings of this study are available within the article and its supplementary materials.
References

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48
Bradbury P, Zhang Z, Kroon D, Casstevens T, Ramdoss Y, Buckler E (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635
Chen J, Zhou H, Xie W, Xia D, Gao G, Zhang Q, Wang G, Lian X, Xiao J, He Y (2019) Genome-wide association analyses reveal the genetic basis of combining ability in rice. Plant Biotechnol J 17(11):2211–2222
Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. PLOS ONE 11(6):e0156744
Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquin D, Campos G, Burgueño J, González-Camacho J, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney R (2017a) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22(11):961–975
Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquin D, Campos G, Burgueño J, González-Camacho M, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney K (2017b) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975
Cui Y, Li R, Li G, Zhang F, Zhu T, Zhang Q, Ali J, Xu S (2020) Hybrid breeding of rice via genomic selection. Plant Biotechnol J 18(1):57–67
Doyle J, Doyle J (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemistry Bulletin 19:11–15
García-Ruiz A, Cole J, anRaden P, Wiggins G, Ruiz-López F, Tassell C (2016) Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. PANS 113(28):3995–4004
Grenier C, Cao T, Ospina Y, Quintero C, Châtel M, Tohme J, Courtois B, Ahmadi N (2015) Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. PLOS ONE 10(8):e0136594
Hassen M, Cao T, Bartholome J, Orasen G, Colombi C, Rakotomalala J, Razafinimpiasa L, Bertone C, Biselli C, Volante A, Desiderio F, Jacquin J, Vale G, Ahmadi N (2018) Rice diversity panel provides accurate genomic predictions for complex traits in the progenies of biparental crosses involving members of the panel. Theor Appl Genet 131:417–435
Huang X, Yang S, Gong J, Zhao Y, Feng Q, Gong H, Li W, Zhan Q, Cheng B, Xia J, Chen N, Hao Z, Liu K, Zhu C, Huang T, Zhao Q, Zhang L, Fan D, Zhou C, Lu Y, Weng Q, Wang Z, Li J, Han B (2015) Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat Commun 6:6258
Huang M, Balimponya E, Mgonja E, McHale L, Kihupi A, Wang G, Sneller C (2019) Use of genomic selection in breeding rice (Oryza sativa L.) for resistance to rice blast (Magnaporthe oryzae). Molecular Breeding 39:114
International Rice Genome Sequencing Project, Sasaki T (2005) The map-based sequence of the rice genome. Nature 436:793–800
Isidro J, Jannink J, Akdemir D, Poland J, Heslot N, Sorrels M (2015) Training set optimization under population structure in genomic selection. Theor Appl Genet 128:145–158
Iwata H, Ebana K, Uga Y, Hayashi T (2015) Genomic prediction of biological shape: elliptic Fourier analysis and kernel partial least squares (PLS) regression applied to grain shape prediction in rice (Oryza sativa L.). PLOS ONE 10(3):e0120610
Jena K, Mackill D (2008) Molecular markers and their use in marker-assisted selection in rice. Crop Sci 48(4):1266–1276
Labroo M, Ali J, Aslam M, de Asis E, dela Paz M, Sevilla M, Lipka A, Studer A, Rutkoski J (2021) Genomic prediction of yield traits in single-cross hybrid rice (Oryza sativa L.). Frontiers in Genetics 12:692870
Li J, Wang J, Zeigler R (2014) The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience 3:8
Li F, Xie J, Zhu X, Wang X, Zhao Y, Ma X, Zhang Z, Rashid MAR, Zhang Z, Zhi L, Zhang S, Li J, Li Z, Zhang H (2018) Genetic basis underlying correlations among growth duration and yield traits revealed by GWAS in rice. Front Plant Sci 9:650
López M, Neira R, Yáñez J (2015) Applications in the search and prospect in China. J Integr Agric 14(2):197–205
Lu F, Yang F, Fan T, Liu J, Li Q, Wang L, Long X (2019) Analysis of rice variety approval data from 1977 to 2018. China Seed Industry 2:29–40
Ma G, Yuan L (2015) Hybrid rice achievements, development and prospect in China. J Integr Agric 14(2):197–205
McCouch S, Committee on Gene Symbolization NaL, Rice Genetics Cooperative, (2008) Gene nomenclature system for rice. Rice 1:72–84
Meuwissen T (2007) Genomic selection: marker assisted selection on a genome wide scale. J Anim Breed Genet 124(6):321–322
Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Code availability Not applicable.

Declarations

Ethics approval This article does not contain any studies with animals performed by any of the authors.

Consent to participate Not applicable.

Consent for publication All authors are consent to publication.

Competing interests The authors declare no competing interests.
Misra G, Anacleto R, Badoni S, Butardo V, Molina J, Graner A, Demont M, Morell M, Sreenivasulu N (2019) Dissecting the genome-wide genetic variants of milling and appearance quality traits in rice. J Exp Bot 70(19):5115–5130

Misztal I, Legarra A (2017) Invited review: efficient computation strategies in genomic selection. Animal 11(5):731–736

Mucha S, Mrode R, MacLaren-Lee I, Coffey M, Conighton J (2015) Estimation of genomic breeding values for milk yield in UK dairy goats. Dairy Science 98(11):8201–8208

Onogi A, Ideta O, Inoshita Y, Ebana K, Yoshioka T, Yamasaki M, Iwata H (2015) Exploring the areas of applicability of whole-genome prediction methods for Asian rice (Oryza sativa L.). Theor Appl Genet 128:41–53

Proost S, Mutwil M (2017) PlaNet: comparative co-expression network analyses for plants. Methods Mol Biol 1533:213–227

Robertsen C, Hjortshøj R, Janss L (2019) Genomic selection in cereal breeding. Agronomy 9:1–16

Samore A, Fontanesi L (2015) Genomic selection in pigs: state of the art and perspectives. Ital J Anim Sci 15(2):211–232

Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, Atlin G, Jannink J, McCouch S (2015) Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite tropical rice breeding lines. PLOS Genetics 11(2):e1004982

Su J, Xu K, Li Z, Hu Y, Hu Z, Zheng X, Song S, Tang Z, Li L (2021) Genome-wide association study and Mendelian randomization analysis provide insights for improving rice yield potential. Sci Rep 11:6894

Wang X, Li L, Yang Z, Zheng X, Yu S, Xu C, Hu Z (2017) Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. Heredity 118:302–310

Wang W, Mauleon R, Hu Z, Chebotarod D, Tai S, Wu Z, Li M, Zheng T, Fuentes R, Zhang F, Mansueto L, Copetti D, Sanjicangco M, Palis K, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrina D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat V, Borja F, Mendoza J, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo M, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann J, Zhang J, Li J, Hamilton R, Wing R, Ruan J, Zhang G, Wei C, Alexandrov N, McNally K, Li Z, Leung H (2018) Genomic variation in 3010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Wilson S, Zheng C, Maliepaard C, Mulder H, Visser R, Burgt A, Eeuwijk F (2021) Understanding the effectiveness of genomic prediction in tetraploid potato. Front Plant Sci 12:1–13

Wolc A, Zhao H, Arango J, Settar P, Fulton J, O’Sullivan N, Preisinger R, Stricker C, Habier D, Fernando R, Garrick D, Lamont S, Dekkers J (2015) Response and inbreeding from a genomic selection experiment in layer chickens. Genet Sel Evol 47:59

Xiao N, Pan C, Li Y, Wu Y, Cai Y, Lu Y, Wang R, Yu L, Shi W, Kang H, Zhu Z, Huang N, Zhang X, Chen Z, Liu J, Yang Z, Ning Y, Li A (2021) Genomic insight into balancing high yield, good quality, and blast resistance of japonica rice. Genome Biol 22:283

Xu S, Zhu D, Zhang Q (2014) Predicting hybrid performance in rice using genomic best linear unbiased prediction. PNAS 111(34):12456–12461

Xu Y, Wang X, Ding X, Zheng X, Yang Z, Hu Z (2018) Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice 11(1):32

Xu S (2017) Predicted residual error sum of squares of mixed models: an application for genomic pPrediction. G3 (Bethesda) 7 (3):895–909

Yabe S, Yoshida H, Kajiya-Kanegae H, Yamasaki M, Iwata H, Ebana K, Hayashi T, Nakagawa H (2018) Description of grain weight distribution leading to genomic selection for grain-filling characteristics in rice. PLOS ONE 13(11):e0207627

Yao W, Li G, Yu Y, Ouyang Y (2018) FunRiceGenes dataset for comprehensive understanding and application of rice functional genes. GigaScience 7(1):gix119

Zhang Z, Zhang Q, Ding X (2011) Advances in genomic selection in domestic animals. Chin Sci Bull 56(25):2655–2663

Zhen G, Qin P, Liu K, Nie D, Yang Y, Deng X, He H (2017) Genome-wide dissection of heterosis for yield traits in two-line hybrid rice populations. Sci Rep 7:7635

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.