

# The hybrid protein interactome contributes to rice heterosis as epistatic effects

Hong Li<sup>1,5,6,†</sup> , Shuqin Jiang<sup>1,†</sup>, Chen Li<sup>2</sup>, Lei Liu<sup>3</sup>, Zechuan Lin<sup>4</sup>, Hang He<sup>4</sup>, Xing-Wang Deng<sup>4</sup>, Ziding Zhang<sup>5,\*</sup>  and Xiangfeng Wang<sup>1,\*</sup>

<sup>1</sup>National Maize Improvement Center, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China,

<sup>2</sup>Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China,

<sup>3</sup>Beijing Key Laboratory of Plant Resources Research and Development, School of Sciences, Beijing Technology and Business University, Beijing 100048, China,

<sup>4</sup>State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, School of Advanced Agriculture Sciences and School of Life Sciences, Peking University, Beijing, China,

<sup>5</sup>State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China, and

<sup>6</sup>Key Laboratory of Tropical Biological Resources of Ministry of Education, School of Life and Pharmaceutical Sciences, Hainan University, Haikou 570228, China

Received 5 May 2019; revised 27 October 2019; accepted 1 November 2019; published online 17 November 2019.

\*For correspondence (e-mails xwang@cau.edu.cn and zidingzhang@cau.edu.cn).

†These authors contributed equally to this work.

## SUMMARY

Heterosis is the phenomenon in which hybrid progeny exhibits superior traits in comparison with those of their parents. Genomic variations between the two parental genomes may generate epistasis interactions, which is one of the genetic hypotheses explaining heterosis. We postulate that protein–protein interactions specific to F<sub>1</sub> hybrids (F<sub>1</sub>-specific PPIs) may occur when two parental genomes combine, as the proteome of each parent may supply novel interacting partners. To test our assumption, an inter-subspecies hybrid interactome was simulated by *in silico* PPI prediction between rice *japonica* (cultivar Nipponbare) and *indica* (cultivar 9311). Four-thousand, six-hundred and twelve F<sub>1</sub>-specific PPIs accounting for 20.5% of total PPIs in the hybrid interactome were found. Genes participating in F<sub>1</sub>-specific PPIs tend to encode metabolic enzymes and are generally localized in genomic regions harboring metabolic gene clusters. To test the genetic effect of F<sub>1</sub>-specific PPIs in heterosis, genomic selection analysis was performed for trait prediction with additive, dominant and epistatic effects separately considered in the model. We found that the removal of single nucleotide polymorphisms associated with F<sub>1</sub>-specific PPIs reduced prediction accuracy when epistatic effects were considered in the model, but no significant changes were observed when additive or dominant effects were considered. In summary, genomic divergence widely dispersed between *japonica* and *indica* rice may generate F<sub>1</sub>-specific PPIs, part of which may accumulatively contribute to heterosis according to our computational analysis. These candidate F<sub>1</sub>-specific PPIs, especially for those involved in metabolic biosynthesis pathways, are worthy of experimental validation when large-scale protein interactome datasets are generated in hybrid rice in the future.

**Keywords:** protein interactome, heterosis, epistatic effects, protein–protein interaction prediction, genomic selection analysis.

## INTRODUCTION

Heterosis (or hybrid vigor) refers to the phenomenon in which hybrid progeny exhibits superior traits in terms of growth, yield, biomass, stress tolerance or disease resistance compared with parents (Shull, 1908). Early in the 1870s, Charles Darwin documented that the offspring of

cross-pollinated maize were almost 25% taller than the offspring of self-pollinated maize. Maize and rice are among the earliest crops for which heterosis was utilized for breeding, and the crop yield and grain quality of these crops have been tremendously improved in the past century. Since the 1980s, hybrid maize has accounted for 80%

of the total maize production in China, and the yield of super hybrid rice peaked at 1149 kg per mu in 2017. However, the genetic and molecular mechanisms underlying heterosis remain poorly understood despite intensive investigations performed in recent decades.

Explanations for heterosis from quantitative genetics aspect include dominance, overdominance and epistasis hypotheses. The dominance hypothesis states that the recessive deleterious alleles from one parent can be complemented by the dominant beneficial alleles from another parent (Bruce, 1910; Jones, 1917). In the overdominance hypothesis, both alleles in hybrid progeny are equally important, and intra-allele interaction is the reason for expression of superior traits, as the fitness of individuals carrying heterozygous genotypes is higher than those carrying homozygous genotype (Shull, 1908; East, 1936; Hull, 1945; Crow, 1948). Both dominance and overdominance hypotheses are based on the action of a single gene, but most heterosis-expressing traits are quantitative and thus must involve multiple genes with different extents of effect (Hua *et al.*, 2003). The epistasis hypothesis emphasizes the role of inter-allele interactions among multiple genes and pathways, which may include all possible forms of molecular interactions, such as physical interactions among proteins and small molecules, transcriptional regulation in *cis* and *trans*, and regulatory interactions from the posttranscriptional, posttranslational and epigenetic machinery (Minvielle, 1987; Schnell and Cockerham, 1992; Sun and Kardja, 2010; Jiang *et al.*, 2017). Evidence supporting the role of epistasis in heterosis has been increasingly reported in recent years, but the precise manner through which epistasis acts via molecular interactions remains largely elusive due to the complexity of gene interactions (Yu *et al.*, 1997; Tang *et al.*, 2010; Zhou *et al.*, 2012; Schnable and Springer, 2013).

With the availability of high-throughput biotechnology, genome-wide investigations of heterosis mechanisms have been intensively conducted at the genomic, transcriptomic, proteomic and metabolomic levels (Baranwal *et al.*, 2012). Transcriptome comparison of the elite hybrid rice cultivar LYP9 and its two parental lines (maternal PA64 and paternal 9311) showed that differentially expressed genes accounting for 10.6% of the genome were mostly enriched in energy metabolism and energy transportation pathways (Wei *et al.*, 2009). In addition, the LYP9 transcriptome resembled that of PA64 during early developmental stages, but resembled that of 9311 at later stages, implying that complex regulatory processes underlie heterosis (Wei *et al.*, 2009). Comparisons of the metabolomes of parents and offspring also revealed higher metabolic activity in F<sub>1</sub> hybrids in comparison with their parents, indicating that heterosis may be at least partially attributed to enhanced energy use efficiency in hybrids (Meyer *et al.*, 2012). Genome resequencing of maize core germplasm revealed

widespread genomic variations, including a large number of presence and absence variations (PAV) in protein-coding genes, which may act as either deleterious or beneficial alleles. These alleles may complement with each other in the hybrid genome to result in heterosis (Springer *et al.*, 2009; Lai *et al.*, 2010). PAVs frequently occur on paralogous gene copies in maize. Loss and gain events for different paralog sets among inbred lines due to long-term artificial selection may form diverse genetic resources to produce a more complete proteome and more complex protein interactome in the hybrid genome in comparison with parental genomes, and thus heterosis occurs (Schnable and Springer, 2013).

Correlations between environmental responses and heterosis have also been reported in plants. Transcriptome analysis of the allotetraploid formed between *Arabidopsis thaliana* and *Arabidopsis arenosa* showed that genes encoding morning-phased clock regulators were downregulated in the allotetraploid in comparison with the two diploid parents during daytime (Ng *et al.*, 2017). This effect was coincident with upregulation of genes in photosynthetic and starch biosynthetic pathways, which promotes biomass heterosis (Ng *et al.*, 2017). Another study in natural allopolyploid *Arabidopsis suecica* and its progenitor species *A. thaliana* and *A. arenosa* showed that the allopolyploid assimilates more CO<sub>2</sub> per unit of chlorophyll in comparison with either of the two progenitor species under high-intensity light (Solhaug *et al.*, 2016). Under low- and high-intensity light conditions, starch accumulation exhibited significant differences between the allopolyploid and its diploid progenitors, suggesting that heterosis is perhaps a result of increased energy efficiency in combination with environmental responses.

Mechanisms of heterosis must be very complex, involve many genes with diverse genetic effects, and are associated with biological processes and molecular pathways. Over the past few decades, comprehensive evidence supporting dominance and overdominance theories has been provided, but evidence for the involvement of epistasis in heterosis, especially in connection with genomic and molecular machinery, remains poorly reported. Protein–protein interactions (PPIs), as one form of inter-allelic gene interaction, definitely contribute to heterosis via epistatic effects. However, it is extraordinarily difficult to obtain an accurate profile of the entire protein interactome due to its complex and dynamic nature, so the role of PPIs in heterosis remains essentially unknown. In this work, we applied a systems biology approach by incorporating protein interactome simulation and genomic selection (GS) prediction to investigate the role of the hybrid protein interactome in heterosis using rice as a model. Our analyses reveal that the protein interactome in F<sub>1</sub> hybrids is indeed more diverse than that of either parent and may contribute to heterosis via epistatic effects.

## RESULTS

### Simulation of the protein interactomes of F<sub>1</sub> hybrid rice

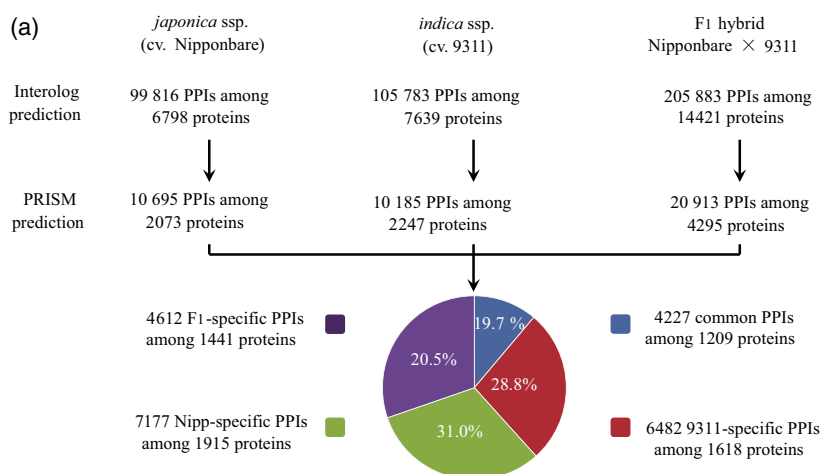
The proteome of an F<sub>1</sub> hybrid is a combination of the parental proteomes. Thus, more diversified PPIs and networks may exist in F<sub>1</sub> offspring in comparison with each parent. To simulate the F<sub>1</sub> (Nipponbare × 9311) interactome, *in silico* prediction of PPIs was conducted in two steps: within each parental proteome and between the two parental proteomes of *japonica* Nipponbare and *indica* 9311 (Experimental Procedures). In the first step, Interolog (Garcia-Garcia *et al.*, 2012) was used to build three preliminary interactomes, including 99 816 PPIs formed by 6798 proteins in Nipponbare, 105 783 PPIs formed by 7639 proteins in 9311, and 205 883 PPIs formed by 7590 and 6831 proteins in Nipponbare and 9311, respectively (Figure 1a). In the second step, PRISM (Tuncbag *et al.*, 2011) was used to screen high-confidence PPIs from Interolog's predictions, resulting in 10 695 PPIs formed by 2073 proteins in Nipponbare, 10 185 PPIs formed by 2247 proteins in 9311, and 20 913 PPIs formed by 2039 and 2256 proteins in Nipponbare and 9311, respectively (Figure 1a).

The simulated F<sub>1</sub> interactome was then categorized into four groups, namely 4612 (20.5%) F<sub>1</sub>-specific PPIs, 6482

(28.8%) 9311-specific PPIs, 7177 (31.0%) Nipponbare-specific PPIs, and 4227 (19.7%) common PPIs with the following rule that assumes A/A' and B/B' are two pairs of orthologs in Nipponbare and 9311:

1. Common PPIs: if A interacts with B in parent Nipponbare, and A' interacts with B' in parent 9311, while A interacts with B' and A' interacts with B in the F<sub>1</sub> hybrid;
2. F<sub>1</sub>-specific PPIs: if A does not interact with B in parent Nipponbare, and A' does not interact with B' in parent 9311, while A interacts with B' and A' interacts with B in the F<sub>1</sub> hybrid;
3. Nipponbare-specific PPIs: if A interacts with B in parent Nipponbare, and A' does not interact with B' in parent 9311, while A does not interact with B' and A' does not interact with B in the F<sub>1</sub> hybrid;
4. 9311-specific PPIs: if A does not interact with B in parent Nipponbare, but A' interacts with B' in parent 9311, while A does not interact with B' and A' does not interact with B in the F<sub>1</sub> hybrid.

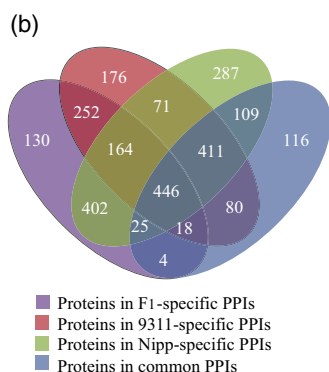
Proteins participating in the four groups of PPIs were then correspondingly categorized as 1441 proteins in F<sub>1</sub>-specific PPIs, 1618 proteins in 9311-specific PPIs, 1915 proteins in Nipponbare-specific PPIs, and 1209 proteins in common PPIs. Proteins may participate in the PPIs of more



**Figure 1.** Simulation of a hybrid protein interactome in the F<sub>1</sub> progeny of *japonica* ssp (cv. Nipponbare) and *indica* ssp (cv. 9311).

(a) Interolog software was first used to predict protein–protein interactions (PPIs) within the 9311 and Nipponbare proteomes, as well as between the two proteomes. PRISM was used to identify high-confidence PPIs from Interolog's predictions. High-confidence PPIs were classified as common PPIs, F<sub>1</sub>-specific PPIs, Nipponbare-specific PPIs or 9311-specific PPIs.

(b) Overlapping numbers of proteins in the four groups of PPIs.



than one group. While 1982 proteins (73.7%) participated in at least two groups of PPIs, 130 proteins only participated in F<sub>1</sub>-specific PPIs that were not found in the other three groups of PPIs (Figure 1b). These 130 proteins are worthy of special attention, as they may generate novel biological functions unique to F<sub>1</sub> hybrids. Functional annotation of these 130 proteins indicates that most of them possess enzymatic activities, such as serine acetyltransferase activity, nucleoside diphosphate kinase activity, or cysteine synthase activity (Table S1).

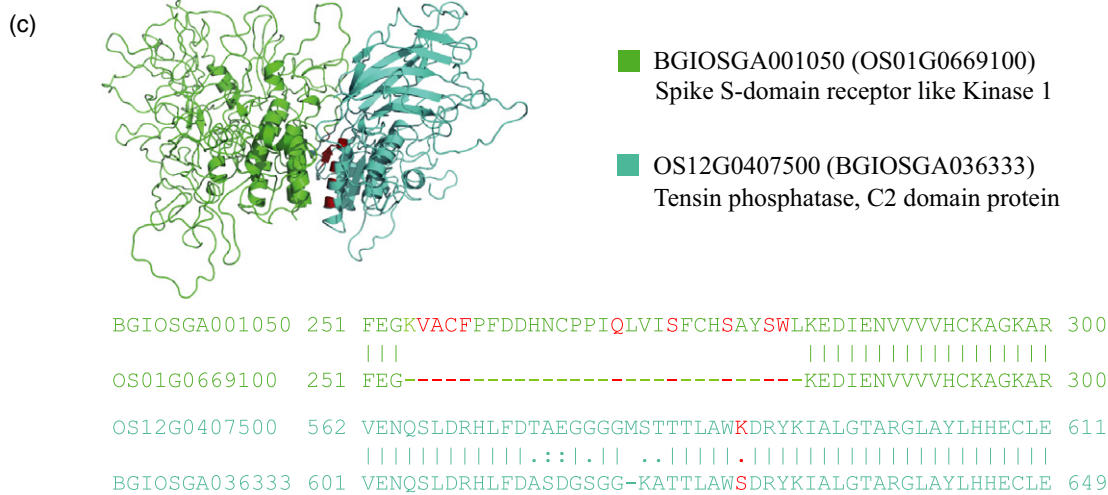
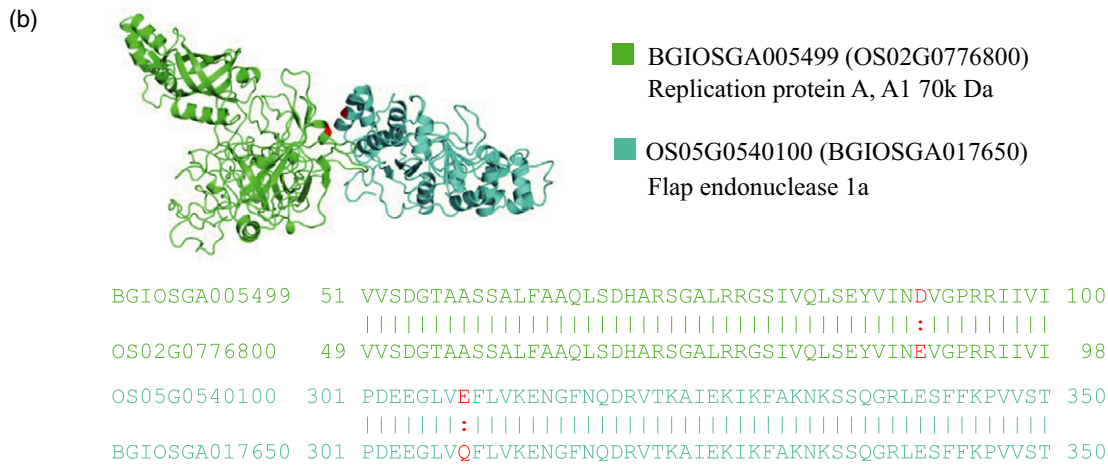
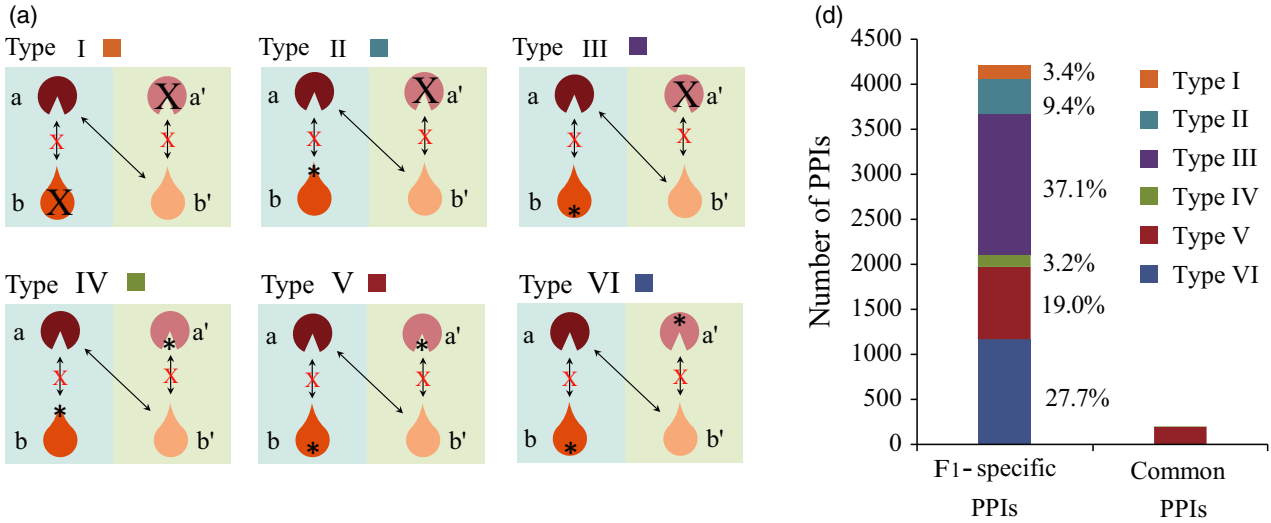
To assess the reliability of predicted F<sub>1</sub>-specific PPIs, we further examined the subcellular co-localization and functional similarity of interacting protein pairs in comparison with 1000 randomly generated PPI networks using the same set of genes involved in F<sub>1</sub>-specific PPIs. The results indicate that the number of co-localized PPIs and the GO similarities of F<sub>1</sub>-specific PPIs are higher than those of any random network (empirical  $P < 0.001$ ). Further, we examined the co-expression of *japonica* and *indica* alleles from the genes involved in F<sub>1</sub>-specific PPIs by using the RNA-seq data in accession of GS113769 (Chen *et al.*, 2018) in NCBI GEO (Experimental Procedures). In GSE113769, the RNA-seq samples were collected from the endosperm tissues of F<sub>1</sub> hybrids generated from two distinct, reciprocal crosses of *japonica* and *indica* rice. Out of the 1001 and 988 endosperm-expressed F<sub>1</sub>-specific genes, 996 (99.5%), 991 (99.0%), 979 (99.1%) and 965 (97.7%) were expressed (FPKM > 0) from both *japonica* and *indica* alleles in the four crosses (Table S2). The results imply that the majority of the genes involved in F<sub>1</sub>-specific PPIs exhibit co-expression of both *japonica* and *indica* alleles in the endosperm tissue of F<sub>1</sub> hybrids.

*F<sub>1</sub>-specific PPIs are attributed to PAVs and missense variations.* F<sub>1</sub>-specific PPIs may arise from two sources of genomic variations between *japonica* and *indica* rice: PAVs and missense variations occurring on protein-coding genes that cause allelic difference between the two parental genomes. For instance, if protein A in Nipponbare lacks its ortholog A' in 9311, and protein B (protein A's interacting protein) in Nipponbare has a defective missense variation, but its ortholog B' in 9311 is intact, an inter-subspecies PPI between A in Nipponbare and B' in 9311 may occur as an F<sub>1</sub>-specific PPI. In other words, if no interactions between A and B in Nipponbare or between A' and B' in 9311 are observed, but an interaction between A in Nipponbare and B' in 9311 is observed, this PPI is defined as an F<sub>1</sub>-specific PPI. To test this conjecture, 4612 F<sub>1</sub>-specific PPIs were classified into six types of PPIs based on PAVs and missense variations (Figure 2a). Figure 2(b) shows an example of a Type IV F<sub>1</sub>-specific PPI, in which a single non-synonymous single nucleotide polymorphism (SNP) that changes one amino-acid residue at the interaction interface may alter the interactions of ortholog pairs. BGIOGA005499 in 9311

and OS05G0540100 in Nipponbare form an F<sub>1</sub>-specific PPI. However, no PPI between OS02G0776800 (an ortholog of BGIOGA005499) and OS05G0540100 (an ortholog of BGIOGA017650) was predicted in Nipponbare, because one missense variation occurred on the 90th amino-acid residue, which was predicted to be an interacting residue between BGIOGA005499 and OS05G0540100. In 9311, no PPI between BGIOGA005499 (an ortholog of OS02G0776800) and BGIOGA017650 (an ortholog of OS05G0540100) was predicted due to the presence of another missense variation on the second interacting residue (the 308th amino acid). The second example in Figure 2(c) shows another F<sub>1</sub>-specific PPI formed by BGIOGA001050 and OS12G0407500, for which no PPI was predicted due to an 84-nt-long indel and a series of non-synonymous SNPs.

Almost half of the F<sub>1</sub>-specific PPIs were caused by PAVs, while the other half were attributed to missense variations caused by either non-synonymous SNPs or indel fragments on either orthologous pair (Figure 2d). It is worth noting that missense variations occurring on PPI interfaces and non-interface regions may abolish PPIs. Among the 27.7% of F<sub>1</sub>-specific PPIs in which both interacting partners have orthologs in the two parental genomes, missense variations were found outside the predicted PPI interface regions (Figure 2d). Missense variations outside the PPI interface may be due to protein structural changes or the failure of PRISM to predict the PPI interface region. To exclude the possibility that causation of F<sub>1</sub>-specific PPIs by missense variations is a coincidence, we also examined Type IV, Type V and Type VI PPIs among the 4227 common PPIs, which revealed that only 191 (4.5%) common PPIs exhibited interface missense variations on at least one partner protein in each PPI (Figure 2d). The significant difference between common and F<sub>1</sub>-specific PPIs indicates that PAVs and missense variations are the main sources of F<sub>1</sub>-specific PPIs in F<sub>1</sub> hybrids.

*Part of F<sub>1</sub>-specific PPIs are involved in metabolic pathways.* The enriched pathways of genes involved in F<sub>1</sub>-specific PPIs may reflect novel biological functions that enhance the trait performance of F<sub>1</sub> hybrids. Therefore, we subsequently analyzed the functional enrichment of proteins participating in F<sub>1</sub>-specific PPIs by comparing them with proteins in common PPIs using the PANTHER protein classification system (Mi *et al.*, 2016). When functional enrichments among the proteins participating in F<sub>1</sub>-specific PPIs, common PPIs and subspecies-specific PPIs were compared, no significantly enriched classes were found (Figure 3a). We suspect that this finding may be due to the large portion of proteins shared by the four groups of PPIs. Next, we focused our attention on the 130 proteins that only participate in F<sub>1</sub>-specific PPIs, and compared their functional annotations with the 493 and 818 proteins that





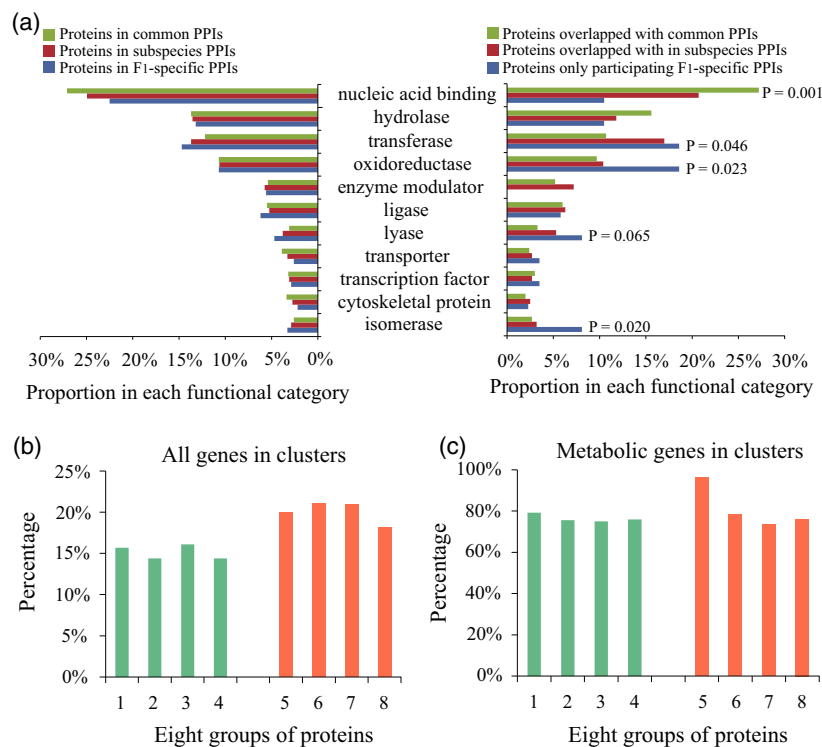
**Figure 2.** F<sub>1</sub>-specific protein–protein interactions (PPIs) arise from ortholog absence and/or missense variations.

(a) Scheme to illustrate six types of situations in which F<sub>1</sub>-specific PPIs are generated in the F<sub>1</sub> genome. Type I, a pair of interacting proteins in an F<sub>1</sub>-specific PPI both lack orthologs; Type II, one protein lacks an ortholog and its partner has an ortholog containing missense variations on the interaction interface; Type III, one protein lacks an ortholog and its partner has an ortholog containing missense variations not on the interaction interface; Type IV, both interacting proteins have orthologs, and both contain missense variations on the interface; Type V, both interacting proteins have orthologs, and one contains missense variations on the interface while the other does not; Type VI, both interacting proteins have orthologs, and both contain missense variations not on the interface.

(b) One example of a Type IV F<sub>1</sub>-specific PPI. BGIOSGA005499 and OS02G0776800 are a pair of orthologs in 9311 and Nipponbare, with only one non-synonymous single nucleotide polymorphism (SNP) causing an Asp-to-Glu substitution on the interacting residue. OS05G0540100 and BGIOSGA017650 are orthologs with a Glu-to-Gln substitution on the other interacting residue. Because of the two substitutions on the predicted interacting interface, BGIOSGA005499 and OS05G0540100 formed a between-subspecies PPI, but no PPIs were predicted between BGIOSGA005499 and BGIOSGA017650 in 9311 or between OS02G0776800 and OS05G0540100 in Nipponbare. The residues marked in red in the sequence alignment indicate the interaction interface that is also marked in red in the 3D structures of the interacting protein pair.

(c) Another example of an F<sub>1</sub>-specific PPI between BGIOSGA001050 in 9311 and OS12G0407500 in Nipponbare. Because of a 24-AA deletion on OS01G0669100, no PPI was detected between OS01G0669100 and OS12G0407500 in Nipponbare. Because of another Lys-to-Ser substitution on the interacting residue, no PPI was detected between BGIOSGA001050 and BGIOSGA036333 in 9311.

(d) Proportions of the six types of genomic variations causing F<sub>1</sub>-specific PPIs.

**Figure 3.** F<sub>1</sub>-specific protein–protein interactions (PPIs) are likely enriched in enzyme interactions and metabolic pathways.

(a) Functional categorization of the proteins in different types of PPIs. The top 11 functional categories are shown. Each category contains at least 100 proteins out of the total set of 6183 proteins. The histogram on the left shows the proportions of the 1441 proteins in F<sub>1</sub>-specific PPIs, 1209 proteins in common PPIs, and 3533 proteins in subspecies-specific PPIs in each of the 11 protein functional categories. No significantly enriched category was detected in any of the three types of proteins. Because a large number of proteins in the four groups overlapped, the 1441 proteins in the F<sub>1</sub>-specific PPIs were further classified into a set of 130 proteins only participating in F<sub>1</sub>-specific PPIs, as well as 493 and 818 proteins also participating in common and subspecies-specific PPIs, respectively. The proportions of these three types of proteins in the 11 categories are shown on the right histogram.

(b) Percentages of the eight groups of proteins present in metabolic gene clusters (MGCs). 1. Proteins involved in F<sub>1</sub>-specific PPIs; 2. Proteins involved in Nipponbare-specific PPIs; 3. Proteins involved in 9311-specific PPIs; 4. Proteins involved in common PPIs; 5. Proteins only involved in F<sub>1</sub>-specific PPIs; 6. Proteins only involved in Nipponbare-specific PPIs; 7. Proteins only involved in 9311-specific PPIs; 8. Proteins only involved in common PPIs.

(c) Percentages of genes in the clusters that were identified as metabolic genes. The eight groups are the same as those shown in (b).

also participate in common PPIs and subspecies-specific PPIs, respectively (Figure 3a). While the ‘oxidoreductase’, ‘transferase’ and ‘isomerase’ categories showed statistically significant enrichment ( $P \leq 0.05$ ) for the 130 proteins that only participate in F<sub>1</sub>-specific PPIs, the category of

‘lyase’ showed weak significance ( $P = 0.065$ ) in the analysis of the 493 proteins also participating in common PPIs. These four enriched functional classes indicate that F<sub>1</sub>-specific PPIs are potentially involved in metabolite biosynthesis pathways. Because enzymes involved in metabolite

biosynthesis are usually composed of multiple members of a gene family and mostly occur as gene clusters on chromosomes, we subsequently examined whether the proteins involved in  $F_1$ -specific PPIs tend to be localized in metabolic gene clusters (MGCs; Hen-Avivi *et al.*, 2016; Nutzmans *et al.*, 2016; Schlapfer *et al.*, 2017).

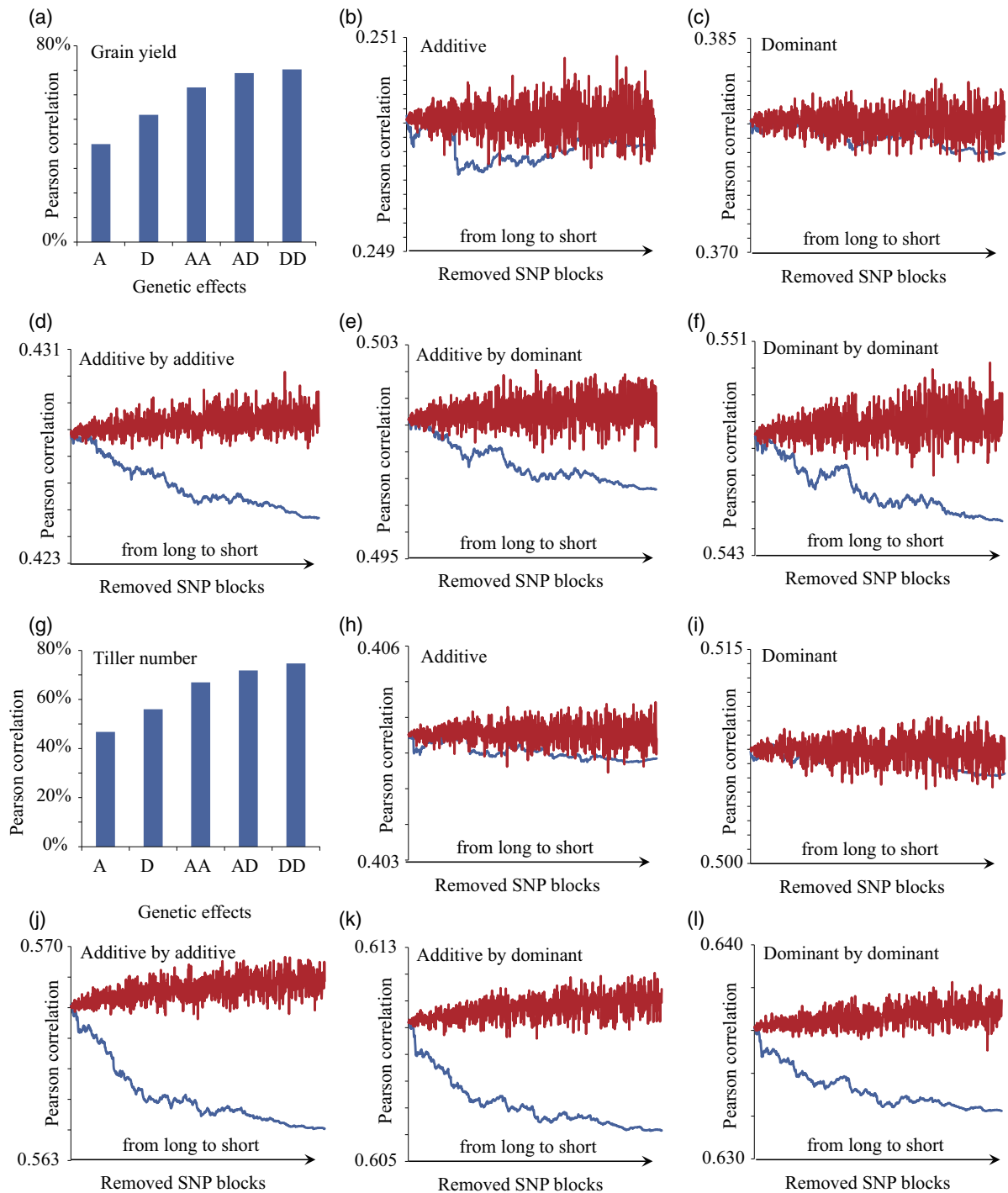
We used the software pipeline PlantClusterFinder to annotate MGCs in the *japonica* (cv. Nipponbare) and *indica* (cv. 9311) reference genomes. Next, proteins involved in  $F_1$ -specific PPIs, Nipponbare-specific PPIs, 9311-specific PPIs and common PPIs were mapped to the genomic regions harboring MGCs. On average, 15.1% of the proteins from the four groups matched with the genes in MGCs. When only the non-overlapping genes in each group were analyzed, the average percentage of genes that occurred in MGCs increased to 20.2% (Figure 3b). Because a region harboring MGCs may also contain non-metabolic genes, we also examined the matching rates of the genes in MGCs with metabolic genes. Among the 130 genes that only participate in  $F_1$ -specific PPIs, 96.2% encode metabolic enzymes (Figure 3c). These results indicate that proteins participating in  $F_1$ -specific PPIs, and especially those that only participate in  $F_1$ -specific PPIs, tend to be clustered and functionally involved in metabolic pathways.

*F<sub>1</sub>-specific PPIs accumulatively contribute to heterosis as epistasis effects.* Epistasis has been hypothesized as part of the genetic components of heterosis, which arises from a variety of inter-subspecies interactions due to the widely dispersed genetic variations between the two parental genomes. Although our computational simulation of the hybrid interactome has indicated the existence of  $F_1$ -specific PPIs, whether they contribute to heterosis or not requires further analysis. To test the genetic effects of  $F_1$ -specific PPIs in heterosis, we used a set of breeding data from a real-world hybrid rice GS-assisted breeding project conducted in Changsha, a provincial capital in southern China (Table S3). One of the unique features of this  $F_1$  population is that hybrids of these breeding lines frequently express superior heterosis performance. The lines include 43 maternal lines and 66 paternal lines that were crossed with a North Carolina II (NCII) design. This GS-assisted breeding project used 925  $F_1$  hybrids, accounting for one-third of all hybridization combinations, as a training population to derive a GS model, which was used to predict two yield-related traits, grain yield (GY) and tiller number (TN), for the remaining hybridization combinations. The genotypes of the 109 parental lines were profiled using Illumina Rice Genotyping Arrays containing 50 000 SNPs, and the genotypes of 925  $F_1$  hybrids were computationally inferred by combining the alleles from each parent.

To dissect the genetic components behind GY and TN traits, the GS model considered additive (A), dominant (D) and epistatic (E) effects, which were further decomposed into additive by additive ( $A \times A$ ) interactions, additive by dominant ( $A \times D$ ) interactions, and dominant by dominant ( $D \times D$ ) interactions. To test the overall prediction accuracy of the model, the five types of genetic effects were sequentially added in the following order: A, D,  $A \times A$ ,  $A \times D$  and  $D \times D$ . Fivefold cross-validation was used to compute an average Pearson correlation between the observed and predicted values. When the GS model considered only additive effects, the prediction accuracies for GY and TN were only 0.398 and 0.475, respectively (Figure 4a,g). When the GS model was adjusted to also consider D,  $A \times A$ ,  $A \times D$  and  $D \times D$  effects, the prediction accuracies gradually increased to 0.712 and 0.758 for GY and TN, respectively, as more effects were added. These results indicate that non-additive effects, especially epistatic effects, account for a considerable proportion of the genetic components of GY and TN, and thus superior heterosis performance is expressed in this hybrid population with high frequency.

We subsequently validated how  $F_1$ -specific PPIs function in heterosis for this population by excluding SNPs adjacent to genes in  $F_1$ -specific PPIs from the GS model and testing its prediction accuracy. If the prediction accuracy drops after SNPs are excluded, then the removed genes may contribute to the traits. Because SNPs may not be evenly distributed in the genome, and some of the genes in  $F_1$ -specific PPIs may be localized in clusters, we first identified genomic blocks with the potential to include more than one SNP using the procedure described in the Experimental Procedures. Next, SNPs were removed from the GS model in a block-by-block manner from the longest to the shortest. For the 1441 genes in 4612  $F_1$ -specific PPIs, we identified 998 blocks covering these genes, including 4004 SNPs. Every time a block was removed, a Pearson correlation was computed to indicate the model performance, until all 998 blocks were removed. For each removal, a set of SNPs equal to the number of SNPs in the removed block was randomly picked from the array for calculation of the Pearson correlation as contrasts.

When the GS model only took additive effects into account, removal of the SNP blocks covering the 1441 genes in  $F_1$ -specific PPIs did not reduce the prediction accuracy in comparison with that obtained using the GY and TN traits (Figure 4b,h). The same patterns were observed for the GS model that took only dominant effect into account (Figure 4c,i). For  $A \times A$ ,  $A \times D$  and  $D \times D$  epistatic effects, the gradual removal of SNP blocks caused a gradual decrease in the prediction accuracy for the GY and TN traits (Figure 4d–f, j–l). Therefore, these results support the assumption that the  $F_1$ -specific



**Figure 4.** F<sub>1</sub>-specific protein–protein interactions (PPIs) may contribute to heterosis via epistatic effects.

(a) The prediction accuracy of the grain yield trait gradually increases when the genomic selection (GS) model accumulatively takes additive effects, dominant effects and three types of epistatic effects into consideration. The highest prediction power was reached when all five types of genetic effects were considered. For the GS model only considering (b) additive or (c) dominant effects, removal of single nucleotide polymorphisms (SNPs) adjacent to genes involved in F<sub>1</sub>-specific PPIs did not reduce prediction accuracy. For the GS model only considering the three types of epistatic effects ((d) additive × additive, (e) additive × dominant and (f) dominant × dominant), removal of SNPs adjacent to genes involved in F<sub>1</sub>-specific PPIs reduced prediction accuracy. The red lines represent the control dataset as a contrast (generated by removing the same number of SNPs, which were randomly picked from the genome). The blue lines represent the actual dataset, which was generated by accumulatively removing SNPs adjacent to genes (within 5 kb) involved in F<sub>1</sub>-specific PPIs. SNPs adjacent to each other (within 5 kb) were merged as one block. The blocks containing SNPs were removed in order from the longest block to the shortest block, as indicated by the arrow on the X-axis.

(g–l) The same analysis was performed for the tiller number trait.



PPIs may play an epistatic role in enhancing the heterosis performance of  $F_1$  hybrids, in addition to the additive and dominant effects. Moreover, the contribution of  $F_1$ -specific PPIs to heterosis may act in an accumulative fashion, reflected by the gradual decrease of GS prediction accuracy when genes involved in  $F_1$ -specific PPIs were removed from the prediction model one by one.

## DISCUSSION

### Insights into heterosis mechanisms from simulated protein interactome

Heterosis has been investigated for more than a century from genetic, molecular and genomic aspects, but the underlying mechanisms remain elusive (Schnable and Springer, 2013). Expression of heterosis may be a result of any possible change occurred during transcriptional and posttranscriptional, translational and posttranslational processes, and physical interactions between any large or small molecules in a hybrid genome (Minvielle, 1987; Schnell and Cockerham, 1992; Sun and Kardina, 2010; Jiang *et al.*, 2017). In addition, although three main quantitative genetics hypotheses, dominance, overdominance and epistasis, have been proposed for decades, researchers have recently attempted to find connections between the genetic and molecular mechanisms behind heterosis. Especially for epistasis, which may involve different types of molecular interactions, any single or accumulative changes in the biological networks of hybrids may cause heterosis to different extents (Birchler *et al.*, 2006, 2010). The effects of dominance and overdominance have been revealed through studying genomic variation from the aspect of population genetics. However, the relationship between protein interactions and heterosis has been overlooked due to the complexity and difficulty in profiling a *bona fide* and comprehensive interactome in  $F_1$  hybrids. In addition, the contribution of individual  $F_1$ -specific PPI to heterosis may be weaker than the effects arising from genetic variation on major loci, but cumulative effects arising from the combination of dozens to thousands of  $F_1$ -specific PPIs with minor contributions may contribute to the yet undiscovered component of heterosis. Our study utilizes computational approaches to simulate the protein interactome in  $F_1$  hybrids together with GS methods to evaluate the contribution of PPIs to heterosis. Although this systems biology approach may involve a certain proportion of false positives that are difficult to estimate, the results pinpoint an interesting direction for deciphering complex heterosis mechanisms. Future experimentally generated PPI data in hybrid rice are expected to validate the conclusion drawn from the current study, providing further understanding of heterosis at the level of protein interactions.

*F<sub>1</sub>-specific PPIs may complement metabolic pathways to enhance F<sub>1</sub> fitness.* We simulated the protein interactomes of the  $F_1$  hybrid of *indica* ssp. (cv. 9311) × *japonica* ssp. (cv. Nipponbare) and its parents, which are elite inbred varieties of two rice subspecies with comprehensive reference genomes and annotations. A previous comparative analysis of the 9311 and Nipponbare genomes revealed extensive structural variations in addition to SNPs, including copy number variations, PAV, segmental and tandem duplications, and chromosomal inversion and translocations (Wang *et al.*, 2005; Yu *et al.*, 2005). These genomic variations may cause differential gene expression and molecular interactions between these two subspecies of rice, resulting in more diversified biological networks in their  $F_1$  hybrids in comparison with their own (Wang *et al.*, 2012). Comparison of the three sets of interactomes showed that common PPIs formed by the same pair of interacting proteins in 9311, Nipponbare and the  $F_1$  hybrid only accounted for 19.7% of all PPIs, while  $F_1$ -specific PPIs accounted for 20.5% of all PPIs (Figure 1a). This result indicates that the hybrid interactome is indeed diversified due to the combination of the two parental proteomes. In addition, the hybrid interactome includes novel PPIs, as each parent may supply novel interacting partners. In addition, nearly half of the  $F_1$ -specific PPIs (the sum of Type I, II and III PPIs) were identified because the interacting orthologs were missing from at least one parental genome, which indicated that PAV for protein-coding genes is the major source of  $F_1$ -specific PPIs (Figure 2b). Moreover, the remaining  $F_1$ -specific PPIs were due to missense variations caused by non-synonymous SNPs or indels on interaction interfaces or elsewhere on the protein sequences, which cause protein structural changes to form  $F_1$ -specific PPIs (Figure 2b). This result supports previous observations that PAVs are widespread among the genomes of maize inbred lines and probably contribute to heterosis through complementation of non-allelic paralogs in hybrid genomes (Eichten *et al.*, 2011; Hufford *et al.*, 2012; Schnable and Springer, 2013). With the newly available Mo17 reference genome (Sun *et al.*, 2018), a similar interactome comparison among B73, Mo17 and their  $F_1$  hybrid can be conducted to illustrate the role of  $F_1$ -specific PPIs in maize heterosis.

We also found that genes participating only in  $F_1$ -specific PPIs are functionally enriched in the categories of transferase, oxidoreductase, lyase and isomerase in comparison with common PPIs. These functional categories mostly consist of metabolic enzymes that are responsible for producing metabolites or small molecules to participate in various cellular functions necessary for survival, stress responses or environmental fitness. Because metabolite biosynthesis genes in plant genomes are usually organized in clusters, and some genes in  $F_1$ -specific PPIs are metabolic enzymes, we were inspired to examine the enrichment of  $F_1$ -specific

PPIs in MGCs (Hen-Avivi *et al.*, 2016; Nutzmann *et al.*, 2016; Schlapfer *et al.*, 2017). Our analysis showed that, in comparison with genes in common PPIs, genes in F<sub>1</sub>-specific PPIs tend to be located in MGCs. Taking the fact that metabolic genes are a group of rapidly evolving genes that proliferate via tandem duplication to diversify the metabolome to enhance environmental fitness (Ober, 2005; Hollister, 2015), metabolic genes specific to *japonica* and *indica* rice may be complemented by each other to form relatively more complete metabolic networks in F<sub>1</sub> hybrids in comparison with their parents. As a result, the environmental fitness of F<sub>1</sub> hybrids is enhanced, and this change manifests as heterosis in various traits. To a certain extent, our analysis of F<sub>1</sub>-specific PPIs supports the previous finding that heterosis may be a result of environmental fitness, as beneficial alleles from both parents favoring fitness are combined and complement deleterious alleles in the offspring genome (Chen, 2013; Li *et al.*, 2018).

*F<sub>1</sub>-specific PPIs contribute to heterosis in a quantitative, epistatic manner.* Our analysis showed that PAVs and missense variations are the primary causes of F<sub>1</sub>-specific PPIs and seem to be an effect of superior alleles complementing deleterious alleles (Springer *et al.*, 2009; Schnable and Springer, 2013). However, further evidence was required to determine whether the contributions of F<sub>1</sub>-specific PPIs to heterosis fit into the dominance, overdominance and/or epistasis models. To decompose the genetic effects of F<sub>1</sub>-specific PPIs, we applied GS-based trait prediction using an F<sub>1</sub> population consisting of 925 hybrids from a real-world breeding project based on a NCII mating design. In contrast with the other four selected populations, this population was constructed with modern inbred lines by a company to breed superior hybrid rice varieties for many years. Two interesting results were observed when SNP markers adjacent to genes in F<sub>1</sub>-specific PPIs were gradually removed from GS prediction models that separately considered additive, dominant and epistatic effects (Figure 4). First, removal of these SNPs did not reduce the prediction accuracy for the GS model when only additive and dominance effects were considered. In contrast, when the same group of SNPs was removed, the prediction accuracy of the GS model considering all three types of interactions in epistasis effects (additive × additive, additive × dominant, dominant × dominant) was clearly reduced. Second, gradual removal of SNPs produced a gradual decline in prediction accuracy within a range of 0.05–0.1 differences, suggesting that the epistatic effects of F<sub>1</sub>-specific PPIs act in a quantitative manner. Nevertheless, the simulated interactomes were based on PPI prediction using a stringent pipeline by combining Interolog and PRISM results using reference/conserved PPIs collected from other model species. Thus, the 4612 F<sub>1</sub>-specific PPIs *in silico* predicted from the 9311 and Nipponbare

reference genomes might only account for a small fraction of the actual number of PPIs. Therefore, the number of F<sub>1</sub>-specific PPIs and their contributions to heterosis are difficult to determine and may be underestimated. Based on these findings, we conclude that F<sub>1</sub>-specific PPIs likely play a minor epistatic role in contributing to heterosis in a quantitative manner.

## EXPERIMENTAL PROCEDURES

### Simulation of the F<sub>1</sub> hybrid interactome using Interolog

The Interolog was used to simulate the protein interactome of the F<sub>1</sub> hybrid of *indica* ssp. (cv. 9311) × *japonica* ssp. (cv. Nipponbare), based on orthologous PPIs (Garcia-Garcia *et al.*, 2012). *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Escherichia coli* and *Homo sapiens* with abundant experimentally validated PPIs were selected as reference organisms, whose protein sequences were obtained from the UniProt database (The UniProt Consortium, 2017). The protein sequences of Nipponbare and 9311 were obtained from the Ensembl Plants database (Bolser *et al.*, 2016). The Inparanoid algorithm (Remm *et al.*, 2001) with the BLOSUM80 substitution matrix was used to identify orthologs between Nipponbare/9311 and *A. thaliana*, whereas the BLOSUM62 between Nipponbare/9311 and *S. cerevisiae* (and *C. elegans*, *D. melanogaster* and *H. sapiens*), BLOSUM45 between Nipponbare/9311 and *E. coli*. The PPIs of the above six species were collected from the BioGRID (Chatr-Aryamontri *et al.*, 2017), DIP (Xenarios *et al.*, 2002), MINT (Licata *et al.*, 2012) and IntAct (Orchard *et al.*, 2014) databases. Additional PPIs of *H. sapiens* and *A. thaliana* were collected from the HPRD (Keshava Prasad *et al.*, 2009) and TAIR (Lamesch *et al.*, 2012) databases, respectively.

*Simulation of the F<sub>1</sub> hybrid interactome using PRISM.* To screen for high-confidence PPIs, PRISM was applied on Interolog-predicted PPIs, based on the rationale that if certain surface regions of two proteins resemble each side of a known interface of a template complex, the two proteins may interact via similar surface regions (Keskin *et al.*, 2008; Tuncbag *et al.*, 2011; Baspinar *et al.*, 2014). For each pair of interacting proteins predicted by Interolog, the corresponding protein complex structure was constructed. Two residues from a PPI were defined as the interface residues if the shortest distance between any of their atoms was less than 4 Å (Li *et al.*, 2016). All interface residues of a PPI constituted its interaction interface.

As the PRISM requires to know the 3D structures of two proteins, we first predicted 3D structures. BLAST was used to search the homologous templates of proteins in the PDB database for homology modeling. Three criteria were considered: (1) the alignments between proteins and templates had at least 30% sequence identity and covered at least 40% of the protein lengths; (2) X-ray structures as templates were preferred over NMR structures; (3) templates with resolutions below 5 Å were prioritized. Then, five models for each protein were produced using Modeller (Sali and Blundell, 1993) according to the best template. The model with the lowest DOPE score was regarded as the best 3D structure of the protein after truncating unaligned residues at the N and C termini (Mosca *et al.*, 2013).

*Reliability assessment of the predicted F<sub>1</sub> hybrid interactome.* We used subcellular localization data and Gene

Ontology (GO) annotations to assess the reliability of  $F_1$ -specific PPIs. The methods used for the assessment have been widely used in Interolog-based PPI predictions (He *et al.*, 2008; Gu *et al.*, 2011; Zhang *et al.*, 2016, 2017). First, we used the proteins in the predicted  $F_1$ -specific PPIs to generate 1000 random PPI networks by randomly rewiring the protein pairs. Then, we predicted the subcellular localizations by the MultiLoc2 software (Blum *et al.*, 2009) and obtained the GO terms from the Ensembl Plants database. Further, the number of co-localized PPIs and the GO similarities between  $F_1$ -specific PPIs and randomly generated PPIs were compared. We downloaded previously published RNA-seq dataset of *japonica* × *indica*  $F_1$  hybrids profiled in endosperm from NCBI GEO GSE113769 (Chen *et al.*, 2018). Based on GSE113769, the co-expression of *japonica* and *indica* alleles from the genes involved in  $F_1$ -specific PPIs can be verified by using the SNPs between *japonica* and *indica*, which can be used to differentiate allelic expression. In GSE113769, the samples were collected from the early developing endosperm tissues (7 days after pollination) of  $F_1$  hybrids. The  $F_1$  hybrids were generated from two distinct, reciprocal crosses of *japonica* and *indica* rice between Liuqianxin-A and Rongfeng-B ( $L \times R$ ), Rongfeng-A and Liuqianxin-B ( $R \times L$ ), Yu6-A and Wufeng-B ( $Y \times W$ ), and Wufeng-A and Yu6-B ( $W \times Y$ ). Liuqianxin and Yu6 are *japonica* rice, Rongfeng and Wufeng are *indica* rice, whose  $F_1$  combinations generate strong heterosis. In  $L \times R$  and  $R \times L$  crosses, a total of 62 092 SNPs identified from 12 295 endosperm-expressed (FPKM > 0) genes were used to distinguish *japonica* and *indica* alleles; and a total of 59 097 SNPs were identified from the 12 373 endosperm-expressed genes in  $Y \times W$  and  $W \times Y$  crosses. This result indicates that over 63% of the total 19 459 endosperm-expressed genes possess at least one SNP that can be used to differentiate the *japonica* and *indica* alleles. In the four sets of crosses, 11 823 (96.16%), 11 863 (96.49%), 11 805 (95.41%) and 11 647 (94.13%) genes were expressed (FPKM > 0) from both *japonica* and *indica* alleles. Examination of the expression of  $F_1$ -specific genes showed that 1001 (69.47%) and 988 (68.56%) out of 1441  $F_1$ -specific genes possessing at least one SNP were expressed in the endosperm of the two reciprocal crosses, which could be used to evaluate allelic expression in  $F_1$  hybrid.

**Identification of metabolic gene clusters.** The MGCs in the rice genome were identified *in silico* by four steps. First, Ensemble Enzyme Prediction Pipeline (E2P2) software was used to predict potential enzymes based on homology transfer (Chae *et al.*, 2014). Second, Pathway Tools software assigned reactions to these enzymes and identified potential metabolic pathways (Karp *et al.*, 2011). Third, protein paralogs were identified by all-against-all BLAST, and a Markov cluster algorithm (MCL, <https://micans.org/mcl/>). Fourth, DNA sequences, gene positions, Pathway Tools and MCL results were used as input for PlantClusterFinder software to identify genome-scale MGCs (Schlapfer *et al.*, 2017). Each MGC satisfied the following four criteria: (1) contained at least three metabolic genes; (2) contained at least two reactions; (3) all genes were located contiguously on the same chromosome; (4) was not composed solely of repetitive metabolic genes (Schlapfer *et al.*, 2017). PlantClusterFinder use multiple iterations to merge adjacent metabolic genes into clusters. Each cluster may contain non-metabolic genes in addition to metabolic genes.

**Trait prediction analysis using GS models.** The  $F_1$  population used for GS analysis included 925  $F_1$  hybrids from the crossing of 43 maternal lines and 66 paternal lines based on a NCII

design scheme from an actual rice breeding project, which was carried out by a breeding company in Changsha in 2015. The 109 parental lines were genotyped using 50K-SNP Illumina Arrays. GY and TN were the two yield-related traits used for the GS prediction. SOMMER software was used to estimate additive, dominant and epistatic effects, as well as to construct the best linear unbiased prediction (BLUP) model (Covarrubias-Pazarán, 2016). The model prediction accuracy was evaluated by fivefold cross-validation. First, 4/5 of the  $F_1$  hybrids were used as a training set to derive the BLUP model, after which the remaining 1/5 of the  $F_1$  hybrids were used as a test set for trait prediction. This procedure was repeated five times, and the average of the five Pearson correlation values was taken as the model accuracy.

To evaluate the effects of additive, dominant and epistatic components in correlation with  $F_1$ -specific PPIs, SNPs adjacent to 1441  $F_1$ -specific genes were removed from the GS model, after which the prediction accuracy was recalculated. In the first round, SNPs falling within 5 kb upstream and downstream of  $F_1$ -specific genes were merged as initial blocks. In the second round, SNPs falling within 2.5 kb upstream and downstream of the initial blocks were merged into additional blocks. In the third round, blocks adjacent to each other (within 2.5 kb upstream and downstream) were merged as the final blocks. Using this procedure, a total of 998 blocks containing 4004 SNPs were selected for the analysis. For each GS prediction, one block of SNPs was removed, while the same number of SNPs was also removed from the model as a contrast, until all 998 blocks were removed. This procedure was performed in an accumulative manner so that blocks of SNPs were removed in order from the longest block to the shortest block.

## ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation of China (31871706 and 31970645), the Department of Agriculture of Guangdong Province (2018-36), the National High Technology Research and Development Program of China (2014AA10A602), the National Key Research and Development Program of China (2016YFD0100801), the Open Research Fund Program (PRRD-2018-YB6) of the Beijing Key Lab of Plant Resource Research and Development, Beijing Technology and Business University, and the Priming Scientific Research Foundation of Hainan University (KYQD(ZR)1929).

## AUTHOR CONTRIBUTIONS

XW and ZZ designed and supervised the project; HL simulated protein interactomes, and performed the analyses of genetic variation and metabolic pathway; XWD, HH and ZL processed the phenotype and genotype data of hybrid population; SJ, CL, ZL and HH performed genomic selection and genetic effect analysis; HL, SJ, LL, ZL, HH and XWD drafted the manuscript; XW and ZZ revised the manuscript.

## CONFLICT OF INTEREST

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The authors confirm that the article's supporting data are available in the supporting tables. The corresponding authors can be contacted when requesting to access the data.



## SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

**Table S1.** The list of 130 proteins only present in F<sub>1</sub>-specific PPIs.

**Table S2.** Allelic expression statistics of endosperm-expressed genes involved in F<sub>1</sub>-specific PPIs.

**Table S3.** Breeding data from GS-assisted breeding project.

## REFERENCES

- Baranwal, V.K., Mikkilineni, V., Zehr, U.B., Tyagi, A.K. and Kapoor, S. (2012) Heterosis: emerging ideas about hybrid vigour. *J. Exp. Bot.* **63**, 6309–6314.
- Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O. and Gursoy, A. (2014) PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. *Nucleic Acids Res.* **42**, W285–289.
- Birchler, J.A., Yao, H. and Chudalayandi, S. (2006) Unraveling the genetic basis of hybrid vigor. *Proc. Natl. Acad. Sci. USA*, **103**, 12 957–12 958.
- Birchler, J.A., Yao, H., Chudalayandi, S., Vaiman, D. and Veitia, R.A. (2010) Heterosis. *Plant Cell*, **22**, 2105–2112.
- Blum, T., Briesemeister, S. and Kohlbacher, O. (2009) MultiLoc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinform.* **10**, 274.
- Bolser, D., Staines, D.M., Pritchard, E. and Kersey, P. (2016) Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods Mol. Biol.* **1374**, 115–140.
- Bruce, A.B. (1910) The Mendelian theory of heredity and the augmentation of vigor. *Science*, **32**, 627–628.
- Chae, L., Kim, T., Nilo-Poyanco, R. and Rhee, S.Y. (2014) Genomic signatures of specialized metabolism in plants. *Science*, **344**, 510–513.
- Chatr-Aryamontri, A., Oughtred, R., Boucher, L. et al. (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379.
- Chen, Z.J. (2013) Genomic and epigenetic insights into the molecular bases of heterosis. *Nat. Rev. Genet.* **14**, 471–482.
- Chen, C., Li, T., Zhu, S. et al. (2018) Characterization of imprinted genes in rice reveals conservation of regulation and imprinting with other plant species. *Plant Physiol.* **177**, 1754–1771.
- Covarrubias-Pazaran, G. (2016) Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS ONE*, **11**, e0156744.
- Crow, J.F. (1948) Alternative hypotheses of hybrid vigor. *Genetics*, **33**, 477–487.
- East, E.M. (1936) Heterosis. *Genetics*, **21**, 375–397.
- Eichten, S.R., Foerster, J.M., de Leon, N., Kai, Y., Yeh, C.T., Liu, S., Jeddeloh, J.A., Schnable, P.S., Kaepler, S.M. and Springer, N.M. (2011) B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. *Plant Physiol.* **156**, 1679–1690.
- Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J. and Oliva, B. (2012) BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic Acids Res.* **40**, W147–151.
- Gu, H., Zhu, P., Jiao, Y., Meng, Y. and Chen, M. (2011) PRIN: a predicted rice interactome network. *BMC Bioinform.* **12**, 161.
- He, F., Zhang, Y., Chen, H., Zhang, Z. and Peng, Y.L. (2008) The prediction of protein-protein interaction networks in rice blast fungus. *BMC Genom.* **9**, 519.
- Hen-Avivi, S., Savin, O., Racovita, R.C. et al. (2016) A metabolic gene cluster in the wheat W1 and the barley Cer-cqu loci determines beta-diketone biosynthesis and glaucousness. *Plant Cell*, **28**, 1440–1460.
- Hollister, J.D. (2015) Polyploidy: adaptation to the genomic environment. *New Phytol.* **205**, 1034–1039.
- Hua, J., Xing, Y., Wu, W., Xu, C., Sun, X., Yu, S. and Zhang, Q. (2003) Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. USA*, **100**, 2574–2579.
- Hufford, M.B., Xu, X., van Heerwaarden, J. et al. (2012) Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811.
- Hull, F.H. (1945) Recurrent selection for specific combining ability in corn. *Agron. J.* **37**, 134–145.
- Jiang, Y., Schmidt, R.H., Zhao, Y. and Reif, J.C. (2017) A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. *Nat. Genet.* **49**, 1741–1746.
- Jones, D.F. (1917) Dominance of linked factors as a means of accounting for heterosis. *Proc. Natl. Acad. Sci. USA*, **3**, 310–312.
- Karp, P.D., Latendresse, M. and Caspi, R. (2011) The pathway tools pathway prediction algorithm. *Stand Genomic Sci.* **5**, 424–429.
- Keshava Prasad, T.S., Goel, R., Kandasamy, K. et al. (2009) Human protein reference database–2009 update. *Nucleic Acids Res.* **37**, D767–772.
- Keskin, O., Nussinov, R. and Gursoy, A. (2008) PRISM: protein-protein interaction prediction by structural matching. *Methods Mol. Biol.* **484**, 505–521.
- Lai, J., Li, R., Xu, X. et al. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030.
- Lamesch, P., Berardini, T.Z., Li, D. et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–1210.
- Li, H., Yang, S., Wang, C., Zhou, Y. and Zhang, Z. (2016) AraPPISite: a database of fine-grained protein-protein interaction site annotations for *Arabidopsis thaliana*. *Plant Mol. Biol.* **92**, 105–116.
- Li, Z., Coffey, L., Garfin, J. et al. (2018) Genotype-by-environment interactions affecting heterosis in maize. *PLoS ONE*, **13**, e0191321.
- Licata, L., Briganti, L., Peluso, D. et al. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–861.
- Meyer, R.C., Witucka-Wall, H., Becher, M. et al. (2012) Heterosis manifestation during early Arabidopsis seedling development is characterized by intermediate gene expression and enhanced metabolic activity in the hybrids. *Plant J.* **71**, 669–683.
- Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T. and Thomas, P.D. (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* **44**, D336–D342.
- Minvielle, F. (1987) Dominance is not necessary for heterosis a two-locus model. *Genet Res.* **49**, 245–247.
- Mosca, R., Ceol, A. and Aloy, P. (2013) Interactome3D: adding structural details to protein networks. *Nat. Methods*, **10**, 47–53.
- Ng, D.W., Chen, H.H. and Chen, Z.J. (2017) Heterologous protein-DNA interactions lead to biased allelic expression of circadian clock genes in interspecific hybrids. *Sci. Rep.* **7**, 45 087.
- Nutzmann, H.W., Huang, A. and Osbourn, A. (2016) Plant metabolic clusters – from genetics to genomics. *New Phytol.* **211**, 771–789.
- Ober, D. (2005) Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends Plant Sci.* **10**, 444–449.
- Orchard, S., Ammari, M., Aranda, B., et al. (2014) The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**.
- Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052.
- Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
- Schlapfer, P., Zhang, P. and Wang, C. (2017) Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* **173**, 2041–2059.
- Schnable, P.S. and Springer, N.M. (2013) Progress toward understanding heterosis in crop plants. *Annu. Rev. Plant Biol.* **64**, 71–88.
- Schnell, F.W. and Cockerham, C.C. (1992) Multiplicative vs. arbitrary gene action in heterosis. *Genetics*, **131**, 461–469.
- Shull, G.H. (1908) The composition of a field of maize. *J. Hered.* **4**, 296–301.
- Solhaug, E.M., Ihinger, J., Jost, M., Gamboa, V., Marchant, B., Bradford, D., Doerge, R.W., Tyagi, A., Replogle, A. and Madlung, A. (2016) Environmental regulation of heterosis in the allopolyploid *Arabidopsis suecica*. *Plant Physiol.* **170**, 2251–2263.
- Springer, N.M., Ying, K., Fu, Y. et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734.
- Sun, Y.V. and Kardia, S.L. (2010) Identification of epistatic effects using a protein-protein interaction database. *Hum. Mol. Genet.* **19**, 4345–4352.
- Sun, S.L., Zhou, Y.S., Chen, J. et al. (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* **50**, 1289–1295.

- Tang, J., Yan, J., Ma, X., Teng, W., Wu, W., Dai, J., Dhillon, B.S., Melchinger, A.E. and Li, J. (2010) Dissection of the genetic basis of heterosis in an elite maize hybrid by QTL mapping in an immortalized F2 population. *Theor. Appl. Genet.* **120**, 333–340.
- The UniProt Consortium (2017) UniProt: the universal protein knowledge-base. *Nucleic Acids Res.* **45**, D158–D169.
- Tuncbag, N., Gursoy, A., Nussinov, R. and Keskin, O. (2011) Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nat. Protoc.* **6**, 1341–1354.
- Wang, X., Shi, X., Hao, B., Ge, S. and Luo, J. (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* **165**, 937–946.
- Wang, Y., Wang, X. and Paterson, A.H. (2012) Genome and gene duplications and gene expression divergence: a view from plants. *Ann. N. Y. Acad. Sci.* **1256**, 1–14.
- Wei, G., Tao, Y., Liu, G. et al. (2009) A transcriptomic analysis of superhybrid rice LYP9 and its parents. *Proc. Natl. Acad. Sci. USA*, **106**, 7695–7701.
- Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. and Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305.
- Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q. and Saghai Maroof, M.A. (1997) Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. USA*, **94**, 9226–9231.
- Yu, J., Wang, J., Lin, W. et al. (2005) The Genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38.
- Zhang, F., Liu, S., Li, L., Zuo, K., Zhao, L. and Zhang, L. (2016) Genome-wide inference of protein-protein interaction networks identifies crosstalk in abscisic acid signaling. *Plant Physiol.* **171**, 1511–1522.
- Zhang, K., Li, Y., Li, T., Li, Z.G., Hsiang, T., Zhang, Z. and Sun, W. (2017) Pathogenicity genes in *Ustilaginoidea virens* revealed by a predicted protein-protein interaction network. *J. Proteome Res.* **16**, 1193–1206.
- Zhou, G., Chen, Y., Yao, W., Zhang, C., Xie, W., Hua, J., Xing, Y., Xiao, J. and Zhang, Q. (2012) Genetic composition of yield heterosis in an elite rice hybrid. *Proc. Natl. Acad. Sci. USA*, **109**, 15847–15852.