Additive and Dominant Loci Jointly Pyramiding the Grain Quality of Hybrid Rice

Lanzhi Li (lancy0829@163.com)
College of Plant Protection, Hunan Agricultural University https://orcid.org/0000-0002-1531-8500
Xingfei Zheng
Hubei Academy of Agricultural Sciences
Jiabo Wang
Northeast Agricultural University
Xueli Zhang
Hunan Agricultural University
Xiaogang He
Hunan Agricultural University
Liwen Xiong
Hunan Agricultural University
Shufeng Song
Hunan Hybrid Rice Research Center
Jing Su
Hunan Agricultural University
Ying Diao
School of Pharmaceutical Sciences Wuhan University
Zheming Yuan
Hunan Agricultural University
Zhiwu Zhang
Washington State University https://orcid.org/0000-0002-5784-9684
Zhongli Hu
Wuhan University https://orcid.org/0000-0002-7258-3829

Article

Keywords: Rice, hybrid, grain quality, genome-wide association study, genomic prediction

Posted Date: April 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1565349/v1

Version of Record: A version of this preprint was published at Nature Communications on July 4th, 2023.
See the published version at https://doi.org/10.1038/s41467-023-39534-x.

Abstract

Hybrid rice has an advantage in its heterosis, resulting in high yield and resistance to biotic and abiotic stress. However, genetic improvement of grain quality is more challenging in hybrid rice than in inbred rice due to additional complexity, such as the dominant effect. It is critical to identify a path to efficiently develop inbreds and identify their superior crosses for hybrid production. Here, we developed a pipeline for joint analysis of phenotypes, effects, and generations (JPEG) and analysed 113 inbred varieties as male parents, five tester varieties as female parents, and their 565 (113×5) hybrid testcrosses for 12 grain quality traits, including grain length and width, chalkiness, and amylose content. A total of 1,619,588 single nucleotide polymorphisms were obtained for the parent varieties and inferred for the hybrids using whole-genome sequencing with an average sequencing depth of 11×. Genome-wide association studies with JPEG identified 128 loci associated with at least one of the 12 traits. There were 44 and 97 loci with additive and dominant effects, respectively, including 13 overlaps. Among the 128 associated loci, 42 loci are located in or near (< 200 kb) 17 known genes. Pyramiding genetic loci with additive and/or dominant effects in genomic selection dramatically improves the accuracy of predicting hybrid performances. More than 45% of genetic variation was explained by the identified markers for all traits except one (30%). These results demonstrate that pyramiding genetically identified additive and dominant loci can afford substantial power for predicting hybrid performances, presenting opportunities to identify superior crosses from existing phenotypes and genotypes of inbreds and hybrids.

Introduction

More than half of the worldwide population consumes rice (*Oryza sativa* L.) as their staple food. With improvements in living standards, people are paying more attention to end-use cooking quality[1]. Rice quality is typically negatively correlated with yield[1]. The yield of many rice hybrids is higher than that of conventional inbred rice varieties, but the grain quality needs more improvement. To obtain both high-quality and high-yield rice hybrids, the quality traits of hybrids need to be genetically dissected thoroughly so that they can be used to identify superior hybrids[2,3].

Rice grain quality has been classified into milling, appearance, cooking and eating, and nutritional categories[1,4]. The rice milling quality determines the yield and appearance of rice after the milling process. Milling quality comprises the brown rice ratio, milled rice ratio, and head rice ratio (BRR, MRR, and HRR). A large HRR (all quality trait abbreviations are listed in **Box 1**) is one of the most important criteria for measuring milled rice quality. Appearance quality is how rice appears after milling and is associated with grain length (GL), grain width (GW), grain length–width ratio (GLWR), and translucency/chalkiness of the endosperm. Generally, most markets prefer translucent rice as opposed to chalky rice. Cooking and eating quality are the easiness of cooking as well as the texture, springiness, stickiness, and chewiness of cooked rice, which are controlled by starch physical–chemical properties and comprise amylose content (AC), alkali spreading value (ASV), and paste viscosity properties. The amylose content of rice is known to play a crucial role in determining its cooked texture. The alkali

spreading value is a standard assay used to classify processing and cooking quality. It provides a simple means of classifying rice into high, intermediate, and low gelatinization temperature types. Protein content (PC) is a major index of rice grain nutritional quality. Since storage protein affects rice texture and processing quality, an intermediate PC is preferred[4].

Many rice grain quality traits in inbreds have been well studied, and multiple genes have been cloned and localized. Their mechanisms and functions have also been investigated. Grain size is closely related to yield[4,5]. At present, dozens of grain size-related genes have been isolated from multiple rice germplasm resources, such as the genes *GS3*, *GL3.1*, and *GW7/GL7*, which control grain length[6–8]; the genes *GW2*, *GW5/qSW5*, and *GS5*, which control grain width[9–11]; and the genes *GS6*, *GS9*, *TGW6*, and *GW8/SPL16*, which control grain size[5,12–14]. Chalky grains are considered low quality because of their poor appearance and undesirable cooking and milling qualities[15]. The rice gene *OsRab5a* regulates endomembrane organization and storage protein trafficking in rice endosperm cells, which affects the formation of amyloplasts[16]. *Chalk5* encodes a vacuolar H+-translocating pyrophosphatase, which influences grain chalkiness in rice[17].

The amylose content (AC) has the greatest influence on the cooking and consumption qualities of rice. The synthesis of rice amylose is catalysed by granule-bound starch synthase protein, which is encoded by the *Waxy* and *Wx* [18]. There are several alleles of the *Wx* gene, including Wx^{a}, Wx^{b}, *wx*, Wx^{mp}, Wx^{op}, Wx^{in}, and Wx^{mq}. Zhou *et al.*[19] cloned a practical resistant starch gene. Rice varieties with an intermediate gel temperature, which is predominantly determined by the amylopectin structure, are generally preferred by consumers. The gene (chr06:6748398_6753302 (+ strand)), starch synthase II (*OsSSIIa*), is the major determinant of gel temperature [4].

In a hybridization program, recognition of the best combination of two (or more) parental inbreds to recognize superior hybrids is one of the most critical challenging problems[20]. Genetic dissection of hybrids is more difficult than that of inbreds, as nonadditive effects are involved in addition to additive genetic effects, such as dominant genetic effects. Furthermore, the joint analyses of these genetic effects require the integration of both inbred and hybrid populations. In many analyses, heterosis is derived as the difference between the hybrid and middle parent from parent–child trios to map loci associated with dominant effects separately. For example, heterozygous genotypes were coded to be homozygous genotypes of the reference allele in the dominant model and homozygous genotypes of the alternative allele in the recessive model[21].

An additive and dominant effect joint model demonstrated superiority over models with separated effects. A GWAS with 130,725 cattle demonstrated that the additive and nonadditive joint model identified six dominant loci with impacts exceeding the largest effect variant identified by the additive effect model[22]. When both hybrid and inbred parent populations are available, the differences and similarities among parental inbred phenotypes, hybrid phenotypes, general combining ability, and hybrid heterosis can be used to infer genetic effects. In maize, 1428 maternal inbred varieties were crossed with

30 paternal inbred varieties to generate 42,840 (1428×30) hybrids. There were 166 QTLs associated with three traits (days to tasseling, plant height, and ear weight). These QTLs were categorized into three classes, additive, dominant, or epistatic effects[23], using comparisons among models with a single effect in a single population.

Ideally, both additive and dominant genetic effects should be analysed simultaneously with both parental inbred and hybrid populations to maximize statistical power. In this study, we crossed 113 male inbreds with five female inbred parents and generated 565 (113×5) hybrid testcrosses. Both parental inbred varieties (V) and hybrid testcrosses (T) were phenotyped for 12 quality traits, including grain length and width, chalkiness, and amylose content. The parental inbred varieties were genotyped using whole-genome sequencing with an average sequencing depth of 11×, resulting in a total of 1,619,588 single nucleotide polymorphisms (SNPs). The genotypes of hybrids were inferred from the genotypes of their parents. General combinability (G) and heterosis (H) were derived for maternal inbreds and hybrids, respectively. Our objectives are 1) to develop a statistical model and a computing pipeline to simultaneously analyse additive and dominant genetic effects using both original phenotypes (V and T) and derived phenotypes (G and H) from parental inbreds and hybrids; 2) to identify genetic loci associated with additive and/or dominant genetic effects on the 12 grain quality traits; and 3) to develop and validate a pipeline to predict superior crosses based on phenotypes and genotypes of parental inbreds and partial phenotypes of hybrids.

Results

Rice grain quality traits in inbreds and hybrids

Most of the 12 rice quality traits had normal distributions in inbred varieties, such as PC, GW, MRR, and BRR (**Figure S1**). There were two traits that were skewed to one end (ASV towards the lower end and PGWC towards the upper end). High ASV and low PGWC are preferred by customers. Similarly, low CD and AC are preferred by customers[1]. These two traits exhibit a bimodal distribution. The deviation from normality among the four traits reveal the selection for quality improvement. High TD is preferred by customers; unfortunately, the five female parent inbred varieties are below average compared to the male parent inbred varieties. Consequently, hybrids had undesirable TD compared to their male parent inbred varieties.

In general, the performance of the hybrid lies between those of their male and female parents, suggesting an additive genetic effect and the critical role of selecting inbred varieties to improve hybrids. For example, all the female parent inbred varieties are more desirable than the average male parent inbred varieties in terms of CD and PGWC. Consequently, the hybrids were more desirable than their male parents for these two traits. Similar phenomena were also observed for other traits. The higher the performance of the female parent was, the higher the performance in the hybrids compared with their male parent inbred varieties. However, the order of female parents did not exactly remain in the hybrids, suggesting the complexity of hybrid quality traits (**Figure S2**). Taking PGWC as an example, the value of

Y58S was higher than the value of 3A. The median PGWC of hybrids parented by Y58S was lower than that of the hybrids parented by 3A. Another example is HRR. The HRR of inbred 3A was higher than that of GZ63. The median HRR of hybrids parented by 3A was lower than that of the hybrids parented by GZ63.

Relationship among quality traits

There was a strong positive correlation between CD and PGWC for the four phenotypes: V(.93), T(.92), G(.93), and H(.77) (**Figure S3**). There were no other traits that had a strong correlation with CD and PGWC. TD only had a strong positive correlation with CD and PGWC in datasets T and G. PC and ASV barely had correlations with any other traits. These relationships suggest the possibility of improving PC and ASV simultaneously without affecting other traits.

Similar to the literature, there was a moderate negative correlation between GL and GW for the four datasets, V(-.5), T(-.5), G(-.57), and H(-.37). GL had moderate negative correlations with CD and PGWC, while GW had moderate positive correlations with CD and PGWC. As a function of GL and GW, their ratio, GLWR (GL/GW), had a strong positive correlation with GL and a strong negative correlation with GW. GL, GW, and GLWR had a very weak correlation with the remaining traits, including AC, ASV, BRR, HRR, and MRR. The relationship among the traits indicated that the germplasm materials in this study may have been preserved by long artificial selection. These relationships should be continually considered during rice breeding for high yield and grain quality.

Properties of Single Nucleotide Polymorphisms

A total of 7,734,465 raw SNPs were obtained from 120 parental varieties genotyped by whole genome sequencing with the Illumina HiSeq2500 platform. Among all the raw SNPs, 1,619,588 SNPs passed filters and quality control (missing rate < 20%, MAF > 5%). Marker distributions are displayed as a heatmap for 12 chromosomes based on minor allele frequency (MAF). Most (95%) intervals of adjacent SNPs were less than 200 kb. The average distance between markers was 196.8 bp. Approximately 58.88% of distances between pairs of SNPs were less than 50 bp. As a strict inbreeding species, rice has a low LD decay rate. In our population, the average LD did not decay below a r^2 value of 0.3 on average within 200 kb in parent inbred varieties (**Figure S4**).

Population Structure Analyses

Most of the female parents (four out of five) belong to one end of the phylogenetic tree developed from the 1,619,588 filtered SNPs. The others lie in the middle with the other end of the phylogenetic tree free of female parents, which creates the potential for heterosis (**Figure S5A**). Similar phenomena were also revealed by the principal component analysis (PCA based on the 1,619,588 filtered SNPs. The first three principal components (PCs) explained 43.4% of the total variation in hybrid test crosses, compared to 27.6% in inbred varieties (**Figure S6**). The population structure in hybrid testcrosses is stronger than the one in inbred varieties. All the female parents are in the middle of PC1 and the lower part of the PC2 (**Figure S6A to D**). Both the phylogenetic tree and principal component plots demonstrate that female

parents AS and GZ63 are genetically similar to each other. Similarly, the female parent Y56S is also similar to PA64. Consequently, the hybrids of Y56S and the hybrids of PA64 are close to each other. The hybrids of AS and GZ63 are completely mixed in the principal component plots (**Figure S6E to H**).

Four inbred varieties are differentiated from the rest as outliers. They are the four inbred varieties with the lowest PC3. Varylava has the lowest PC3, and IRAT109 is almost identical to Varylava for all PCs. Nanjing 11 has the third lowest PC3. This variety is a type of early-season rice. N22 has the fourth lowest PC3, with partial japonica components. All four outliers have foreign origins. The remaining 111 male parents were medium- or late-season indica rice varieties. The subpopulation structure was investigated using ADMIXTURE software. The cross-validation demonstrated that K = 5 reached the minimum error (**Figure S5B** and **C**).

Whole-genome Association Studies

We developed the JPEG pipeline (Fig. 1) and conducted three analyses with inbred varieties, hybrid testcrosses, and their combination. The combined analysis performed on the four datasets (V, T, G, and H) gave a higher power than analyses based on inbred varieties (V and G) or hybrid testcrosses (T and H). The larger the sample size of the dataset was, the more significant the loci that were identified. In total, 192 SNPs were identified to be associated with at least one of the 12 traits. After merging SNPs within 200 kb, 128 loci remained in the end. The middle positions of the merged SNPs were used as the locations of the loci. The analyses with inbreds, hybrids, and their combination identified 14, 68, and 89 loci, respectively, that associated with at least one of the 12 traits in formats of additive or dominant effects. Most of these loci demonstrate additive and dominant effects independently, while 13 of them demonstrated both (Figs. 2 and 3, **Table S6**). There were 44 and 97 loci with additive and dominant effects, respectively. Among the 13 loci with both additive and dominant effects, eight were identified by all three analyses. Four of the eight loci hit known genes, including *GS3* for GL and *GW5* for GW. The other two genes *(Wx/SSG6* and *AlaAT/OsAlaAT1)* are responsible for the AC. Among the 128 associated loci, 42 loci are located in or near (< 200 kb) 17 known genes (**Table S6**).

Many known genes are located within 200 kb of the identified loci. Three major genes of quality traits were all identified, including *Wx* (for AC), *GS3* (for GL and GLWR), and *GW5* (for GW and GLWR). Additionally, *BG1* (for GW and GL) was found in the hybrid and combination dataset[24]. *OsGRF8* was found in the hybrids for GLWR and in the combined dataset for AC. This gene is a growth-regulating factor that affects grain length, grain width, and grain starch granule size[25]. For the milling quality traits (MRR, BRR, and HRR), one significant locus was detected in the inbred dataset, compared with 19 in the hybrid dataset and 24 in the combined dataset.

Genetic effect/locus-based prediction

With the associated loci identified by GWAS, an immediate practical question is how to identify the new superior testcross from the existing phenotypes and genotypes of inbred varieties and hybrid testcrosses. To answer this question, we first estimated the heritability of each trait to set up the goal of the upper

bound for predictability. Heritabilities were estimated in inbred varieties for additive genetic effects only and in hybrid testcrosses for additive, dominant, and total (additive + dominant) genetic effects (Fig. 4). The total heritability is the upper bound of accuracy used to predict hybrid performance using genetic effect loci that can be operated through breeding.

As the 192 associated SNPs were identified by using the observations of the hybrids, the predictions made with these SNPs cannot be used to assess applicability for the future to select superior testcrosses due to the overfitting problem. Therefore, we divided the entire population into training and testing populations and conducted GWAS on the training population only. The training population contained all inbred varieties. The hybrids were randomly divided into two groups. One group was taken as the testing population, and the other group was combined with inbred varieties as the training population. The selected SNPs were used to predict the phenotypes of the testing population. The observed phenotypes of the testing population were not used to determine the associated SNPs. They were used only to calculate the correlation coefficient with the predicted phenotypes. Cross validations were iterated until all folds were used as the testing population. This process was repeated 100 times. The mean of the squared correlation coefficients (R2_CV) for all iterations and replicates was used as the final assessment. The ratio of R2_CV to total heritability was used as the proportion of genetic variance explained by the selected SNPs.

Pyramiding genetic loci for the 12 traits with additive and/or dominant effects in GS dramatically improved the accuracy of predicting hybrid performance. More than 45% of the total genetic variation was explained by the identified markers for all traits except one (30%). These results demonstrate that pyramiding genetically identified additive and dominant loci can afford substantial power to predict hybrid performances, presenting opportunities to identify superior crosses from existing phenotypes and genotypes of inbred varieties and hybrids (Fig. 4).

Discussion

Hybrid phenotypes and genotypes

In hybrid rice production, hybrid seeds are consumer products, and their qualities are important to consumers. Hybrid plants are F1s, while the seeds they produce are F2s. The genotypes of F1 plants can be inferred from parent inbred varieties. The genotypes of F2s remain dynamic due to gamete recombination and chromosomal crossover. Although hybrid seeds have different genotypes, their phenotypes are substantially determined by the mother plant F1s. Although the seeds produced by the F1 plant are different from each other both genetically and phenotypically, these differences are part of the errors. The cancelation of such errors depends on the number of seeds phenotyped. Therefore, this study investigated the association between F1 genotypes and F1 phenotypes regarding the seeds they produce.

Integration of additive and dominant effects in multiple loci models

Reducing false-positives and increasing statistical power are both critical in GWAS. Multiple loci modes, such as MLMM and FarmCPU, and BLINK have advantages over single-locus models in both reducing false-positives and increasing statistical power. In multiple loci models, markers are iteratively incorporated as covariates when their associations with the trait are identified while testing markers one at a time. This makes it feasible for multiple loci models to incorporate both additive and dominant effects simultaneously by staging additive and dominant marker genotypes together. Genetic additive and dominant effects can be realized by parallelizing the additive genotypes and dominant genotypes side by side into one genotype matrix for testing. We chose BLINK over FarmCPU and MLMM because both FarmCPU and MLMM involve kinship derived from all markers. When additive and dominant marker genotypes are merged, the derived kinship loses its original property of being either additive kinship or dominant kinship. This process of using the multiple loci models avoids using the hard coding additive genetic effect coefficient and dominant genetic effect coefficient simultaneously, which is not an option for many GWAS software packages.

Known gene confirmation

In this study, the GWAS method was used to analyse the genetic basis of 12 rice grain quality traits in 113 parental varieties and their 565 hybrid testcrosses. A total of 192 significant SNPs clustered into 128 significant loci were associated with at least one of the 12 traits. Among these loci, 42 loci are located in or near (< 200 kb) 17 known genes (**Table S6**), while the potential roles of the other SNPs in affecting rice grain quality were newly discovered. They provide new information for rice quality trait breeding. Several of these associated loci were detected simultaneously in multiple analyses of the same trait or correlated traits. Some of them were in or near the cloned quality genes (*Wx*, *GW5*, *GS3*, *BG1*, etc.). In 2018, Liu et al.[26] reported that the rice grain yield quantitative trait locus *qLGY3* encodes the MADS-domain transcription factor *OsMADS1*, which acts as a key downstream effector of G-protein βγ dimers. They demonstrated that combining the *OsMADS1* lgy3 allele with high-yield-associated dep1-1 and gs3 alleles represents an effective strategy for simultaneously improving both the productivity and end-use quality of rice. This study identified a locus near *OsMADS15*, a family gene of *OsMADS1*, that was significantly associated with two closely related-to-yield traits, GW and GLWR, in hybrid and combined datasets. *OsMADS15* may be another candidate gene that simultaneously affects grain yield and quality, so it needs to be further investigated.

Potential to incorporate epistatic genetic effects

In the current study, we limited our analyses to additive and dominant effects only. JPEG can be extended to incorporate epistatic effects. For the additive genetic effects, genotypes are coded as 0 and 2 for the two homozygotes and 1 for the heterozygotes to form additive genotypes (A). For the dominant effect, genotypes are coded as 0 for two homozygous and 1 for heterozygous to form dominant genotypes (D). The full model codes the genotypes as dummy variables to represent the interaction between two loci. There are multiple reduced models, including additive by additive interaction (AA), dominant by dominant interaction (DD), and additive by dominant interaction (AD and DA) models. For example, in the DD model, only an individual of being heterozygous for both loci is coded as 1, and the rest are coded as 0. In

the AA model, an individual of being homozygous for the alternative alleles at both loci is coded as 4. An individual of being homozygous for the alternative alleles at one locus and being heterozygous at the other locus is coded as 2. An individual of being heterozygous for both loci is coded as 1. The rest are coded 0. Even with the reduced models, the computation increases dramatically. Restriction to loci with additive or dominant effects can make computation feasible at the cost of missing loci with epistatic effects only.

Conclusion

Dominant genetic effects contribute to many complex human diseases and phenotypic variations among animals and plants. Additive-nonadditive joint analysis is more powerful than additive effect analysis. When both parent inbred varieties and hybrid testcrosses are available, general combined ability and heterosis can be derived to further assist in dissecting additive and dominant genetic effects. However, both methods and computing tools are lacking for the joint analysis of phenotypes, effects, and generations (JPEG). Here, we developed the JPEG pipeline using BLINK software and analysed 113 male inbred varieties, five female inbred varieties, and their 565 (113×5) hybrid testcrosses for 12 rice quality traits and 1,619,588 SNPs. Numerous loci were identified, which not only included many known cloned genes but also demonstrated the feasibility of identifying superior testcrosses from parent inbred varieties and their partial hybrid testcrosses.

Material And Methods

Materials and field management

Following the North Carolina II (NC II) genetic mating design, 115 indica rice accessions (**Table S1**), as male parents, were crossed with five male sterile varieties, as female parents, to produce hybrid testcrosses. All seeds of the 115 male parental varieties were obtained from the China National Crop GenBank (https://www.cgris.net/cgris_english.html), which includes restorer varieties of 29 three-line wild-deficient hybrid rice and 86 accessions of microcore germplasm. In 2013, male parental varieties and hybrid testcrosses were planted in Wuhan, China. The field trials were designed as randomized blocks and repeated twice, with 20 plants per field planted at a density of 5 × 8 inches. Field management, including irrigation, fertilizer application, and pest control, followed normal agricultural practices. The grains were harvested when fully ripe. In total, phenotypes on 12 rice quality traits were obtained on 113 parental varieties (**Table S1** and **S2**) and 565 (113 × 5) hybrid testcrosses (**Table S4**). Two samples (V2 and V3) with missing phenotypes were excluded from the analyses.

Phenotyping grain quality traits

After the materials matured, three plants with uniform growth were selected from the middle eight plants, dried, and stored at room temperature for three months. They were then sent to the Agricultural College of Hunan Agricultural University for rice grain quality evaluation. Following the method of Lou et al.[27], 12

rice quality traits of male parental varieties and hybrid testcrosses, including the grain length (GL, mm), grain width (GW, mm), GL/GW ratio (GLWR), chalkiness degree (CD, %), percentage of grains with chalkiness (PGWC, %), transparency (transparent degree, TD), amylose content (AC %), alkali spreading value (ASV), PC (total protein content %), brown rice ratio (BRR, %), milled rice ratio (MRR, %) and whole milled rice rate (head rice ratio, HRR, %), were determined.

Resequencing, genotyping and imputation

For each parental variety, genomic DNA was prepared from a single plant for sequencing. A sequencing library was established following the Illumina protocol. The genomes of 118 parental varieties were sequenced on an Illumina HiSeq2500 platform with a 11× sequencing depth on average. '*Nipponbare*' was used as the reference genome, and BWA software[28] was used for all paired-end read mapping. Then, SNP calling and quality control were conducted by deleting SNPs with a missing rate > 20% and minor allele frequency < 5%. In total, 1,619,588 SNPs were obtained. Missing genotypes were imputed by NPUTE[29] (version 4.0). Hybrid testcross genotypes were inferred using parental SNP genotypic information. The genotypes were coded for both additive genotypes and dominant genotypes. The additive genotypes were coded as 0 and 2 for the two homozygotes and 1 for the heterozygotes. The dominant genotypes were coded as 0 for the two homozygotes and 1 for the heterozygotes.

Phylogenetic and population structures

Neighbour-joining (NJ) trees and principal component analysis (PCA) plots were used to infer the structures of the 118 rice parental varieties and 565 hybrid testcrosses. A pairwise distance matrix derived from the simple matching distance for all the SNP sites was calculated to construct unweighted NJ trees using the software SNPhylo[30] (version 20140116), and phylogenetic trees were drawn by iTOL online (http://itol.embl.de/). The program Admixture (version 1.3) was used to perform optimization on the levels of K from 2 to 10[31]. PCA was conducted using GAPIT[32] (version 3.0) and performed separately for parental varieties and hybrid testcrosses. Genome-wide linkage disequilibrium (LD) was estimated using pairwise r^2 values between SNPs, which were calculated using the –r2 –ld-window99999 –ld-window-r2 0 command in PLINK[33] (version 1.9).

Phenotypic and heterotic statistics

The phenotypes of a trait are denoted V for the inbred varieties and T for the testcross. Two additional types of observations were derived from V and T, including general combining ability (G) for inbred varieties (**Table S3**) and heterosis (H) for hybrid testcrosses (**Table S5**).

The general combined ability of an inbred variety i was derived from its testcross using the following formula:

$$G_i = \bar{y_i} - \bar{y_{..}}$$

1

where G_i represents the G of the ith paternal parent, $\bar{y}_i$ represents the mean phenotypic value of a testcross derived from the ith parent, and $\bar{y}_{..}$ represents the mean phenotypic value of all the hybrid testcrosses.

Heterosis was defined as the difference between a testcross and the average of its parents as indicated by the following formula:

$$H_{ij} = T_{ij} - (V_i + V_j)/2 \quad (2)$$

where T_{ij} represents the phenotypic value of the testcross hybrid derived from the ith male parent and jth female parent, V_i represents the phenotype of the ith male parental varieties, and V_j represents the jth female parent phenotypic value.

GWAS pipeline across parents and hybrids for both additive and dominant effects

The four types of observations (V, T, G, and H) were normalized within each type to eliminate the difference between scales and averages. They were staged together as a single phenotype vector (Y) in the combined analysis. Their corresponding genotype matrix was staged correspondingly. V and G share the same genotype matrix, while T and H share the same genotype matrix (Fig. 1). Homozygous genotypes are coded as 0 and 2, while heterozygous genotypes are coded as 1 in the additive genotype matrix (A). Similarly, both homozygous genotypes are coded as 0, while the heterozygous genotype is coded as 1 in the dominant genotype matrix (D). Note that the dominant genotypes are all 0s for V and G. The additive genotype matrix and dominant genotype matrix were staged together (left and right) for GWAS with the BLINK[34] multiple loci model implemented in GAPIT[32] (version 3.0). At the end of the iterations for testing additive and dominant marker effects, the associated marker effects, either additive or dominant, were selected as covariates to test the model. We named this pipeline the joint analysis of phenotypes, effects, and generations (**JPEG**). The additional fixed effect covariates included the first three principal components derived from the additive genotypes of all SNPs and the dummy variables of female parents for the testcross. The association was determined with the threshold of 1% type I error after Bonferroni multiple test correction on both additive and dominant markers.

Heritability estimation among inbred varieties and hybrids

Additive genotype matrices were used to derive the additive genetic kinship matrices among inbred varieties and hybrid testcrosses. Similarly, a dominant genetic marker matrix was used to derive dominant kinship among hybrid testcrosses. Additive kinship and dominant kinship matrices define the covariance structure of the additive genetic effect and the dominant genetic effect. A mixed linear model with random additive and dominant genetic effects was solved by the BGLR[35] software package using the Gaussian processes (RKHS) algorithm to estimate the additive variance, dominant genetic variance, and residual variance. The proportions of the additive genetic variance, dominant genetic variance, and total genetic variance (additive + dominant) over the total variance (residual variance included) were

calculated as the corresponding additive heritability, dominant heritability, and total heritability, respectively.

Genetic effect-locus-based prediction

A critical breeding goal is to identify new superior crosses with existing genotypes of inbred varieties and phenotypes of inbred varieties and their hybrids. The predictions for hybrids using the associated additive and dominant loci cannot be used to evaluate their predictability because of the overfitting that these loci were derived from all hybrids. To assess the capability, we reconducted GWAS with cross-validations.

We randomly divided the hybrids into two groups. One group of hybrids was selected as a testing population. The remaining hybrid group and the parent inbred varieties were used as the training population. The full model with both additive and dominant genetic effects was used to conduct GWAS with the training population by BLINK on the full phenotypes, including V, T, G, and H. The selected SNPs with additive effects and SNPs with dominant effects in the final iteration of BLINK were used to predict the phenotypes of the hybrids in the testing population based on their genotypes. The correlation coefficient between observed and predicted observations was calculated for the hybrids in the testing population. The testing population was iterated until all groups were tested. This process was repeated 100 times. The mean squared correlation coefficient across all groups and replicates was used to assess the capability to identify new superior crosses.

Declarations

Author contribution

LL, XZ, and JW performed data analyses and drafted the manuscript. XZ, XH, LX, SS, JS, WT, ZY and ZT assisted in data analysis and the discussion. YD and XZ developed the populations. XZ and XH assisted in field data collection. ZH conceived the project and planned and secured extramural funds. ZZ designed the data analyses and revised the manuscript.

Acknowledgements

This work was partially supported by the National Key Technology Research and Development Program (2016YFD0100101), the Open Research Fund of State Key Laboratory of Hybrid Rice (Hunan Hybrid Rice Research Center (2019KF05), Wuhan University (KF201912), the Research Foundation of Education Bureau of Hunan Province (19A244), the Natural Science Foundation of Hunan Province (2020JJ4039), Special Funds for Construction of Innovative Provinces in Hunan Province (2021NK1011), the Key Research and Development Program of Hubei Province (2021BBA223), the Sichuan Science and Technology Program, China (Award #s 2021YJ0269 and 2021YJ0266), the USDA National Institute of Food and Agriculture (Hatch project 1014919 and Award # 2018-70005-28792) and the Washington Grain Commission (Award #s 126593 and 134574). The authors thank Prof. Sibin Yu and Tongmin Mou

from Huazhong Agricultural University, Prof. Shuangcheng Li from Sichuan Agricultural University, Prof. Shuzhu Tang from Yangzhou University, Prof. Guanghua He from Southwest University, and Prof. Huaxiong Qi from Hubei Academy of Agricultural Sciences for providing materials.

Data availability

All phenotypic data are included in the supplementary. The genotype data for the inbred varieties is available in HapMap format (hmp11.txt.zip) at https://zzlab.net/share.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Huggins, T. D. *et al.* Association Analysis of Three Diverse Rice (Oryza sativa L.) Germplasm Collections for Loci Regulating Grain Quality Traits. Plant Genome **12**, 170085 (2019).
2. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet **42**, 961–967 (2010).
3. McCouch, S. R. *et al.* Development of genome-wide SNP assays for rice. Breed. Sci. **60**, 524–535 (2010).
4. Asante, M. D. Breeding Rice for Improved Grain Quality. Adv. Int. Rice Res. (2017). doi:10.5772/66684
5. Che, R. *et al.* Control of grain size and rice yield by GL2-mediated brassinosteroid responses. Nat. Plants **1**, 1–7 (2015).
6. Mao, H., Ren, J., Ding, N., Xiao, S. & Huang, L. Genetic variation within coat color genes of MC1R and ASIP in Chinese brownish red Tibetan pigs. Anim. Sci. J. (2010). doi:10.1111/j.1740-0929.2010.00789.x
7. Wang, S. *et al.* The OsSPL16-GW7 regulatory module determines grain shape and simultaneously improves rice yield and grain quality. Nat. Genet. **47**, 949–954 (2015).
8. Wang, Y. *et al.* Copy number variation at the GL7 locus contributes to grain size diversity in rice. Nat. Genet. **47**, 944–948 (2015).
9. Song, X. J., Huang, W., Shi, M., Zhu, M. Z. & Lin, H. X. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. Nat. Genet. **39**, 623–630 (2007).
10. Wang, E. *et al.* Control of rice grain-filling and yield by a gene with a potential signature of domestication. Nat. Genet. **40**, 1370–1374 (2008).

11. Li, Y. *et al.* Natural variation in GS5 plays an important role in regulating grain size and yield in rice. Nat. Genet. **43**, 1266–1269 (2011).
12. Sun, L. *et al.* GS6, A Member of the GRAS gene family, negatively regulates grain size in rice. J. Integr. Plant Biol. **55**, 938–949 (2013).
13. Ishimaru, K. *et al.* Loss of function of the IAA-glucose hydrolase gene TGW6 enhances rice grain weight and increases yield. Nat. Genet. **45**, 707–711 (2013).
14. Song, X. J. *et al.* Rare allele of a previously unidentified histone H4 acetyltransferase enhances grain weight, yield, and plant biomass in rice. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 76–81 (2015).
15. Zhao, D. S. *et al.* GS9 acts as a transcriptional activator to regulate rice grain shape and appearance quality. Nat. Commun. **9**, (2018).
16. Wang, S. *et al.* Control of grain size, shape and quality by OsSPL16 in rice. Nat. Genet. **44**, 950–954 (2012).
17. Li, Y. *et al.* Chalk5 encodes a vacuolar H + -translocating pyrophosphatase influencing grain chalkiness in rice. Nat. Genet. **46**, 398–404 (2014).
18. Smith, A. M., Denyer, K. & Martin, C. The synthesis of the starch granule. Annu. Rev. Plant Biol. **48**, 67–87 (1997).
19. Zhou, H. *et al.* Critical roles of soluble starch synthase SSIIIa and granule-bound starch synthase Waxy in synthesizing resistant starch in rice. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12844–12849 (2016).
20. Fasahat, P. Principles and Utilization of Combining Ability in Plant Breeding. Biometrics Biostat. Int. J. **4**, 1–24 (2016).
21. Huang, X. *et al.* Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat. Commun. **6**, 1–9 (2015).
22. Reynolds, E. G. M. *et al.* Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nat. Genet. **53**, 949–954 (2021).
23. Xiao, Y. *et al.* The genetic mechanism of heterosis utilization in maize improvement. Genome Biol. **22**, 1–29 (2021).
24. Xie, W. *et al.* Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc. Natl. Acad. Sci.* **112**, E5411–E5419 (2015).
25. Yang, X. *et al.* OsmiR396/growth regulating factor modulate rice grain size through direct regulation of embryo-specific miR408. Plant Physiol. **186**, 519–533 (2021).
26. Liu, Q. *et al.* G-protein βγ subunits determine grain size through interaction with MADS-domain transcription factors in rice. Nat. Commun. **9**, 1–12 (2018).
27. Lou, J. *et al.* QTL mapping of grain quality traits in rice. J. Cereal Sci. **50**, 145–151 (2009).
28. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
29. Roberts, A. *et al.* Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. Bioinformatics **23**, 401–407 (2007).

30. Lee, T. H., Guo, H., Wang, X., Kim, C. & Paterson, A. H. SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics **15**, 1–6 (2014).
31. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**, 1655–1664 (2009).
32. Wang, J. & Zhang, Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics. Proteomics Bioinformatics **19**, 1–12 (2021).
33. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**, 559–575 (2007).
34. Huang, M., Liu, X., Zhou, Y., Summers, R. M. & Zhang, Z. BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience **8**, (2018).
35. Perez, P. BGLR: A Statistical Package for Whole Genome Regression and Prediction. Genetics **198**, 483–495 (2014).

Figures

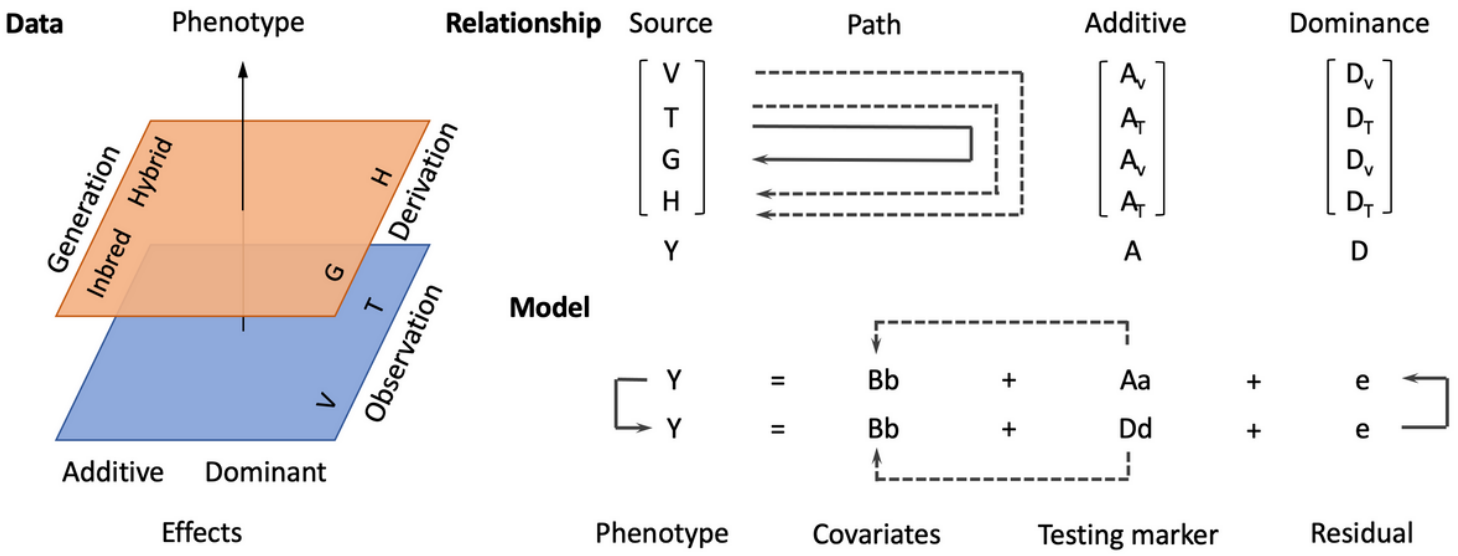

Figure 1

Analysis pipeline of joint phenotypes, effects, and generations (JPEG). The pipeline is summarized in the data panel on the left for genetic effects, generations from inbred varieties to hybrids, and phenotypes from original observations to their derivations. Original observations are measured for inbred varieties (V) and test crosses (T). The general combinability (G) of a variety was defined as the average performance of its testcross (the solid line in the relationship panel). Heterosis (H) of a test cross was defined as its difference from the middle parents, which is the mean performance of the parent inbred varieties (the dashed lines in the relationship panel). Standardization is performed within each of V, T, G, and H to form the joint observations (Y). Additive genotypes (A) and dominant genotypes (D) are recorded or derived for

the inbred variety (V) and test crosses (T), respectively (the data panel at the top). For an SNP, the additive genetic effect (a) is the regression performed on additive genotypes (A) and is coded as 0 and 2 for the two homozygotes and 1 for the heterozygotes. The dominant genetic effect (d) is the regression performed on dominant genotypes (D), which is coded as 0 for the two homozygotes and 1 for the heterozygotes. The JPEG pipeline can be conducted with multiple loci models such as BLINK to iteratively add associated additive effects and dominant effects as covariates (the black arrows in the model panel at the bottom). At the end of each iteration, the associated A or D was added to the covariates (B) to control the population structure and cryptic association among individuals (the dashed lines from A/D to B). The initial covariates include the first three principal components derived from all markers and the dummy variables of female parents for the testcross. The iterations stopped when no additive or dominant effect could be added to covariates.

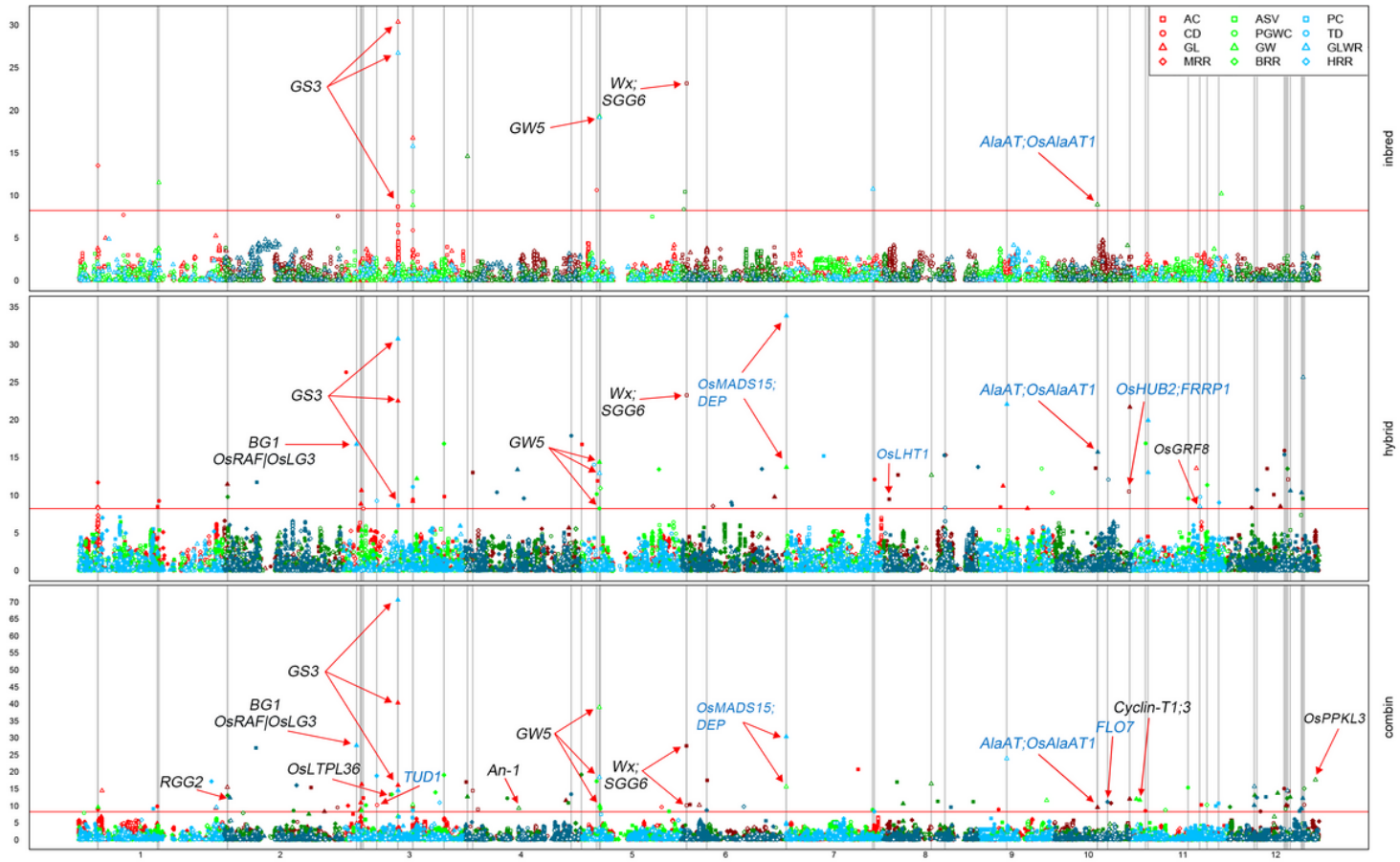

Figure 2

Associations of additive and dominant effects analysed using the JPEG pipeline among inbred varieties, hybrids, and their combination. The associations are illustrated as the negative -log10 P values of association tests on the 12 traits plotted against SNP positions for the 12 rice chromosomes. SNPs with additive and dominant effects are illustrated as open and filled symbols, respectively. The association analyses were conducted among inbred, hybrid, and their combined datasets. The inbred datasets include inbred observations and general combined ability. The hybrid datasets include hybrid observations and

heterosis. The whole-genome association threshold was set to 1% after Bonferroni correction (the horizontal red lines). Known cloned genes are labelled at the nearest associated SNPs. Clone genes influencing the same quality trait are labelled in black. Genes reported to influence related quality traits are labelled in blue. An associated SNP is marked with a vertical line if it has significant effects in more than one analysis or is associated with more than one trait.

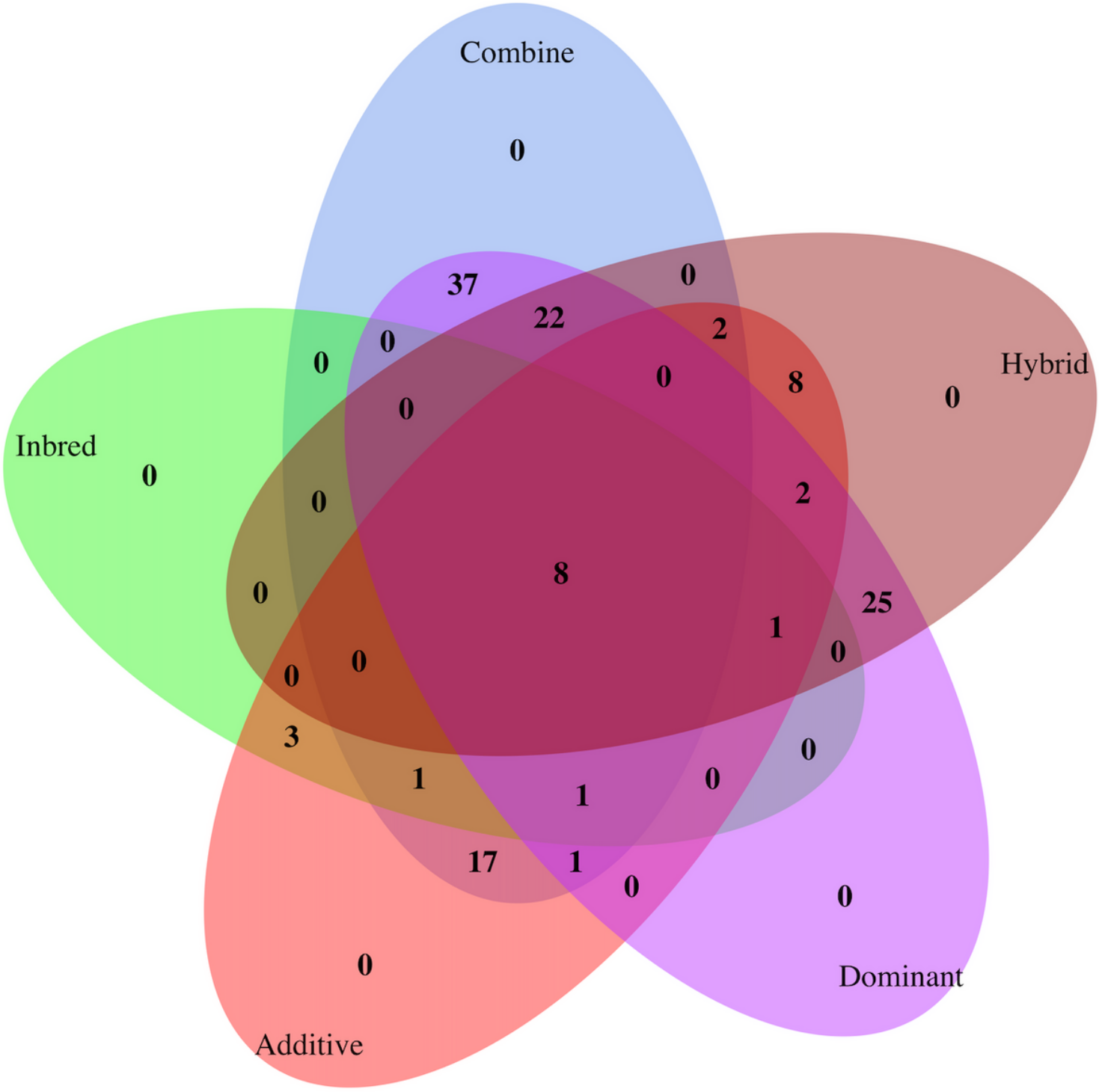

Figure 3

Venn diagram of the numbers of associated loci identified by analyses with inbred varieties, hybrids, and their combination for additive and dominant genetic effects. There were 128 loci associated with at least one of the 12 traits. Among these loci, 14, 68, and 89 were identified by analyses with inbred varieties, hybrids, and their combination, respectively. There are 9, 10, and 32 overlaps between inbred varieties and hybrids, between inbred varieties and the combination, and between hybrids and the combination, respectively. There were 44 and 97 loci with additive and dominant effects, respectively. There were 13 overlaps with both additive and dominant effects, including eight loci detected by all three analyses.

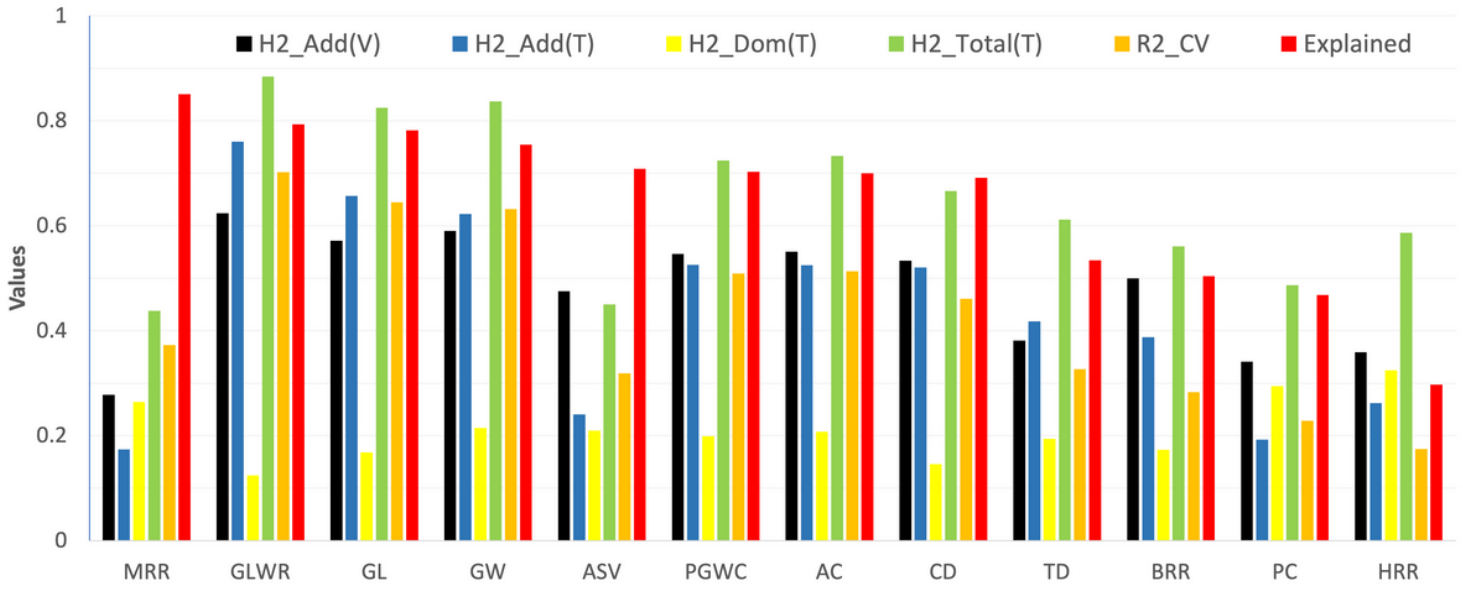

Figure 4

Genomic heritability and predictability using identified additive and dominant loci. The phenotypes of inbred varieties (V) were analysed to estimate additive heritability using additive kinship derived from the additive genotypes. The phenotypes of hybrid testcrosses (T) were analysed to estimate additive heritability and dominant heritability using additive kinship derived from the additive genotypes and dominant kinship derived from the dominant genotypes, respectively. The total heritability (additive + dominant) was calculated as the sum of the additive heritability and dominant heritability. The predictability of hybrids was evaluated through twofold cross-validation (CV). The squared correlation coefficients (R2) between the observed and predicted values were used to examine predictability (R2_CV). The ratio of R2_CV to the total heritability (H2_Total) indicated the percentages of the genome additive and dominant variances explained by the associated additive and dominant loci, which were over 50% for all traits except PC (45%) and HRR (30%).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RiceGWASXGSSupplementary0408.xlsx

- Supplementary.docx