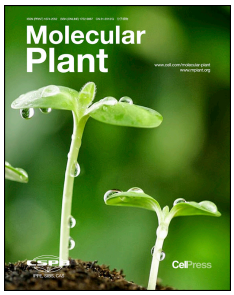
Journal Pre-proof



Dissection of genetic basis of 1,392 rice landraces and 770 hybrid combinations reveal great potential of rice landraces in hybrid rice improvement

Yinting Wang, Zixuan Li, Bi Wu, Xueqiang Wang, Wenyan Yang, Danjing Lou, Jinyue Ge, Ziran Liu, Wenlong Guo, Neng Zhao, Jun Yang, Weiya Fan, Kai Wang, Fei Li, Weihua Qiao, Hongbo Pang, Leina Zhou, Qingwen Yang, Chenwu Xu, Dingyang Yuan, Yang Xu, Qian Qian, Xiaoming Zheng

PII: S1674-2052(25)00354-5

DOI: https://doi.org/10.1016/j.molp.2025.10.004

Reference: MOLP 2004

To appear in: *Molecular Plant*

Received Date: 1 April 2025

Revised Date: 30 July 2025

Accepted Date: 8 October 2025

Please cite this article as: Wang Y., Li Z., Wu B., Wang X., Yang W., Lou D., Ge J., Liu Z., Guo W., Zhao N., Yang J., Fan W., Wang K., Li F., Qiao W., Pang H., Zhou L., Yang Q., Xu C., Yuan D., Xu Y., Qian Q., and Zheng X. (2025). Dissection of genetic basis of 1,392 rice landraces and 770 hybrid combinations reveal great potential of rice landraces in hybrid rice improvement. Molecular Plant doi: https://doi.org/10.1016/j.molp.2025.10.004.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2025 Published by Elsevier Inc. on behalf of CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, and Chinese Society for Plant Biology.

Journal Pre-proof

Research Article

Dissection of genetic basis of 1,392 rice landraces and 770 hybrid combinations reveal great potential of rice landraces in hybrid rice improvement

Yinting Wang[1,3†], Zixuan Li[1,2,3†], Bi Wu[3†], Xueqiang Wang[3†], Wenyan Yang[4†], Danjing Lou[1], Jinyue Ge[1], Ziran Liu[1,4], Wenlong Guo[1,3], Neng Zhao[3], Jun Yang[3], Weiya Fan[1], Kai Wang[1], Fei Li[1], Weihua Qiao[1], Hongbo Pang[5], Leina Zhou[1], Qingwen Yang[1], Chenwu Xu[4], Dingyang Yuan[6], Yang Xu[4*], Qian Qian[1,3*], Xiaoming Zheng[1,2,7*]

[1.] State Key Laboratory of Crop Gene Resources and Breeding/ Key laboratory Grain Crop Genetic Resources Evaluation and Utilization Ministry of Agriculture and Rural Affairs, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, 100081, China

[2.] National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya, 572000, China

[3.] Yazhouwan National Laboratory, Sanya, 572024, China

[4.] Key Laboratory of Plant Functional Genomics of the Ministry of Education, College of Agriculture, Yangzhou University, Yangzhou 225009, China

[5.] College of Life Science, Shenyang Normal University, Shenyang, 110034, China

[6.] State Key Laboratory of Hybrid Rice, Hunan Hybrid Rice Research Center, Hunan Academy of Agricultural Sciences, 410125 Changsha, China

[7.] International Rice Research Institute, Metro Manila 1301, Philippines

[†]These authors contributed equally to this work

[*]Correspondence: Xiaoming Zheng (zhengxiaoming@caas.cn); Qian Qian (qianqian188@hotmail.com); Yang Xu (xuyang_89@126.com)

Short summary:

Based on a large-scale analysis of 2,088 germplasms and 770 hybrid rice combinations, this study demonstrates that landraces significantly contribute to heterosis and yield-related traits. This research identified 105 novel QTLs and validated the role of *OsGRW5.1* in regulating grain width and weight through gene editing. A predictive platform of phenotypic performance was developed for potential hybrid combinations,

providing valuable resources for future rice breeding.

Abstract

Hybrid rice has made significant contributions to global food security. However, the efficient utilization of landraces to further enhance heterosis remains a key challenge in rice breeding. In this study, we collected a set of 1,392 landraces and 696 hybrid rice parental lines. A total of 770 hybrid combinations were constructed by crossing of 517 accessions selected from 2,088 rice accessions and seven key yield-related traits were collected. Nearly 500,000 potential hybrid combinations were predicted, comprehensive analysis revealed that landraces from South Asia played a significant role in improving multiple traits. Further investigation revealed substantial variation in the landraces contribution to the optimal hybrid combinations for various traits, with landraces particularly contributing to improvements in grain width. We identified 171 QTLs for seven traits using three association analysis methods. Of these QTLs, 77 known genes were identified within 66 QTLs and 105 novel QTLs without any known genes nearing to them. Gene editing based on CRISPR/Cas9 method showed that *OsGRW5.1* plays a critical role in regulating grain width and grain weight. Additionally, a strong correlation was also observed between advantageous haplotypes accumulating and phenotypic performance enhancing in our findings. More than 45% of the advantageous haplotypes derived from landraces are likely to play a significant role in future hybrid rice improvement. A predictive platform was developed that can output all the seven phenotypes of potential hybrid combinations with a given genotype of both parents. This research provides valuable data and practical insights for enhancing heterosis through the efficient utilization of landraces.

Key words: Heterosis; Landraces; Germplasm diversity; Genomic selection; Hybrid rice

Journal Pre-proof

Over the past fifty years, the widespread adoption of hybrid rice has greatly enhanced food security and increased farmers' incomes in China. Heterosis (hybrid vigor) has been widely applied in staple crops such as rice and maize because of its proven ability to boost yield and improve adaptability (Cheng et al., 2007). However, a significant challenge persists in effectively and efficiently utilizing the diversity of landraces to further amplify heterosis (Melchinger, 1999), which continues to limit progress in crop improvement (Hochholdinger and Yu, 2025).

Landraces provide a valuable genetic pool for rice improvement, offering a wide range of genetic diversity that enhances yield, strengthens disease resistance, improves quality, and supports adaptation to climate change (Song et al., 2024; Zheng et al., 2024). The historical progress of hybrid rice has relied heavily on these diverse landraces. For example, the development of cytoplasmic male sterility (Luo et al., 2013) and the introduction of restorer genes from Southeast Asian germplasm make the utilization of heterosis in practicable (Zhao et al., 2023). The discovery of photo-thermosensitive genic male sterility (Ding et al., 2012; Fan et al., 2016; Zhou et al., 2012b; Zhou et al., 2014) have significantly expanded the range of hybrid rice combinations by utilizing broader genetic backgrounds. For instance, the introduction of the *Xa21* gene has greatly enhanced rice disease resistance (Song et al., 1995), while the *Gn1a* gene has significantly increased yield per unit area (Ashikari et al., 2005). These innovations have improved rice adaptability to high disease pressure and high-density planting conditions, laying the foundation for future high-yield, resilient varieties (Zhang et al., 2021).

However, in the current breeding system, most breeders tend to improve existing parental lines that are high-performing, stable, and easily hybridized (Khan et al., 2015; Wu et al., 2018). While this approach has improved breeding efficiency in the short term, it has also led to the underutilization of potentially valuable landraces and wild relatives gene resources. Although a significant exponential relationship persists between hybrid cultivar numbers and total cultivar counts across ten major hybrid rice-producing provinces, the cultivar diversity index has approached saturation thresholds (Huang, 2022). Over the course of breeding, the emphasis on traits such as high yield,

disease resistance, and stress tolerance has caused the parent pool to converge, with many parents sharing highly similar genetic backgrounds (Chen et al., 2021). The cultivation area of currently major rice varieties in China (https://www.natesc.org.cn/) reveal that the genetic similarity among current parents had exceeded 90%, significantly constraining the potential for enhancing heterosis in hybrid breeding. Although broadening the genetic base is widely recognized as a key strategy for enhancing hybrid rice (Gao et al., 2022), the low utilization rate of landraces and the substantial workload involved in hybridization experiments remain significant challenges in the breeding process.

In recent years, modern computational techniques, particularly the application of machine learning, have enabled rapid prediction and optimization of hybrid combinations (Gu et al., 2023). However, the prediction still requires hybrid combinations based on diverse landraces as its foundation (Martini et al., 2021). This study collected 2,088 rice accessions and carefully selected parental lines for hybridization based on their genetic background, geographical distribution, and agronomic traits, resulting in the construction of 770 hybrid combinations. We conducted in-depth analyses of seven key yield agronomic traits and generated a dataset of nearly 500,000 potential hybrid combinations using computational techniques. By analyzing these combinations, we identified distinct patterns of heterosis within the landraces that could potentially enhance parental performance. Furthermore, we developed a hybrid combination prediction platform capable of forecasting the phenotypic performance of potential hybrid combinations based on genotype data. This study provides valuable data support and practical insights for improving heterosis in hybrid rice through the efficient utilization of landraces.

Journal Pre-proof

Results

Genetic Diversity Differences Between Rice Germplasm and Hybrid Rice Parents

To enhance the efficient utilization of landraces in hybrid rice breeding, we sequenced 946 accessions landraces and integrated 1,142 publicly available accessions, resulting in a total of 2,088 rice accessions. This collection includes 1,392 landraces from 18 major rice-producing countries across Asia, representing six global subgroups (Supplemental Table 1). Additionally, it comprises 321 sterile lines and 375 restorer lines, collected since the 1980s. These materials reflect key hybrid rice-growing regions in China, capturing significant ecological and agricultural diversity (Supplemental Table 1), spanning key periods during which human domestication and breeding have evolved substantially.

A total of 11.5 Tb of clean reads were obtained, leading to the identification of 1,365,785 high-quality SNPs following a series of stringent filtering steps. Phylogenetic relationships and principal component analysis (Figure 1A and Supplemental Figure 1A) revealed that 2,088 rice accessions could be divided into six subgroups: 149 *AUS*, 744 *Indica1* (*IND1*), 626 *Indica2* (*IND2*), and 297 *Indica3* (*IND3*), all of which belong to *Oryza sativa* ssp. *indica*; 107 *Tropical japonica* (*TRJ*) and 123 *Temperate japonica* (*TEJ*), all of which belong to *O. sativa* ssp. *japonica*. And it includes 42 intermediate accessions. Among them, the restorer lines were mainly concentrated in *IND1* (345 accessions), with a small number distributed in *IND2* (22 accessions). The sterile lines were predominantly found in *IND2* (278 accessions), with only a limited number observed in *IND1* (34 accessions; Supplemental Table1).

To gain a deeper understanding of the genomic differences between hybrid rice parents and landraces, we compared the differentiated SNPs (Figure 1B) and the nucleotide diversity (π; Figure 1C) among them. The results showed that the most significant SNP differentiated were observed between *TEJ* and the hybrid rice parent (*IND1*), with 1,176,054 differentiated SNPs. Similarly, *TEJ* also had 1,134,240 differentiated SNPs compared to the hybrid rice parent (*IND2*). Following this, the differences between *TRJ* and *IND1* and *IND2* were also significant, with 816,897 *TRJ-IND1* differentiated SNPs and 780,526 *TRJ-IND2* differentiated SNPs. In contrast, the

differences between *AUS* and *IND1* were 399,600 differentiated SNPs, and between *AUS* and *IND2*, 386,886 differentiated SNPs. The smallest differences were found between *IND3* and the sterile and restorer lines of *IND1* and *IND*2, with 22,457 and 23,488 unique SNPs, respectively. GO functional enrichment analysis revealed that genes associated with above SNPs were enriched in key biological processes, such as protein synthesis, metabolic regulation, and energy management (Supplemental Figure 1B). These processes may influence rice growth, development, and stress resistance. We further calculated π values for the landrace subgroups and the sterile and restorer lines. The average π value of landraces in the *IND1* subgroup is 0.001300, compared to 0.001235 for restorers and sterile lines. In *IND2*, the landraces π value is 0.001470, while restorers and sterile lines have a π value of 0.001285. The average π values for *IND3*, *AUS*, *TEJ*, and *TRJ* are 0.003226, 0.003000, 0.000991, and 0.002315, respectively, indicating that landraces exhibited higher genetic diversity. These findings suggest that landraces may provide potential genetic variation and resources for hybrid rice improvement.

Prediction and Heterosis Analysis of Hybrid Combinations Based on Multi-Trait for Landraces

We selected 517 accessions (Supplemental Table 2) from a collection of 2,088 rice accessions as parental lines for subsequent hybrids, including 298 *indica* (*IND*) rice, 92 *AUS*, 72 *Temperate japonica* (*TEJ*), 44 *Tropical japonica* (*TRJ*), and 11 intermediate (IM) rice. These parental lines represent a wide geographic distribution across all target countries and encompass over 99.97% of the diversity in agronomic traits. We successfully generated 770 F_1 hybrids including 335 *IND-IND*, 92 *IND-TEJ*, 60 *IND-TRJ*, 150 *IND-AUS*, 15 *TEJ-AUS*, 16 *TRJ-AUS*, 48 *AUS-AUS*, 15 *TEJ-TEJ*, 6 *TEJ-TRJ* and 1 *TRJ-TRJ* crosses and 32 hybrids based on IM rice (Supplemental Table 2) developed from these selected parental lines and collected 176 commercially cultivated hybrids. These hybrids, spanning multiple countries and rice subgroups, offer a solid foundation for systematically analyzing the agronomic traits of hybrids from diverse landraces for the first time. We collected data on seven key traits from these hybrids in Hainan, including flag leaf width (FLW), grain length (GL), grain width (GW), heading

date (HD), plant height (PH), panicle length (PL), and thousand grain weight (TGW).

To gain a more comprehensive understanding of the agronomic traits of hybrid combinations derived from 2,088 rice accessions, we developed our predictive models based on existing hybrids. Therefore, the simplified linear model, which is usually used to measure the additive effect, is chosen as the prediction method in this study. We conducted a simulated comparison of three main types of simplified linear models (Xu et al., 2016; Xu et al., 2014), including parametric models (e.g., GBLUP, LASSO, EN, and BayesB), semi-parametric models (e.g., RKHS), and machine learning models (e.g., SVM, LightGBM, and XGBoost). The analysis results showed that the prediction data based on GBLUP exhibited the most robust performance across all seven traits (from 0.51 to 0.63; Figure 2A, Supplemental Figure 2B and C) compared to the other seven methods. Based on these results, we concluded that GBLUP outperformed than other models in multi-trait prediction which is suitable as the final prediction tool.

To optimize the accuracy and efficiency of the GBLUP method for rice phenotypic prediction, we assessed the effects of training population size and hybrid combinations. Our results revealed that prediction accuracy improved and stabilized when the training population exceeded 700 hybrids (Figure 2B, Supplemental Table 3-4), ensuring reliable predictions for HD (0.517), FLW (0.278), TGW (0.324), GW (0.611), PH (0.554), PL (0.425), and GL (0.471). Further analysis the effects of genetic background showed that hybrids with both parents present achieved the highest prediction accuracy, such as HD (0.640), PH (0.733), and GW (0.633), while single-parent hybrids demonstrated reduced accuracy, including TGW (0.301) and PL (0.235) indicating the impact of missing genetic information on certain traits (Figure 2C). However, it is noteworthy that hybrids without parental involvement still maintained relatively high accuracy for specific traits, such as HD (0.851), GW (0.992), and GL (0.660), highlighting the robustness of the GBLUP method (Figure 2C). our findings confirmed that an appropriately size of training population is crucial for ensuring GBLUP's accuracy, while genetic background plays a significant role in prediction performance. Even without parental data, GBLUP remains effective, underscoring its broad applicability in genomic selection breeding.

To assess the agronomic performance of all 446,985 hybrid combinations, we calculated the positive mid-parent heterosis (MPH) values for each predicted combination across multiple phenotypic traits (Figure 2D). A total of 125 varieties exhibited positive MPH in more than 40% of combinations for five traits. Additionally, 52 varieties showed positive MPH in more than 60% of combinations for five traits. Only two varieties demonstrated positive MPH in more than 40% of combinations for six traits. Notably, no variety exhibited positive MPH in more than 60% of six traits (Figure 2E). We then conducted an in-depth analysis of the geographical and genetic characteristics of the 52 superior varieties. The results revealed that 38 of these varieties primarily originated from South Asia, 6 of these varieties from East Asia, and 8 of these varieties from South East Asia (Figure 2F) and were classified into four major groups (Figure 2G): *IND* (21 varieties), *AUS* (11 varieties), *TEJ* (4 varieties), and *TRJ* (16 varieties). Further analysis of the distribution of hybrid combinations showed that these superior hybrids were predominantly concentrated in South Asia/East Asia (37.6%), South Asia/South Asia (27.6%), and South Asia/Southeast Asia (18.5%) (Figure 2H). Additionally, the most common hybrid types among these combinations were *IND*/*TRJ* (22.9%), *IND*/*IND* (22.6%), and *IND*/*AUS* (18.6%) (Figure 2I).

The phenotypes of a total of 2,088 × 2,088 hybrid combinations were predicted based on 770 hybrids. Analysis of the top 100 hybrid combinations across six traits revealed the contributions of landraces in hybrids combinations (Supplemental Table 5). For GW and PL, all top-performing combinations (100%) had at least one parent derived from landraces. In contrast, 59% of the top combinations for PH included at least one landrace parent (Figure 2J). The presence of landraces was lower in the top combinations for HD and TGW, at 38% and 35%, respectively. Notably, landraces were completely absent from the top-performing combinations for GL (Figure 2J). These findings suggest that landraces contribute differently to various traits, playing a particularly important role in GW and PL improvement for future breeding programs (Figure 2K-H).

To ensure users easily utilizing the predictive platform of our constructed models, we developed a user-friendly online platform within Predicting Rice Germplasm

Hybrid Phenotype (PRGHP) web platform (https://hybrice.cn/). Anyone can access the nine predictive traits by selecting the parent lines. If users wish to predict their own materials using our models, they can process their genotype data according to the pipeline described in the support webpage, upload the genotype file, and download for the results once the task is complete.

GWAS uncovers potential genes associated with heterosis

Heterosis is generated by dominance effects, overdominance effects, and epistasis. To gain a more comprehensive understanding of the genomic basis of heterosis in rice, we employed three methods to analyze the genetic basis influencing seven key agronomic traits in rice. First, a mixed linear model (MLM) incorporating Q+K covariates was used for genome-wide association studies (GWAS) on 946 landraces, focusing on additive and dominance effects, identifying 8,842 SNPs and 86 QTLs significantly associated with the phenotype ($-\log_{10}(P) > 5$). Second, we applied $F_{1Genotype01}$ GWAS to 770 F_1 populations, encoding SNPs as either 0 or 1 to identify loci associated with dominant effects (Tan and Ingvarsson, 2022). A total of 5,443 SNPs and 62 QTLs were identified that were significantly associated with phenotype. Lastly, we conducted $F_{1Genotype012}$ GWAS on 770 F_1 populations, treating the number of minor alleles (0/1/2) as a continuous variable to assess additive effect loci, assuming each additional minor allele has the same genetic impact, thereby revealing the independent contributions of parental alleles and heterozygous alleles the influence of additive effects on phenotypes, identifying 3,657 SNPs and 51 QTLs significantly associated with seven traits. In total, 171 QTLs were detected across the three methods, of these, 77 known genes located within 66 QTLs, and 105 new QTLs were discovered (Figure 3A and B, Supplemental Table 6-8). We compared the haplotypes of those 77 known genes revealed that the causal variants in 19 genes matched those reported in the previously study (Wei et al., 2021).

***OsGRW5.1* positively regulates grain width and weight in rice**

Grain width is a key quantitative trait that influences thousand-grain weight, which is one of the three major factors determining rice yield. Comparative analysis of QTLs for grain width and thousand-grain weight revealed that significant loci for grain width

Journal Pre-proof

(*qGW5.1*, 130 Kb) and thousand-grain weight (*qTGW5.1*, 132 Kb) were co-localized in the same region, suggesting that they may represent the same QTL (Figure 3A). Gene function annotation and haplotype-phenotype association analysis of the 24 genes within the *qGW5.1* region identified *LOC_Os05g01710*, named *OsGRW5.1* (Grain-related Width and Weight 5.1; Figure 4A-B), as being functionally related to both grain width and grain weight, with its haplotype also significantly associated with these two traits (Supplemental Table 6). This gene encodes a transcription initiation factor IIAγ (OsTFIIAγ5) containing four helix-bundle domains and three β-sheet domains, which suggested to be a cofactor that required for transcription by RNA polymerase II (Figure 4C) (Iyer-Pascuzzi and McCouch, 2007). We identified 24 SNPs in the promoter (2 Kb) and coding regions of *OsGRW5.1*, which can be categorized into 7 major haplotypes (Figure 4D and Supplemental Table 9). Haplotype network analysis revealed that these haplotypes can be grouped into four distinct clusters. Group A, consisting of Hap1 and Hap2, primarily includes *TEJ* and *TRJ* landraces and shows the largest grain width (3.35 mm) and highest thousand-grain weight (26.15 g) compared to other groups. Group C, composed of *AUS* and *IND* landraces, exhibits the narrowest grain width (3.05 mm) and lowest thousand-grain weight (25.70 g). Group D consists solely of Hap3, mainly found in *AUS*, with a slender grain shape (3.09 mm) and lower thousand-grain weight (22.80 g). Group B, composed exclusively of *IND*, shows a similar grain width (2.92 mm) to Group D, but with a thousand-grain weight similar to Group A (22.55 g) (Figure 4E-G). To further validate the function of *OsGRW5.1* in regulating grain width and thousand-grain weight, we used CRISPR/Cas9 technology to knock out *OsGRW5.1* and generated two independent knockout transgenic lines (Figure 4H). The results showed that the grain width (3.35 mm) and thousand-grain weight (20.04 g) of *OsGRW5.1* knock out lines were significantly lower than those of the wild-type Nipponbare variety (3.44 mm and 21.39 g) (Figure 4I-M), confirming the role of *OsGRW5.1* in regulating rice grain width and its potential to improve both rice appearance quality and yield.

Hybrid vigor analysis of genetic loci

To evaluate the dominance effect size of the loci, the heterotic loci and the dominance/overdominance effects were analyzed. We evaluated the effects of

heterozygous loci for the peak SNPs above the suggestive *P* value (FLW, GL, GW, HD, PH, PL, and TGW) in Sanya. 1,361 loci were used for final analysis. We found that most loci (81%) showed incomplete dominance effects in F_1 GWAS. There were 340 loci with strong overdominance effects (86 with positive effect and 254 with negative effect (Figure 5A, Supplemental Figure 3A). To determine whether the accumulation of dominant heterozygous genotypes in the 770 F_1 hybrids is associated with phenotypic enhancement, we performed correlation analysis between the genotypes of significant loci and their corresponding phenotypes. First, dominant heterozygous genotypes were identified among the significant loci, and then the correlation between the number of dominant heterozygous genotypes and phenotypic improvement was calculated. The results show that the accumulation of dominant heterozygous genotypes from both $F_{1Genotype01}$ and $F_{1Genotype012}$ GWAS datasets is correlated with improved phenotypic performance (Figure 5B, Supplemental Figure 3B). Notably, heading date exhibited the highest correlation in both GWAS results (0.40 and 0.33), while flag leaf width showed the lowest correlation in both datasets (0.10 and 0.10). Overall, the accumulation of dominant heterozygous genotypes was strongly correlated with phenotype (Figure 5B, Supplemental Figure 3B), which can enhance phenotypic expression and contribute to hybrid vigor. Additionally, we performed haplotype analysis to further investigate the correlation between gene haplotypes and phenotypes. These results show that the effects of favorable haplotypes accumulation are better than the accumulation of dominant heterozygous locus (Supplemental Figure 3C).

Divergence of favorable alleles accumulation in hybrids parents and landraces

Since rice is a self-pollinating crop, during the process of artificial selection, certain favorable gene combinations become fixed across a large number of materials, resulting in the same beneficial genes appearing in the parental lines. This leads to a limitation in further enhancing the hybrid heterosis. A comprehensive haplotype analysis was performed on 120 candidate genes that previously identified through GWAS in 946 landraces. Among them, 12 are associated with FLW, 14 with GL, 22 with GW, 14 with HD, 15 with PH, 23 with PL, and 20 with TGW. The results showed that 37 candidate genes (30.8%) with superior haplotypes were highly prevalent, present at a frequency

of 80-100%, while 24 candidate genes (20.0%) with superior haplotypes demonstrated substantial presence at 60-80% frequency. Moderate frequency distribution (40-60%) of superior haplotypes was observed in 8 candidate genes, whereas 15 candidate genes (12.5%) show relative low frequency (20-40%). Notably, 36 candidate genes were identified as rare superior haplotypes that present a frequency below 20%. However, the haplotype analysis of 368 restorer lines and 312 maintainer lines showed a distinct distribution pattern compared to the landrace accessions. Among them, approximately half of the candidate genes with superior haplotypes were primarily distributed within a frequency range of 80-100% in the 368 restorer lines (46.7%) and 312 maintainer lines (42.5%). Additionally, candidate genes with superior haplotypes in the frequency range of 80%-20% were relatively lower distributed in the 368 restorer lines (5%) and 312 maintainer lines (18.3%). A total of 58 candidate genes (48.3%) with superior haplotypes were unique to the germplasm resources, mainly distributed on chromosomes 5 and 8 (Figure 5C), and had a high proportion in 52 elite parental lines. Among them, five parental lines had a specific superior haplotype distribution frequency of 0-20%, 29 parental lines had a frequency of 20-40%, and 18 parental lines had a frequency of 40-60% (Figure 5D). These 58 superior haplotypes were primarily derived from different subpopulations: 3 from *AUS*, 9 from *IND*, 24 from *TEJ*, and 20 from *TRJ* (Figure 5E). However, they were not fixed in the 368 restorer lines and 312 maintainer lines.

Discussion

Heterosis, or hybrid vigor, remains a century-old challenge in crop genetics. While numerous studies have supported three major hypotheses-dominance, overdominance, and epistasis (Hua et al., 2003; Huang et al., 2016; Zhou et al., 2012a). Most findings remain at the QTL level, lacking gene-level resolution. In rice, heterosis manifests as a complex organism-wide phenomenon, evident as early as syncytium formation during embryogenesis, influencing gene expression, cell size, and metabolic efficiency (Gu and Han, 2024; Gu *et al.*, 2023; Huang et al., 2015; Jahnke et al., 2010). In this study, we analyzed 946 landraces and 770 F_1 hybrids using GWAS across seven agronomic traits. A total of 171 QTLs were identified (86 in landraces, 62 in $F_{1Genotype012}$, and 51 in

$F_{1Genotype01}$), with 25 QTLs shared across populations. Within these QTL intervals, we identified 66 QTLs co-localized with 77 known genes. Notably, *OsGRW5.1*, located on the short arm of chromosome 5, was validated as a key gene influencing grain width and thousand-grain weight through CRISPR/Cas9 analyses. Particularly, *OsGRW5.1* containing a nonsynonymous variant (T/A, 437,499 bp) in the second exon region (Figure 4D). Further haplotype analysis revealed that Group A and Group B showing strong yield-enhancing effects (Figure 4F-M). Additional investigations, such as genetic complementation assays, the base transversion editing A-to-T could further validate its function in improving grain yield.

Predictive ability is influenced by marker density, training population size, the relationship of training population and the testing sample, heritability, the linkage disequilibrium (LD) of markers and QTL, ranged from 0.15 to 0.85 with different population and phenotype (Voss-Fels et al., 2019; Xu et al., 2021). In this study, we generated a hybrid panel from *IND*, *JAP*, and *AUS* lines and predicted the performance of 2,088 possible hybrid combinations using GBLUP. Prediction accuracies ranged from 0.278 to 0.611. Among the top 100 predicted combinations for grain yield traits, all contained at least one landrace parent, confirming their potential in hybrid improvement. Pan-genome analysis has shown that variable genes, rather than core genes, are critical for crop improvement (Qin et al., 2021). We found that hybrid genomes contain more variable genes than either parent, and their complementary may underpin heterosis (Hochholdinger and Yu, 2025). However, modern parental lines often show fixation of elite alleles, limiting diversity. Analysis of 120 genes significantly associated with local varieties GWAS showed that about 40% of the dominant haplotypes were found in the 696 hybrid rice parent materials, but more than 45% of the dominant haplotypes were endemic to landraces. These findings highlight the untapped value of landrace alleles for improving hybrid vigor.

Genetic dissection of key agronomic traits in rice not only advances the theoretical understanding of the molecular mechanisms underlying yield but also provides practical guidance for breeding programs (Fukuoka et al., 2010). Heterosis leverages diverse variation to enhance not only grain yield but also disease resistance, stress

tolerance, and environmental adaptability. Through whole-genome association analysis, heterosis-related loci can be identified and subsequently utilized via marker-assisted selection, gene editing, and the accumulation of superior haplotypes containing additive-effect QTLs (Liu et al., 2022). These approaches collectively contribute to the improvement of major traits, the enhancement of parental lines, and the acceleration of hybrid breeding cycles. Furthermore, the genetic analysis and predictive modeling of heterosis can facilitate the efficient selection of high-performing hybrid combinations, boosting both yield and grain quality. By integrating genome selection and landrace-derived haplotypes into breeding programs, we can use materials with superior haplotypes improving these male sterile and restorer lines and design new parental combinations with greater heterotic potential, offering a promising strategy to enhance yield and maintain food security under environmental and demographic pressures (Figure 6).

Journal Pre-proof

Materials and Methods

Plant Materials

The 2,088 rice accessions used in this study include 1,392 samples from 18 major rice-producing countries in Asia, representing local varieties from six rice populations worldwide (Supplemental Table 1). Among these, 836 samples were introduced from the International Rice Research Institute (IRRI), sourced from regions such as East Asia, South Asia, and Southeast Asia, primarily consisting of *indica* and *japonica* rice varieties, as well as Aus and Basmati varieties. Additionally, the collection includes 321 cytoplasmic male sterility (CMS) lines and 375 restorer lines, directly collected since the 1980s. These materials include 211 CMS lines, 110 gametocidal male sterility (GMS) lines, 294 three-line hybrid rice (3R) varieties, and 81 two-line hybrid rice (2R) varieties, totaling 696 samples. These materials represent the parental lines of most indica hybrid rice varieties widely cultivated in southern China over the past 50 years. To fully represent the comprehensively diversity of 2,088 local rice varieties, we selected accessions for hybrid combinations based on three key factors: geographical distribution, genetic diversity, and phenotypic variation. As a result, 517 representative accessions were selected for crossing. The selection process was as follows: Geographical distribution: The 517 selected lines were proportionally sampled from major rice-growing regions covered by the 2,088 local varieties. This ensured broad ecological and regional representation among the chosen parents. Genetic diversity: To assess how well the selected lines represented the overall genetic variation, we calculated the Shannon-Wiener diversity index based on seven key agronomic traits. First, the mean and standard deviation of each trait were calculated across all samples. Each trait was then categorized into three levels (low, medium, high) based on its distribution. Diversity indices were computed for each trait and averaged to obtain an overall diversity value. The results showed that the selected 517 lines retained 99.97% of the total phenotypic diversity observed in the 2,088 local varieties. Phenotypic traits: Synchronization of flowering time was also prioritized during selection to facilitate effective field crossing. This approach enabled efficient random mating in the field, ultimately resulting in the development of 770 F_1 hybrid combinations.

Plant materials and phenotyping.

Phenotyping of local landraces: In the winter seasons of 2018 and 2019, we conducted systematic phenotyping of 946 rice landrace accessions in Lingshui, Hainan (18.25°N, 109.51°E). Crossing procedure: In the summer of 2020, the 946 landrace accessions were planted at the Guangxi Academy of Agricultural Sciences to facilitate large-scale crossing and ensure crossing efficiency and success rates. During crossing, hot water emasculation was used to inactivate the pollen viability of the female plants. Once the anthers extruded from the lemma and palea, manual removal was conducted promptly, and the panicles were immediately bagged to prevent contamination from external pollen during the crossing process. This method effectively reduced the difficulty and labor intensity of manual emasculation while enabling the acquisition of sufficient F_1 seeds for subsequent phenotyping. Verification of hybridization: After obtaining the F_1 seeds, we conducted molecular marker genotyping on both the parents and the F_1 hybrids to verify the success of the hybridization. Only those hybrids confirmed to be successful were used for subsequent phenotyping, ensuring the accuracy and reliability of the data. Phenotyping of F_1 hybrids: In the winter of 2020, 770 verified F_1 hybrids were planted in Lingshui, Hainan for systematic phenotyping.

In the spring of 2024, a randomized block design was employed to plant and phenotypically assess T_1 generation *OsGRW5.1* gene knockout mutants and wild-type plants in Beijing (latitude 39.9042°N, longitude 116.4074°E). Each genotype was planted in 2 rows with 10 plants per row and a spacing of 20 cm × 20 cm. The main traits measured included grain width and 1000-grain weight.

DNA isolation and genome sequencing.

Genomic DNA was extracted from leaf samples using the CTAB method (Doyle and Doyle, 1987), and the quality of the extracted DNA was assessed. The DNA library for re-sequencing was constructed using the TruSeq Nano DNA HT kit (Illumina, San Diego). First, the genomic DNA was randomly fragmented using ultrasonic waves, and DNA fragments of approximately 350 bp were selected as the target size. After electrophoretic recovery of these fragments, adapters were ligated to both ends, followed by PCR amplification (Saiki et al., 1988). The size distribution of the

amplification products was assessed using an Agilent 2100 Bioanalyzer, and the library was precisely quantified using quantitative PCR. The constructed library was then subjected to paired-end sequencing according to the standard protocol of the Illumina HiSeq 4000 platform (Illumina, San Diego).

Sequence quality checking and filtering.

We first performed quality control (QC) and adapter trimming on the raw sequencing data using the fastp tool (Chen et al., 2018). During the QC process, the following types of reads were excluded: (1) reads containing more than 10% unrecognized bases; (2) reads shorter than 50 bp; (3) reads with a quality score of $Q < 15$; and (4) low-quality reads where more than 50% of the bases had a Q-score ≤ 5. After quality control processing, the data were aligned to the reference genome NIP (*Oryza sativa* L. var. *Nipponbare*, MSU v7.0) using BWA-MEM (Li and Durbin, 2009) with the parameters: mem-T4-K3-M-R. Subsequently, variant detection was performed using the GATK HaplotypeCaller (McKenna et al., 2010) module, generating initial variant sites. To ensure high-quality variants, four filtering criteria were applied: QD < 4.0, FS > 200.0, SOR > 10.0, and ReadPosRankSum < -20.0, resulting in a final VCF file. In population analysis, low-frequency alleles and alleles with high missing rates or high heterozygosity can impact the accuracy of the results. Therefore, we further filtered the SNP sites using VCFtools (Danecek et al., 2011), removing the following sites that did not meet the criteria: (1) non-biallelic SNP sites; (2) sites with a minor allele frequency (MAF) < 0.05; (3) sites with a missing rate > 0.25; (4) sites with a heterozygosity rate > 0.8.

VCF file integration.

We used the merge command in bcftools (Li, 2011) (v1.10) to merge the VCF files from the two populations. To ensure the accuracy of the merging process, we first indexed each VCF file and generated the corresponding .csi index files using the bcftools index command. Subsequently, the VCF files from the two populations were merged using the bcftools merge command. This command integrated the variation data from all samples into a new VCF file based on the genomic positions and genotypic information of each variant. If the variant sites differed between the populations, the merged file

retained all unique variants and filled missing values for the absent sites. After the merging process was completed, the bcftools stats command was used to perform quality checks on the merged VCF file to verify the success of the merge, including checking the integrity of the variant sites, completeness of the sample information, and ensuring the accuracy of all variant and sample data. Following the quality control of the merged VCF file, we further filtered the variant sites using the following four criteria: 1) minor allele frequency (MAF) ≥ 0.05; 2) missing rate (max-missing) ≤ 0.75; 3) minimum number of alleles (min-alleles) = 2; 4) maximum number of alleles (max-alleles) = 2. These filtering criteria ensured the quality of the variant data while retaining meaningful variant sites for population genetic analysis and association studies. The final VCF file, after rigorous filtering, was suitable for subsequent population structure analysis and association studies.

Clustering and Principal Component Analysis.

The Neighbor-Joining Phylogenetic Tree (NJ) is a tree diagram used to describe the phylogenetic relationships among different varieties or populations, based on genetic characteristics to assess the degree of relatedness between groups. We constructed a phylogenetic tree for the 2,088 samples using the Neighbor-Joining method implemented in the VCF2Dis tool (https://github.com/BGI-shenzhen/VCF2Dis). The resulting tree was then visualized through the iTOL (Interactive Tree of Life) (Letunic and Bork, 2021). platform to display the genetic relationships between different varieties and populations. Principal Component Analysis (PCA) was performed using PLINK software (Purcell et al., 2007), which reduced the dimensionality of the variation data to reveal the primary genetic differences among the population samples. These differences were projected into the principal component space for intuitive visualization of population structure.

Population diversity analysis and specific locus calculation.

SNP data analysis was performed using PLINK software. The population genome was divided using PLINK, employing a sliding window approach with a window size of 100 Kb to calculate π values. The π value for each window was determined by calculating the average sequence difference between all possible base pairs within the

window. π values were generated for the entire genome across different populations, including landraces and hybrid rice parental lines from each subpopulation. Based on the integrated VCF file, the proportion of variant bases at each SNP site contributed by the landraces and hybrid rice parental lines from each subpopulation was calculated. If the contribution of any one population to the variant base at a given SNP site reached or exceeded 95%, the SNP site was classified as a population-specific site.

Trait diversity assessment

Use Excel to calculate the mean and standard deviation of seven quantitative traits. Based on the mean and standard deviation, rank the quantitative traits of all materials and classify them into three levels: low, medium, and high. The Shannon-Wiener diversity index (H') is used to evaluate the diversity of each trait. The calculation formula is as follow (Ortiz-Burgos, 2016):

$$H' = -\sum_{i=1}^{s} p_i \ln(p_i)$$

Where H' is the Shannon-Wiener diversity index, representing the diversity level of the sample; S is the number of different categories in the sample; p_i is the relative frequency of the i-th category, which is the proportion of that category in the total sample; $\ln(p_i)$ is the natural logarithm of the frequency of the i-th category. By calculating the Shannon-Wiener diversity index, we separately compute the diversity index for the seven traits in the 517 parent materials and 2,088 materials, and take their average values. At the same time, we calculate the average diversity index for the 517 parent materials and determine the percentage of this average relative to the average diversity index of the seven traits in the 2,088 materials.

Hybrid genome prediction

Definition and Coding of Hybrid Varieties: In this model, the alleles of the paternal and maternal parents are defined as A_1 and A_2, respectively. When the hybrid variety's allele at a given locus matches the paternal allele (A_1) or the maternal allele (A_2), it is assigned values of 1 and -1, respectively. Two different coding schemes are employed based on the genetic characteristics of the loci: **Additive Coding (Z)**: In the additive model, the hybrid genotype is coded as -1, 0, and 1, corresponding to the homozygous

maternal allele, heterozygous, and homozygous paternal allele, respectively. This method is used to capture the additive effects between alleles in the hybrid. **Dominant Coding (W)**: In the dominant model, the hybrid genotype is coded as 0 or 1, where 0 represents the absence of a dominant effect and 1 represents the presence of a dominant effect. This method is used to capture the contribution of dominant alleles. Based on these coding schemes, a mixed model of additive and dominant effects is used to predict the genotype. The final prediction results are standardized and centered using the scale() function in R to ensure comparability across different traits.

The definition formula for the hybrid prediction variables is:

$$Z\begin{cases} 1=\frac{1}{2}(1+1) & for A_1A_1 \\ 0=\frac{1}{2}[1+(-1)\,] & for A_1A_2 \\ -1=\frac{1}{2}[(-1)+(-1)\,] & for A_2A_2 \end{cases}$$

$$W\begin{cases} 0=\left|\frac{1}{2}(1-1)\right| & for A_1A_1 \\ 1=\left|\frac{1}{2}[1-(-1)\,]\right| & for A_1A_2 \\ 0=\left|\frac{1}{2}[(-1)-(-1)\,]\right| & for A_2A_2 \end{cases}$$

In this study, to further improve the accuracy of hybrid genotype prediction, we extended the existing additive-dominant mixed model by incorporating parental phenotypes as an additional influencing factor. This resulted in a predictive model that integrates additive-dominant effects with parental phenotypes. Based on the best linear unbiased prediction (GBLUP) of the additive-dominant mixed variable model (UV-AD) (Clark and van der Werf, 2013). we included the parental phenotypes as covariates in the model to enhance the accuracy of hybrid genotype prediction. In this model, parental phenotypes are treated as significant variables affecting the hybrid performance. By quantifying the additive and dominant effects of the parents in the model, we further improved the predictive power of the hybrid genotypes. **Additive Effect Integration**: By incorporating the phenotypic data of both the father and mother, the model can more accurately capture the additive effects of parental alleles, thereby

improving the prediction of the hybrid's phenotypic performance. The additive effects are encoded through the parental alleles (A_1 and A_2). **Dominant Effect Integration**: In addition to the additive effects, the model also captures the dominant effects between the parents. Dominant effect encoding (W) is used to describe the influence of parental allele combinations on the hybrid's traits. The inclusion of dominant effects enhances the model's ability to explain the complex traits of the hybrid. **Parental Phenotype Integration**: In the model, parental phenotypic information is incorporated as covariates, allowing the prediction to be based not only on genotype data but also on the potential impact of the parental phenotypes on the hybrid performance. This approach enables the model to more accurately predict the hybrid performance under different environmental conditions.

The definition formula for the hybrid prediction variable becomes:

$$y = X\beta + Z\gamma_a + W\gamma_d + \varepsilon$$

$$y = \frac{1}{2}(P_M + P_F)\beta_a + \frac{1}{2}|P_M - P_F|\beta_d + X\beta + Z\gamma_a + W\gamma_d + \varepsilon$$

In the formula, P_M and P_F represent the parental phenotypes, β_a and β_d represent the additive and dominance fixed effects, respectively. X is the fixed effect structure matrix used to predict y, and m and n denote the total sample size and the total number of markers, respectively. Z and W are the additive and dominance hybrid prediction variables, which are m × n matrices. γ represents the effect values of the markers, and ε is a random error vector, where $\varepsilon \sim N(0, I_n\sigma^2)$.

Two methods were employed in this study to predict the hybrid genotype. The first method ($F_{1Genotype01}$) utilizes the additive-dominance mixed model (UV-AD) based on the best linear unbiased prediction (GBLUP) from the "Predhy" package of R (Xu, 2017). This model integrates parental phenotype data to predict the genotypes of 770 hybrid samples. The second method ($F_{1Genotype012}$) follows the procedure outlined below: First, all heterozygous sites in the 946 parental VCF files are replaced with missing values to avoid interference from heterozygotes in subsequent analyses. Next, the Beagle software is used to impute the missing genotype data. Afterward, the imputed genotype data are converted to a 0-2 encoding format (0 for homozygous major alleles, 1 for heterozygotes, and 2 for homozygous minor alleles) using the Plink software to

ensure compatibility with subsequent analysis tools. Finally, genotype prediction for the hybrids is performed using the synthetic.cross() function from the "ASRgenomics" package of R (https://doi.org/10.32614/CRAN.package.ASRgenomics). This method effectively reduces the bias introduced by missing data, ensuring the accuracy and reliability of genotype predictions.

Hybrid phenotype prediction

Genome selection (GS) utilizes high-density SNP and phenotype information from parental samples and a subset of hybrid samples to establish the association between markers and phenotypes, thereby enabling the prediction of phenotypes for additional hybrids that have not yet undergone field trials. The model expression is as follows:

$$y = X\beta + \sum_{k=1}^{m} Z_k\gamma_k + \varepsilon$$

In this model, X represents the fixed effect structure matrix, β denotes the fixed effects, m is the total number of markers, Z_k represents the genotype vector for n individuals at the k-th marker, γ_k denotes the effect of the k-th marker, and ε is a random error vector, distributed as $\varepsilon \sim N(0, I_n\sigma^2)$. Based on this, the GBLUP prediction model combined with auxiliary traits is expressed as follows:

$$y = P_1\beta + X\beta + Z_k\gamma_k + W_k\gamma_k + \varepsilon$$

In this model, auxiliary traits are treated as fixed effects, with P1 representing the phenotype value of the auxiliary trait. X denotes the fixed effect structure matrix, β represents the fixed effects, γ_k indicates the effect of the k-th marker, and Z_k and W_k represent the additive and dominance genotype vectors for n individuals at the k-th marker, respectively. ε is a random error vector, distributed as $\varepsilon \sim N(0, I_n\sigma^2)$.

Genome-Wide Association Studies

In this study, the QK mixed linear model was used for analysis, and the EMMAX software package was employed (Kang et al., 2010). The expression of the model is as follows:

$$y = X\alpha + Q\beta + K\mu + e$$

In this study, y represents the phenotype vector, X is the genotype matrix, α is the genotype effect vector, Q is the fixed effect vector, β is the fixed effect vector, K is the

630 random effect matrix, μ is the random effect vector, and e is the residual vector. The P-
631 value indicates the likelihood that the genotype effect vector α for each SNP is zero; the
632 smaller the P-value, the stronger the association between the SNP and the phenotype.
633 The association results are visualized using Manhattan and Quantile-Quantile (QQ)
634 plots. The Manhattan plot displays the P-values for each SNP across all chromosomes,
635 while the QQ plot shows the overall association effect for all SNPs. In the initial phase,
636 if the observed values are close to the expected values, it indicates that the model's
637 association effect aligns well with the actual data. As the analysis progresses, if the
638 observed values exceed the expected values, it indicates that the model has accurately
639 identified significant loci. Finally, significant SNPs are annotated for their gene
640 functions based on decay distance or conventional genetic distance. The x-axis of the
641 Manhattan plot corresponds to the 12 chromosomes of rice, while the y-axis shows the
642 -log10(*P*) value of each SNP's effect on the phenotype. We calculated the suggestive
643 threshold using the formula: $-\log_{10}$(1/effective number of independent SNPs) as
644 previously described(Wang et al., 2016). To determine the effective number of
645 independent SNPs, we used PLINK with a window size of 50, step size of 50, and $r^2 \geq$
646 0.3, resulting in 102,526 effective independent SNPs in the full population. Therefore,
647 we set the significance threshold at: $-\log_{10}(P)=5$. SNP blocks within 70KB of a
648 significant SNP are defined as candidate associated regions. Genes within these regions
649 are selected as candidate genes for GWAS association. Based on related SNPs, allele
650 types with favorable agronomic traits (e.g., earlier heading date (HD), shorter plant
651 height (PH), wider flag leaf width (FLW) and grain width (GW), higher 1000-grain
652 weight (TGW), and longer panicle length (PL) and grain length (GL) are considered
653 advantageous.

654 We conducted a genome-wide association study (GWAS) using the rMVP package
655 of R (Yin et al., 2021).The phenotype data included seven traits from 770 hybrids, and
656 the genotype data were derived from the predicted 01-type and 012-type hybrid
657 genotypes. Genotype quality control was performed using VCFtools, retaining SNPs
658 with a minor allele frequency (MAF) ≥ 0.05. After quality control, a total of 5,415,129
659 SNPs and 4,311,972 SNPs were retained for subsequent analysis. The MVP.Data()

Journal Pre-proof

function within the rMVP package was used to convert the genotype and phenotype data into a suitable format for analysis. GWAS analysis was performed using the MVP() function in rMVP. To account for population structure and hidden relatedness, a mixed linear model (MLM) was used for association analysis. In the MLM, the kinship matrix was used to correct for genetic relatedness among individuals, and the first three principal components (PCs) were used to adjust for population structure. The Genotype01 file was used to calculate the dominant effects of the loci, while Genotype012 was used to calculate the additive effects. The x-axis of the Manhattan plot represents the 12 chromosomes of rice, and the y-axis shows the $-\log_{10}(P)$ value for each SNP's effect on the phenotype. A threshold of $-\log_{10}(P) > 5$ was set, with points above this threshold considered significantly associated with the phenotype under investigation.

Degree of Dominance

The degree of dominance 'd/a' was calculated using the peak SNPs of the associated loci from $F_{1Genotype01}$ and $F_{1Genotype012}$ GWAS, where 'd' and 'a' referred to the dominant effect and the additive effect, respectively. The effects of heterozygous alleles were analysed for the GWAS peaks above the suggestive *P* value ($-\log_{10}(P) > 5$, from the linear mixed model) underlying the heading data, grain length, grain width, thousand grain weight, panicle length, plant height and flag leaf width. In the calculation of 'd/a', the lowest p-value of the SNP in each 200 Kb genomic region is recorded as an association signal for that site. Peak SNPs of the top 100 associated sites (sorted by associated signal) were used for analysis. The SNP sites in which heterozygous genotypes or homozygous genotypes of both the minor alleles had a frequency of ≤ 15 in number were excluded. The average phenotypic values of heterozygous and homozygous genotypes were calculated for each significant associated SNP to evaluate the effects of heterozygous and homozygous genotypes.

Accumulation of Superior Locus and Superior Haplotypes

In the GWAS of $F_{1Genotype01}$ and $F_{1Genotype012}$, the significant correlation sites ($-\log_{10}(P) > 5$) were selected, and the phenotypes with better performance such as short heading time, long grain length, wide grain width, large 1000-grain weight, long ear length,

Journal Pre-proof

plant height and blade width were defined as dominant phenotypes. At the same time, the phenotypic grouping significance of heterozygous and homozygous alleles was compared for the dominant phenotypes, and the sites where the heterozygous alleles were significantly superior to the homozygous alleles were identified as the dominant heterozygous alleles. The corr.test() function of the psych package of R was used to calculate the correlation between the accumulation of predominance heterozygous alleles and phenotype (https://doi.org/10.32614/CRAN.package.psych).

Significant correlation sites ($-\log_{10}(P) > 5$) were screened at GWAS in $F_{1Genotype012}$, and each phenotypic significant site was annotated and haplotype analyzed. Excellent haplotypes were identified for candidate genes with significant phenotypic differences. Then the number of dominant haplotypes in 770 hybrids was counted, and the correlation between the cumulative number of dominant haplotypes and phenotypes was tested.

Gene cloning and plant transformation

The sgRNA targeting the exon of *OsGRW5.1* gene was designed using CRISPR-P 2.0 and cloned into the pCAMBIA1300 vector containing Cas9. The recombinant plasmid was sequenced and verified, then transformed into the Agrobacterium strain EHA105. Subsequently, the Agrobacterium-mediated transformation method was used to transform callus tissue of Nipponbare. After co-cultivation, the callus tissue was cultured on a selection medium containing 50 mg/L hygromycin to select for resistant callus. The resistant callus was then transferred to a differentiation medium to induce plant regeneration, and eventually transferred to a rooting medium for root development. Genomic DNA was extracted from T_0 transgenic plants, and PCR amplification of the target *OsGRW5.1* gene fragment was performed, followed by sequencing to verify the mutations. Mutation analysis was conducted using Sequencher (5.4.5) software to confirm the presence of frameshift mutations (Nishimura, 2000). T_0 and T_1 generations were planted in Hainan, China, during the winter of 2023 and the spring of 2024, respectively, for phenotypic evaluation.

CODE AND DATA AVAILABILITY

To facilitate user-friendly access to our hybrid prediction tools and datasets, we

developed the Predicting Rice Germplasm Hybrid Phenotype (PRGHP) web platform, publicly available in (https://hybrice.cn/). All software or scripts used in the study are publicly available as described in methods. The raw sequence data have been deposited in the NCBI GenBank database under accession numbers PRJNA65690, PRJEB6180 and China National Genomics Data Center under accession numbers PRJCA045734. All source datasets were available in the Supplemental tables.

SUPPLEMENTAL INFORMATION

Supplemental information is available at Molecular Plant Online.

FUNDING

This study was supported by the National Natural Science Foundation of China (32261143465, 32350710198), the Project of Sanya Yazhou Bay Science and Technology City (SKJC-2023-02-001 and SCKJ-JYRC-2023-47), the National Key Research and Development Program of China (2021YFD1200101), the Project of Hainan Province Science and Technology Special Fund (ZDYF2022XDNY260 and ZDYF2024KJTPY001), the Nanfan special project, CAAS (YBXM2403, YYLH2501 and YBXM2556), and the Project of Hainan Province Science and Technology Innovation (KJRC2023A01), the Hainan Province International Scientific and Technological Cooperation Talent and Exchange Project (Foreign Expert Program) Plan(G20241024007E).

AUTHOR CONTRIBUTIONS

X.M. Z. and Q. Q. conceived and designed the study. Y.T. W., Z.X. L., B. W. and X.Q. W analyzed the data and wrote the manuscript and edited the manuscript. D.J. L., J.Y. G., Z.R. L., W.Y. F., W.L. G. and F. L. developed the materials and collected the phenotypic data. N. Z., J. Y. and K. W. carried out the transgenic plants. W.Y. Y., Y. X. and C.W. X. developed the R-package predhy and performed the predicting of phenotype. W.H. Q., H.B. P., L.N. Z., Q.W. Y., and D.Y. Y. contributed valuable suggestions for this research. All authors read and approved the final manuscript.

Journal Pre-proof

ACKNOWLEDGMENTS

We thank Danting Li and Baoxuan Nong for their excellent fieldwork. The computations in this paper were performed on the bioinformatics computing platform of Hainan Artificial Intelligence Computing Center. No conflict of interest is declared.

Journal Pre-proof

REFERENCES

Ashikari, M., Sakakibara, H., Lin, S.Y., Yamamoto, T., Takashi, T., Nishimura, A., Angeles, E.R., Qian, Q., Kitano, H., and Matsuoka, M. (2005). Cytokinin oxidase regulates rice grain production. Science **309**:741-745.

Chen, F.D., Yan, B.X., and He, Z.H. (2021). Mechanisms of disease resistance to bacterial blight and perspectives of molecular breeding in rice. Plant Physiology Journal **56**:2533-2542.

Chen, S.F., Zhou, Y.Q., Chen, Y.R., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics **34**:i884-i890.

Cheng, S.H., Zhuang, J.Y., Fan, Y.Y., Du, J.H., and Cao, L.Y. (2007). Progress in Research and Development on Hybrid Rice: A Super-domesticate in China. Annals of Botany **100**:959-966.

Clark, S.A., and van der Werf, J. (2013). Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values. In *Genome-Wide Association Studies and Genomic Prediction* (ed. Cedric, G. and van der Werf, J. and Ben, H.), pp. 321-330. Totowa: Humana Press.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics **27**:2156-2158.

Ding, J.H., Lu, Q., Ouyang, Y.D., Mao, H.L., Zhang, P.B., Yao, J.L., Xu, C.G., Li, X.H., Xiao, J.H., and Zhang, Q.F. (2012). A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proceedings of the National Academy of Sciences **109**:2654-2659.

Doyle, J.J., and Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull **19**:11-15.

Fan, Y.R., Yang, J.Y., Mathioni, S.M., Yu, J.S., Shen, J.Q., Yang, X.F., Wang, L., Zhang, Q.H., Cai, Z.X., Xu, C.G., et al. (2016). *PMS1T*, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. Proceedings of the National Academy of Sciences **113**:15144-15149.

Fukuoka, S., Ebana, K., Yamamoto, T., and Yano, M. (2010). Integration of Genomics into Rice Breeding. Rice **3**:131-137.

Gao, H., Ma, Z.B., Wang, Y.Z., Zhang, M.L., Wang, X.J., Wang, C.H., Tang, Z.Q., Zhang, L.Y., Fu, L., He, N., et al. (2022). Analysis of parental genetic diversity and its impact on grain yield and quality of japonica hybrid rice in northern China. The Crop Journal **10**:904-910.

Gu, Z., and Han, B. (2024). Unlocking the mystery of heterosis opens the era of intelligent rice breeding. Plant Physiology **196**:735-744.

Gu, Z.L., Gong, J.Y., Zhu, Z., Li, Z., Feng, Q., Wang, C.S., Zhao, Y., Zhan, Q.L., Zhou, C.C., Wang, A.H., et al. (2023). Structure and function of rice hybrid genomes reveal genetic basis and optimal performance of heterosis. Nature Genetics **55**:1745-1756.

Hochholdinger, F., and Yu, P. (2025). Molecular concepts to explain heterosis in crops. Trends in Plant Science **30**:95-104.

Hua, J.P., Xing, Y.Z., Wu, W.R., Xu, C.G., Sun, X.L., Yu, S.B., and Zhang, Q.F.

(2003). Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proceedings of the National Academy of Sciences **100**:2574-2579.

Huang, M. (2022). Hybrid breeding and cultivar diversity in rice production in China. Agricultural & Environmental Letters **7**:e20074.

Huang, X.H., Yang, S.H., Gong, J.Y., Zhao, Q., Feng, Q., Zhan, Q.L., Zhao, Y., Li, W.J., Cheng, B.Y., Xia, J.H., et al. (2016). Genomic architecture of heterosis for yield traits in rice. Nature **537**:629-633.

Huang, X.H., Yang, S.H., Gong, J.Y., Zhao, Y., Feng, Q., Gong, H., Li, W.J., Zhan, Q.L., Cheng, B.Y., Xia, J.H., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nature Communications **6**:6258.

Iyer-Pascuzzi, A., and McCouch, S. (2007). Recessive Resistance Genes and the *Oryza sativa*-Xanthomonas oryzae pv. oryzae Pathosystem. Molecular plant-microbe interactions : MPMI **20**:731-739.

Jahnke, S., Sarholz, B., Thiemann, A., Kühr, V., Gutiérrez-Marcos, J.F., Geiger, H.H., Piepho, H.P., and Scholten, S. (2010). Heterosis in early seed development: a comparative study of F_1 embryo and endosperm tissues 6 days after fertilization. Theor Appl Genet **120**:389-400.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature Genetics **42**:348-354.

Khan, M.H., Dar, Z.A., and Dar, S.A. (2015). Breeding strategies for improving rice yield—a review. Agricultural Sciences **6**:467-478.

Letunic, I., and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research **49**:W293-W296.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**:2987-2993.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754-1760.

Liu, W.W., Zhang, Y.L., He, H., He, G.M., and Deng, X.W. (2022). From hybrid genomes to heterotic trait output: Challenges and opportunities. Current Opinion in Plant Biology **66**:102193.

Luo, D.P., Xu, H., Liu, Z.L., Guo, J.X., Li, H.Y., Chen, L.T., Fang, C., Zhang, Q.Y., Bai, M., Yao, N., et al. (2013). A detrimental mitochondrial-nuclear interaction causes cytoplasmic male sterility in rice. Nature Genetics **45**:573-577.

Martini, J.W.R., Molnar, T.L., Crossa, J., Hearne, S.J., and Pixley, K.V. (2021). Opportunities and Challenges of Predictive Approaches for Harnessing the Potential of Genetic Resources. Frontiers in Plant Science **12**:674036.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research **20**:1297-1303.

Melchinger, A.E. (1999). Genetic Diversity and Heterosis. In *Genetics and Exploitation of Heterosis in Crops* (ed. James, G.C. and Shivaji, P.), pp. 99-118. Madison, WI: ASA, CSSA, and SSSA.

Nishimura, D. (2000). Sequencher 3.1.1. Biotech Software & Internet Report **1**:24-30.

Ortiz-Burgos, S. (2016). Shannon-Weaver Diversity Index. In *Encyclopedia of Earth Sciences Series* (ed. Kennish, M.J.), pp. 572-573. Dordrecht: Springer Netherlands.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J.L., Sklar, P., Bakker, P.I.W.D., and Daly, M.J. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics **81**:559-575.

Qin, P., Lu, H.W., Du, H.L., Wang, H., Chen, W.L., Chen, Z., He, Q., Ou, S.L., Zhang, H.Y., Li, X.Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell **184**:3542-3558.e3516.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988). Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. Science **239**:487-491.

Song, S.F., Li, Y.X., Qiu, M.D., Xu, N., Li, B., Zhang, L.H., Li, L., Chen, W.J., Li, J.L., Wang, T.K., et al. (2024). Structural variations of a new fertility restorer gene, *Rf20*, underlie the restoration of wild abortive-type cytoplasmic male sterility in rice. Molecular Plant **17**:1272-1288.

Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T., Gardner, J., Wang, B., Zhai, W.X., Zhu, L.H., et al. (1995). A Receptor Kinase-Like Protein Encoded by the Rice Disease Resistance Gene, *Xa21*. Science **270**:1804-1806.

Tan, B.Y., and Ingvarsson, P.K. (2022). Integrating genome-wide association mapping of additive and dominance genetic effects to improve genomic prediction accuracy in Eucalyptus. The Plant Genome **15**:e20208.

Voss-Fels, K.P., Cooper, M., and Hayes, B.J. (2019). Accelerating crop genetic gains with genomic selection. Theoretical and Applied Genetics **132**:669-686.

Wang, X.L., Wang, H.W., Liu, S.X., Ferjani, A., Li, J.S., Yan, J.B., Yang, X.H., and Qin, F. (2016). Genetic variation in *ZmVPP1* contributes to drought tolerance in maize seedlings. Nature Genetics **48**:1233-1241.

Wei, X., Qiu, J., Yong, K.C., Fan, J.J., Zhang, Q., Hua, H., Liu, J., Wang, Q., Olsen, K.M., Han, B., et al. (2021). A quantitative genomics map of rice provides genetic insights and guides breeding. Nature Genetics **53**:243-253.

Wu, B., Hu, W., and Xing, Y.Z. (2018). The history and prospect of rice genetic breeding in China. Hereditas **40**:841-857.

Xu, S.Z. (2017). Predicted Residual Error Sum of Squares of Mixed Models: An Application for Genomic Prediction. G3 Genes|Genomes|Genetics **7**:895-909.

Xu, S.Z., Zhu, D., and Zhang, Q.F. (2014). Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proceedings of the National Academy of Sciences **111**:12456-12461.

Xu, S.Z., Xu, Y., Gong, L., and Zhang, Q.F. (2016). Metabolomic prediction of yield in hybrid rice. The Plant Journal **88**:219-227.

Xu, Y., Ma, K.X., Zhao, Y., Wang, X., Zhou, K., Yu, G.N., Li, C., Li, P.C., Yang,

Z.F., Xu, C.W., et al. (2021). Genomic selection: A breakthrough technology in rice breeding. The Crop Journal **9**:669-677.

Yin, L.L., Zhang, H.H., Tang, Z.S., Xu, J.Y., Yin, D., Zhang, Z.W., Yuan, X.H., Zhu, M.J., Zhao, S.H., Li, X.Y., et al. (2021). rMVP: A Memory-Efficient, Visualization-Enhanced, and Parallel-Accelerated Tool for Genome-Wide Association Study. Genomics, Proteomics & Bioinformatics **19**:619-628.

Zhang, S.N., Huang, X.H., and Han, B. (2021). Understanding the genetic basis of rice heterosis: Advances and prospects. The Plant Journal **9**:688-692.

Zhao, Z., Ding, Z., Huang, J.J., Meng, H.J., Zhang, Z.X., Gou, X., Tang, H.W., Xie, X.R., Ping, J.Y., Xiao, F.M., et al. (2023). Copy number variation of the restorer Rf4 underlies human selection of three-line hybrid rice breeding. Nature Communications **14**:7333.

Zheng, X.M., Sun, J., Cheng, C., and Qian, Q. (2024). A new gene for restoring wild abortive-type cytoplasmic male sterility in rice. Molecular Plant **17**:1800-1802.

Zhou, G., Chen, Y., Yao, W., Zhang, C.j., Xie, W.b., Hua, J.P., Xing, Y.Z., Xiao, J.H., and Zhang, Q.F. (2012a). Genetic composition of yield heterosis in an elite rice hybrid. Proceedings of the National Academy of Sciences **109**:15847-15852.

Zhou, H., Liu, Q.J., Li, J.L., Jiang, D.G., Zhou, L.Y., Wu, P., Lu, S., Li, F., Zhu, L.Y., Liu, Z.L., et al. (2012b). Photoperiod- and thermo-sensitive genic male sterility in rice are caused by a point mutation in a novel noncoding RNA that produces a small RNA. Cell Research **22**:649-660.

Zhou, H., Zhou, M., Yang, Y.Z., Li, J., Zhu, L.Y., Jiang, D.G., Dong, J.F., Liu, Q.J., Gu, L.F., Zhou, L.Y., et al. (2014). RNase Z^{S1} processes *UbL40* mRNAs and controls thermosensitive genic male sterility in rice. Nature Communications **5**:4884.

Journal Pre-proof

914 **Figure Legends**

915 **Figure 1. Diversity Comparison Analysis of 2,088 accessions.**

916 **(A)** Neighbor-Joining Tree of 2,088 accessions including 1,392 landraces and 696
917 hybrid rice parent lines (HP) that including 321 sterile lines and 375 restorer lines.
918 Yellow lines represent HP, while red, purple, blue, dark blue, and orange represent
919 *Indica1* (*IND1*), *Indica2* (*IND2*), *Indica3* (*IND3*), *AUS*, *Tropical japonica* (*TRJ*) and
920 *Temperate japonica* (*TEJ*), respectively.

921 **(B)** The specific SNPs of hybrid rice parent lines compared to other subgroups (*IND3*,
922 *AUS*, *TEJ*, *TRJ*). Green bars represent the proportion of shared SNPs (share), and
923 yellow bars represent the proportion of unique SNPs (uniq).

924 **(C)** Genome-wide nucleotide diversity (π-value) analysis of landraces and HP,
925 Landraces were categorized into five groups: *AUS*, *IND1*, *IND2*, *IND3*, *TEJ* and *TRJ*.
926 $IND1^{HP}$ and $IND2^{HP}$ represent hybrid rice parent belongs to *IND1* and *IND2*.

927

Journal Pre-proof

928 **Figure 2. Phenotypic Prediction and Heterosis Analysis of Hybrid Combinations**
929 **(A)** Prediction accuracy of seven traits (FLW, flag leaf width; GL, grain length; GW,
930 grain width; HD, heading date; PH, plant height; PL, panicle length; TGW, thousand
931 grain weight) using five-fold cross-validation. 'A' and 'AD' represent the additive effect
932 model and the additive-dominant effect model, respectively.
933 **(B)** Effects of training sample size (ranging from 110 to 770 samples, with increments
934 of 110 samples) on prediction accuracy using GBLUP across seven traits.
935 **(C)** Predictive ability of seven traits across parental presence/absence combinations.
936 **(D)** Statistical analysis of positive mid-parent heterosis across multiple traits for each
937 hybrid combination. Blue indicates hybrid combinations exhibiting positive mid-parent
938 heterosis in a single trait, while red denotes those displaying positive mid-parent
939 heterosis (MPH) across all seven traits. A phylogenetic tree from 946 samples and
940 groups (*IND*, *AUS*, *TRJ*, *TEJ*) are shown on the right.
941 **(E)** Distribution of parents based on the percentage of hybrid combinations exhibiting
942 MPH across multiple traits. The table categorizes parents according to the proportion
943 of their hybrid combinations showing positive MPH for more than 4, 5, or 6 traits,
944 divided into five percentage ranges: 0-20%, 20-40%, 40-60%, 60-80%, and 80-100%.
945 **(F)** Geographical distribution of the 52 superior parents within the 946 population
946 across East Asia (EA), South Asia (SA), and Southeast Asia (SEA). Yellow and green
947 indicate superior parents (adv) and non-superior parents (disadv), respectively.
948 **(G)** Subgroup distribution of the 52 superior parents in the 946 population, classified as
949 *AUS, IND*, *TEJ* and *TRJ*.
950 **(H)** Geographical distribution of all hybrid combinations the 52 superior parents.
951 **(I)** Population distribution of all hybrid combinations involving the 52 superior parents.
952 **(J)** 2,088 parents and 946 hybrids were used to predict the phenotype of hybrids, and
953 the proportion of hybrids involving germplasm resources and non-germplasm
954 resources among the top 100 combinations of six phenotypes was determined.
955 **(K-H)** The phenotypes of the five top 100 combinations with these germplasm
956 resources were significantly different from those of the 176 hybrids formed by the
957 sterile line restorer line (the top 100 combinations with no germplasm resources

Journal Pre-proof

958 involved).

Journal Pre-proof

Figure 3. Genome-wide association study (GWAS) of 946 landraces and 770 F_1.

(A) The circles from inside-out represent GWAS results of 946 landraces (a) 770 $F_{1Genotype01}$ (b) and 770 $F_{1Genotype012}$ (c). Significant SNPs with $-\log_{10}(P)$ values greater than 5 are marked (a, b, and c), with points in different colors representing different traits.

(B) The numbers of candidate QTLs identified in the GWAS of parental lines, $F_{1Genotype012}$, and $F_{1Genotype01}$, along with their co-localization patterns. The bar chart shows the number of co-localized QTLs for each category, with specific numbers provided. The points and lines below indicate the co-localization relationships between different categories, and the right-side bar chart represents the total number of QTLs identified in each of the three categories.

Journal Pre-proof

971 **Figure 4. GWAS Analysis of Seven Traits and Identification and Functional**
972 **Validation of Grain Width Candidate Genes.**
973 **(A)** GWAS of grain width in chromosome 5, the y-axis represents the -$\log_{10}$(*P*) values.
974 Red points represent SNPs significantly associated with grain width.
975 **(B)** The local Manhattan plot for the candidate gene and the LD haplotype block map.
976 **(C)** The predicted domain of protein OsGRW5.1
977 **(D)** Nucleotide variation in promoter and exon region of *OsGRW5.1*.
978 **(E)** Haplotype network of *OsGRW5.1*.
979 **(F)** Comparison of grain width across different haplotypes of *OsGRW5.1*.
980 **(G)** Comparison of thousand grain weight across different haplotypes of *OsGRW5.1*.
981 **(H-M)** Functional analysis of *OsGRW5.1*. Genotypic identification of CRISPR/Cas9
982 knockout mutant. The bold sequence represents the target site, and the red box indicates
983 the PAM sequence (H). Phenotypic comparison of grain width (I), grain length (J) and
984 thousand grain weight between the wild-type and mutant plants. The scale bar is 1 cm.
985 Statistical significance (*P*-value) was calculated using a T-test for K, L, and M.
986

Journal Pre-proof

Figure 5. Hybrid Superiority Analysis of F_1 GWAS and Gene Loci.

(A) The significant SNP loci for seven traits (FLW, GL, GW, HD, PH, PL, and TGW) and their corresponding dominance *(d/a)*. The x-axis represents the -$\log_{10}(P)$ values of the SNP loci, and the y-axis shows the dominance (*d/a*) values.

(B) The Pearson correlation between the cumulative number of superior heterozygous alleles and phenotypic values for seven traits. The x-axis represents the cumulative number of superior heterozygous alleles, and the y-axis represents the corresponding phenotypic values for each trait.

(C) The distribution map of unique advantageous haplotypes on the chromosomes in 946 germplasm resources, where different colors represent different traits.

(D) The distribution of unique advantageous haplotypes in the 52 superior parental lines.

(E) The subgroup origins of unique advantageous haplotypes in 946 germplasm resources.

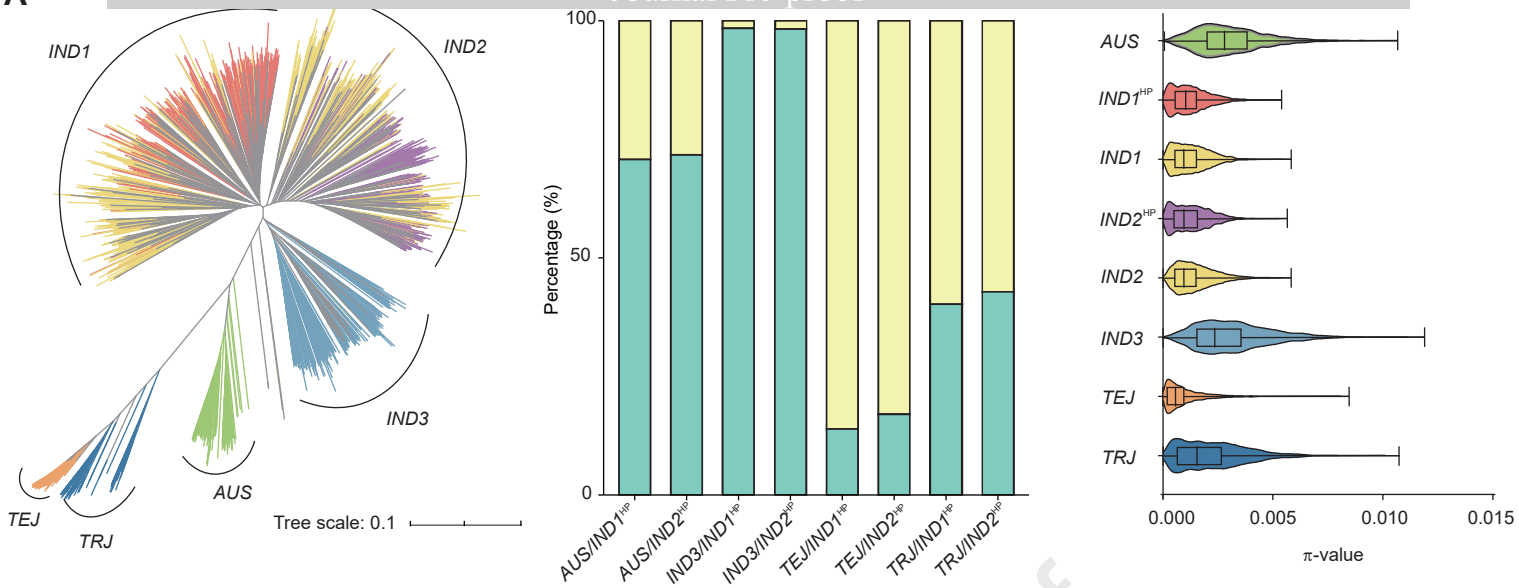
Journal Pre-proof

Figure 6. Utilizing Germplasm Resources for Genetic Improvement of Hybrid Rice.

The collection and introduction of globally diverse germplasm resources have enhanced diversity by over 65%. A GWAS on the germplasm and F_1 populations identified 45% of unique advantageous haplotypes that can be used to improve hybrid parent lines (restorer lines and sterile lines). The hybrid combination phenotypic prediction, using the 770 hybrid F_1 population as a training set, predicted a total of 446,985 hybrid combinations. This approach facilitated the selection of superior hybrid combinations, thereby guiding hybrid breeding and reducing the breeding workload by 95%.

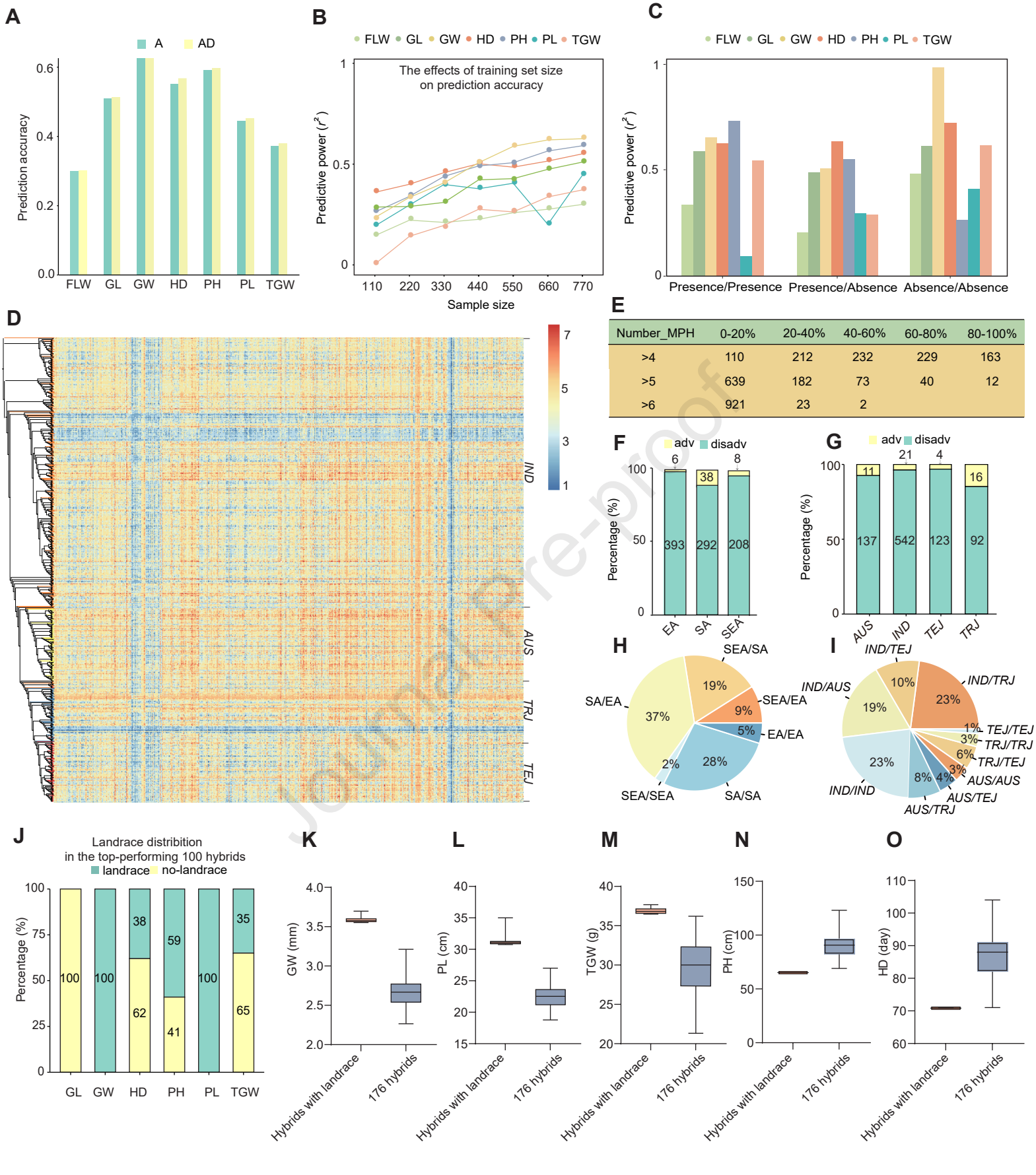
Journal Pre-proof

Journal Pre-proof

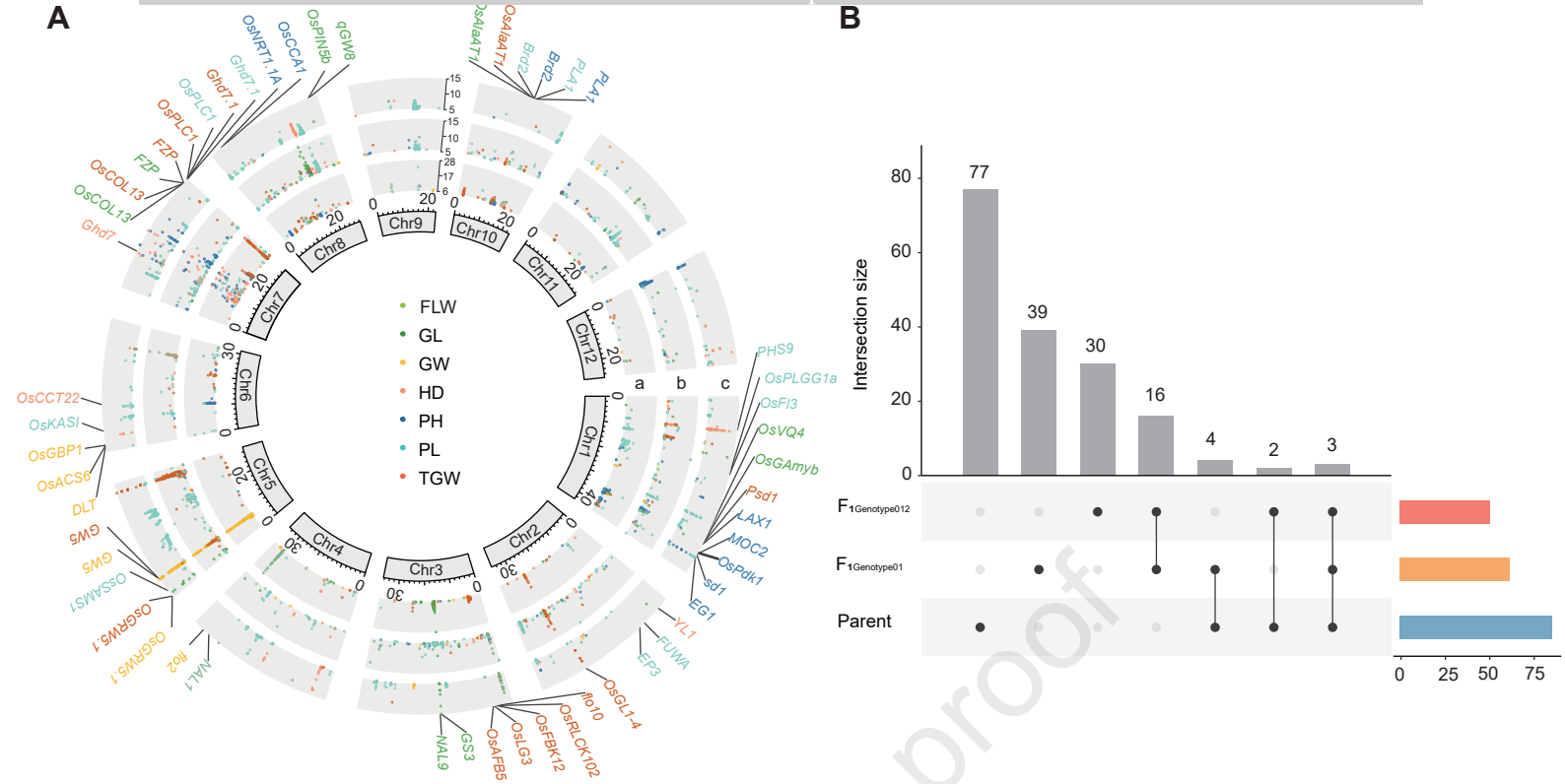


Journal Pre-proof

Journal Pre-proof

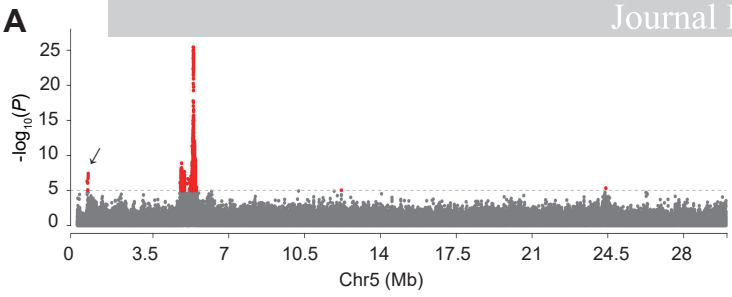


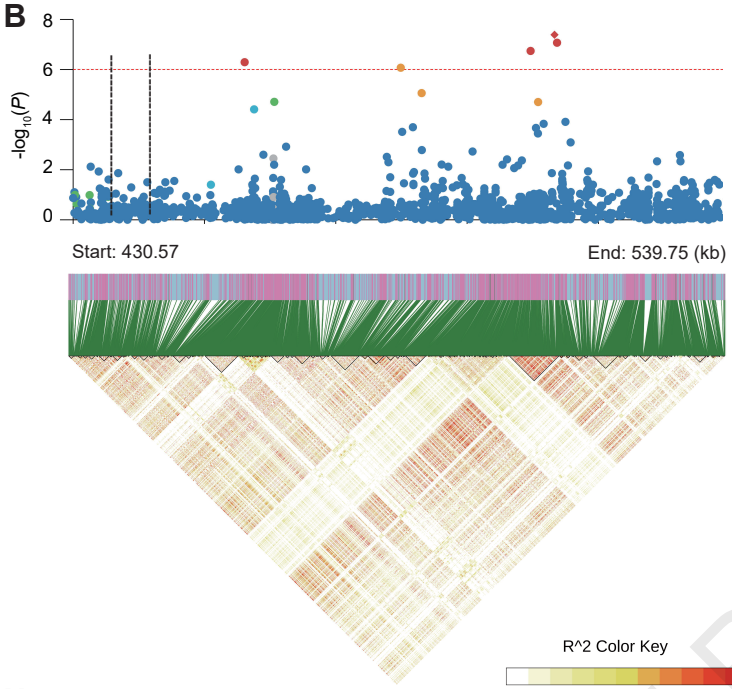
Journal Pre-proof

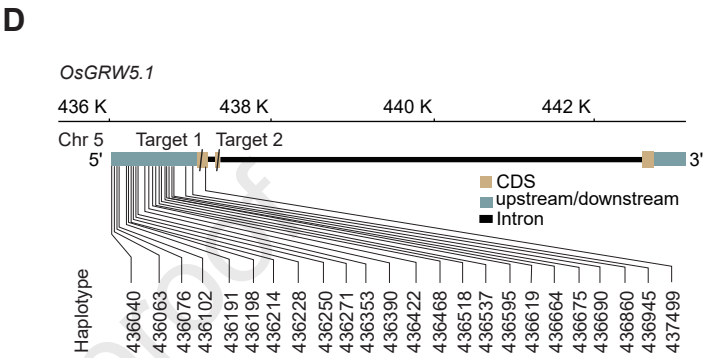


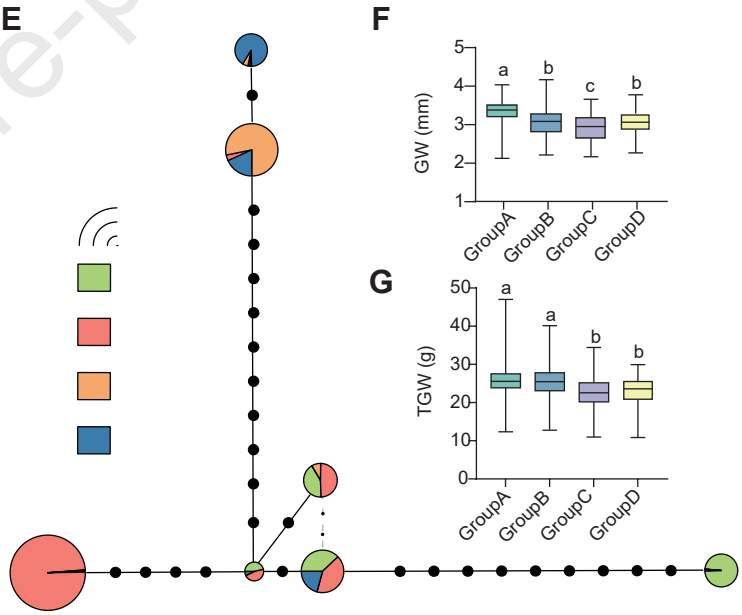
Journal Pre-proof

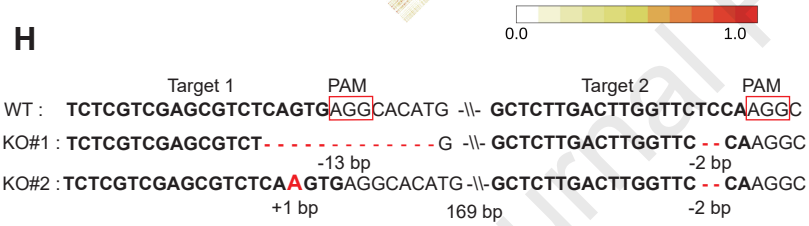
Journal Pre-proof

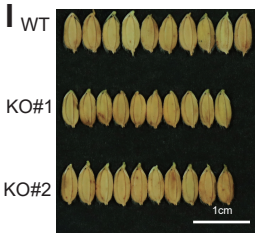


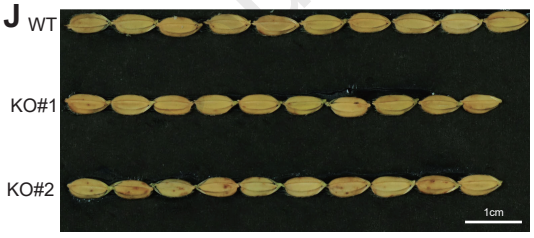


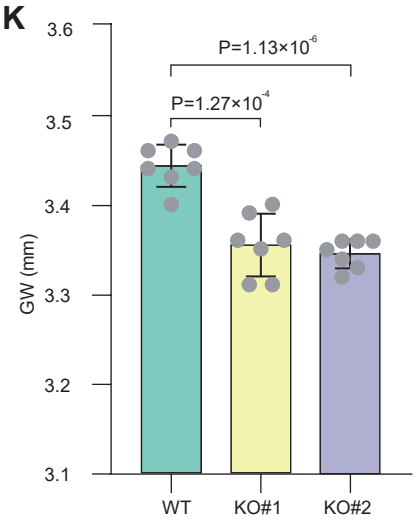


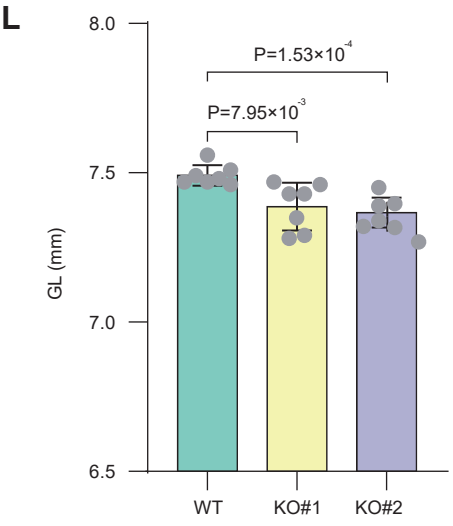


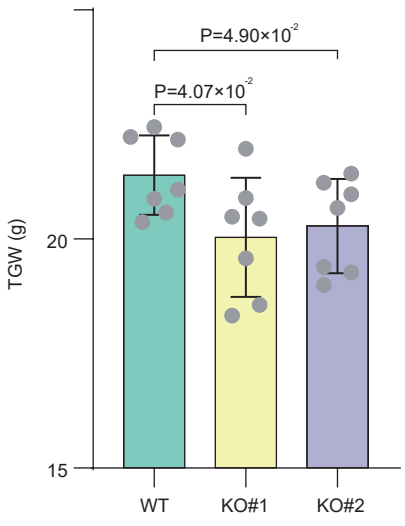


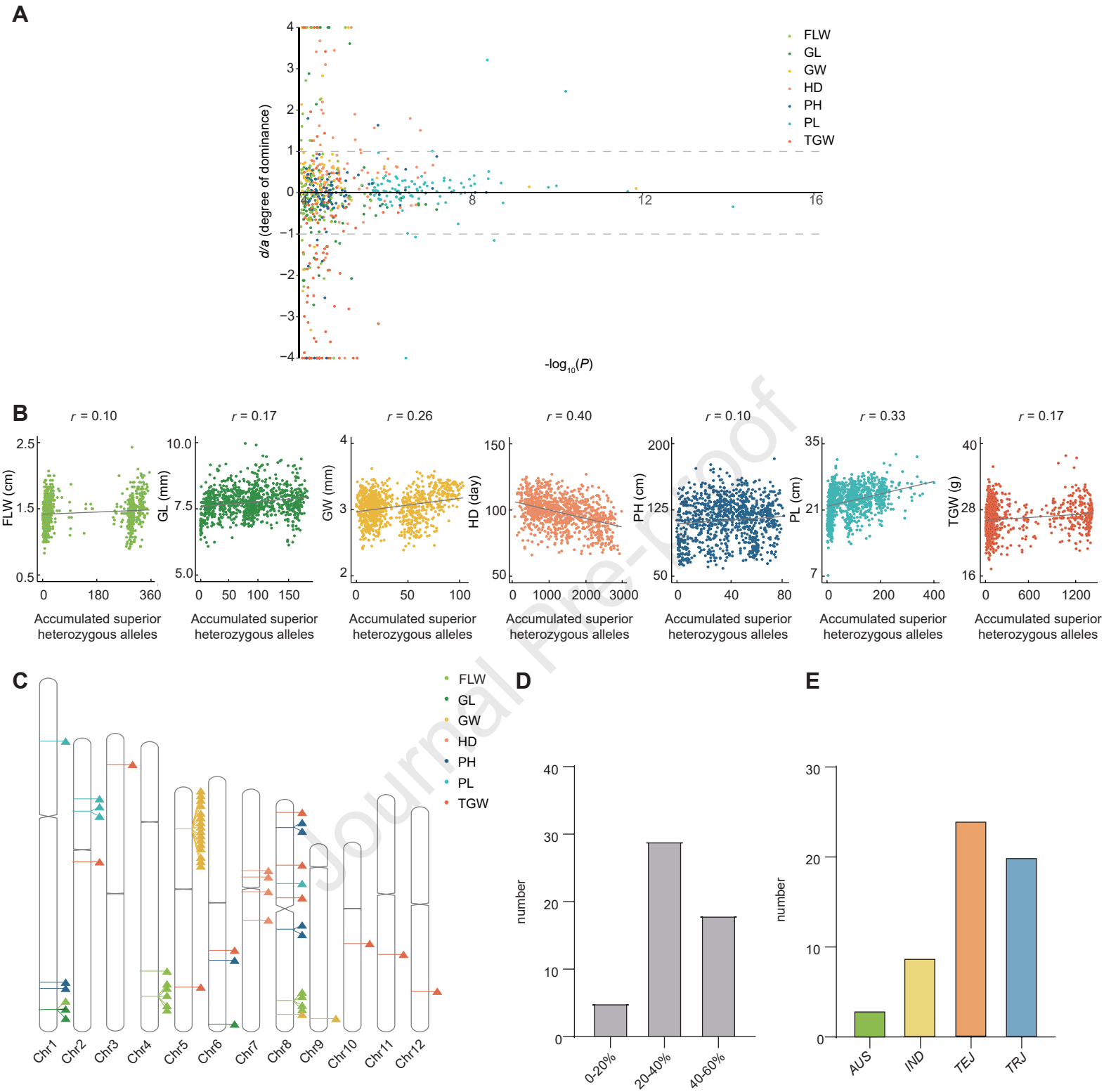






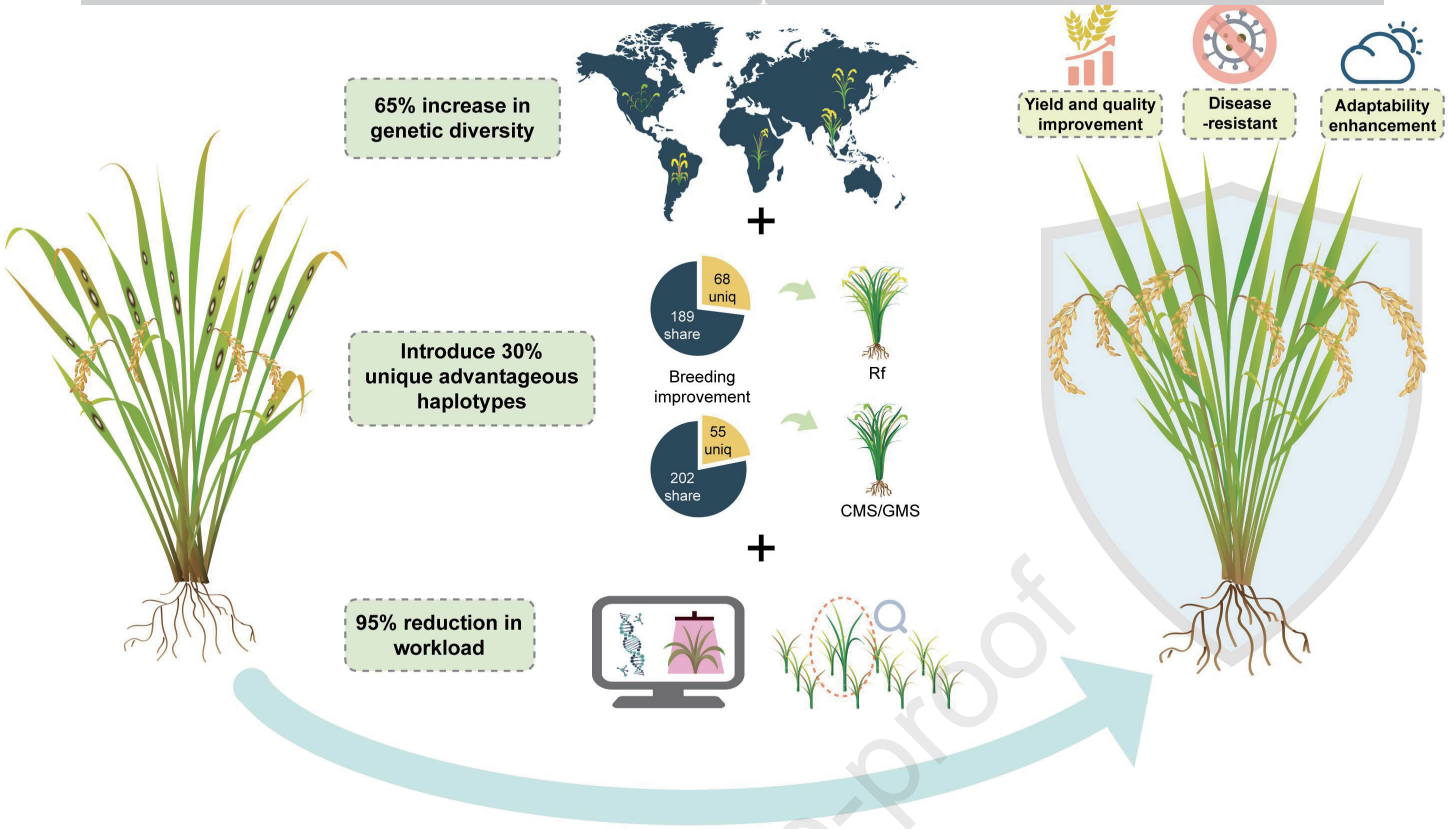






Journal Pre-proof

Journal Pre-proof



Journal Pre-proof