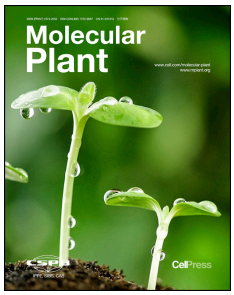



# Large-scale genomic and phenomic analyses of modern cultivars empower future rice breeding design

Xiaoding Ma, Hao Wang, Shen Yan, Chuanqing Zhou, Kunneng Zhou, Qiang Zhang, Maomao Li, Yaolong Yang, Danting Li, Peng Song, Cuifeng Tang, Leiyue Geng, Jianchang Sun, Zhiyuan Ji, Xianjun Sun, Yongli Zhou, Peng Zhou, Di Cui, Bing Han, Xin Jing, Qiang He, Wei Fang, Longzhi Han

PII: S1674-2052(25)00097-8

DOI: https://doi.org/10.1016/j.molp.2025.03.007

Reference: MOLP 1877

To appear in: *MOLECULAR PLANT*

Received Date: 25 September 2024

Revised Date: 25 January 2025

Accepted Date: 10 March 2025

Please cite this article as: Ma X., Wang H., Yan S., Zhou C., Zhou K., Zhang Q., Li M., Yang Y., Li D., Song P., Tang C., Geng L., Sun J., Ji Z., Sun X., Zhou Y., Zhou P., Cui D., Han B., Jing X., He Q., Fang W., and Han L. (2025). Large-scale genomic and phenomic analyses of modern cultivars empower future rice breeding design. Mol. Plant. doi: https://doi.org/10.1016/j.molp.2025.03.007.

# Large-scale genomic and phenomic analyses of modern cultivars empower future rice breeding design

Xiaoding Ma[1, †], Hao Wang[1, †], Shen Yan[1, †], Chuanqing Zhou[2, †], Kunneng Zhou[3], Qiang Zhang[4], Maomao Li[5], Yaolong Yang[6], Danting Li[7], Peng Song[8], Cuifeng Tang[9], Leiyue Geng[10], Jianchang Sun[11], Zhiyuan Ji[1], Xianjun Sun[1], Yongli Zhou[1], Peng Zhou[1], Di Cui[1], Bing Han[1], Xin Jing[2,*], Qiang He[1,*], Wei Fang[1,*], Longzhi Han[1,*]

[1]State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2]Smartgenomics Technology Institute, Tianjin 301700, China
[3]Anhui Province Key Laboratory of Rice Germplasm Innovation and Molecular Improvement, Rice Research Institute, Anhui Academy of Agricultural Sciences, Hefei 230031, China
[4]Jilin Provincial Laboratory of Crop Germplasm Resources, Rice Research Institute, Jilin Academy of Agricultural Sciences, Changchun 136100, China
[5]Jiangxi Research Center of Crop Germplasm Resources, National Engineering Laboratory for Rice (Nanchang), Rice Research Institute, Jiangxi Academy of Agricultural Sciences, Nanchang 330200, China
[6]State key laboratory of rice biology and breeding, China National Rice Research Institute, Hangzhou 310006, China
[7]Guangxi Key Laboratory of Rice Genetics and Breeding, Rice Research Institute, Guangxi Academy of Agricultural Sciences, Nanning 530007, China.
[8]National Key Laboratory of Crop Genetic Improvement, College of Plant Science & Technology, Huazhong Agricultural University, Wuhan 430070, China
[9]Key Lab of Southwestern Crop Gene Resources and Germplasm Innovation, Ministry of Agriculture; Biotechnology and Germplasm Resources Institute, Yunnan Academy of Agricultural Sciences, Kunming 620205, China
[10]Institute of Coastal Agriculture, Hebei Academy of Agriculture and Forestry Sciences, Tangshan 063300, China
[11]Institute of Crop Sciences, Ningxia Academy of Agricultural and Forestry Sciences, Yinchuan 750002, China

[†] These authors contributed equally to this work.

*Correspondence: Longzhi Han (hanlongzhi@caas.cn), Wei Fang (fangwei@caas.cn), Qiang He (heqiang@caas.cn), Xin Jing (jingxin@smartgenomics.cn)

**Short Summary**

We constructed a comprehensive genome variation map of modern rice using resequencing data from 6044 representative modern cultivars across five major rice-growing regions in China, revealing distinct regional breeding preferences. By integrating multiple datasets, we developed the *RiceAtlas* breeding design platform, which, for instance, facilitated the efficient optimization of grain shape in the Suigeng4 cultivar.

**Abstract**

Modern cultivated rice plays a pivotal role in global food security. China accounts for nearly 30% of the world's rice production and has bred numerous cultivated varieties over the last decades that are well adapted to diverse growing regions. However, the genomic bases that underlie the phenotypes of modern cultivars are poorly characterized, limiting access to this vast resource for breeding of specialized, regionally adapted cultivars. In this study, we constructed a comprehensive genetic variation map of modern rice using resequencing datasets from 6044 representative cultivars from five major growing regions in China. Genomic and phenotypic analyses of this diversity panel revealed regional preferences for genomic backgrounds and specific traits, such as heading date, biotic/abiotic stress resistance, and grain shape, associated with adaptation to local growing conditions and consumer preferences. We identified 3131 QTLs associated with 53 phenotypes across 212 datasets under different environmental conditions through genome-wide association studies. Notably, we cloned and functionally verified a novel gene related to grain length, *OsGL3.6*. By integrating multiple datasets, we developed *RiceAtlas*, a versatile multi-scale toolkit for rice breeding design. We rapidly improved the grain shape of the Suigeng4 cultivar using the *RiceAtlas* breeding design function. These valuable resources enhance our understanding of the adaptability and breeding requirements of modern rice and can facilitate advances in future rice-breeding initiatives.

**Key words:** Modern rice cultivar, Genomic bases, Rice-growing region, breeding design

## Introduction

Rice (*Oryza sativa* L.) is the world's most important food crop, serving as the staple food for over 60% of China's population and feeding half of the global population. China, a major rice producer, cultivates more than 26.6 million hectares annually, representing approximately 18% of the world's rice-growing area and contributing 28% of global rice production (Nie and Peng, 2017). With its diverse ecological conditions and agricultural demands, China has released thousands of cultivars through continuous genetic improvement (https://www.ricedata.cn). However, a lack of genomic information on these cultivars poses significant challenges for rice breeding in China, particularly as climate change leads to more frequent and severe stress events, highlighting the limited adaptability of elite regional cultivars.

Rice is cultivated from 18° N to 50° N latitude in China, covering a vast range that includes tropical, subtropical, warm-temperate, temperate, and cold-temperate climate conditions (Lv et al., 2018; Saud et al., 2022). On the basis of their ecological conditions, cropping systems, and rice cultivar types, China's rice-growing areas are classified into six regions: South China (SC), Central China (CC), Southwest Plateau (SW), North China (NC), Northwest Arid (NW) and Northeast (NE) (Ding, 1961). The SC region, which benefits from abundant water, heat, and light, grows mainly *indica* rice. The CC and SW regions grow both *indica* and *japonica* rice, with *indica* rice primarily grown in southern CC and low-elevation SW areas and *japonica* rice in northern CC and high-elevation SW areas. The majority of *japonica* rice is grown in the NC, NW, and NE regions, although an arid climate and limited water resources restrict its cultivation in the NW region (<1.0% of total acreage). The NC region, in the North China Plain, is characterized by abundant arable land, ample water resources, and a favorably warm climate. The NE region, including the Liaodong Peninsula, experiences significantly lower temperatures than other regions and grows mainly *japonica* rice in one-season cropping systems. The varied ecology of rice regions inevitably affects the genetic composition and diversity of the cultivars grown in each region.

Recent decades have seen substantial advances in rice genetics and genomics, largely

driven by the development of sequencing technologies and bioinformatics. The International Rice Genome Sequencing Program completed the first genome of *O. sativa* cv. Nipponbare in 2005 (Sasaki and Burr, 2000), with an update in 2013 (Sakai et al., 2013), and the first complete telomere-to-telomere reference genome was released in 2023 (Shang et al., 2023). Wang et al. (2018) constructed a rice pan-genome that included 3010 accessions, adding 268 Mb of novel sequences (Wang et al., 2018). More recent studies have expanded this estimate to ~1250 Mb and ~1520 Mb (Shang et al., 2022; Zhang et al., 2022). Qin et al. (2021) generated a high-quality pan-genome assembly of 33 accessions and detected 171,072 structural variants and 25,549 gene copy-number variants (Qin et al., 2021). Wei et al. (2021) constructed a map of rice quantitative trait nucleotides (QTNs) using a library of 404 accessions (Wei et al., 2021), and comprehensively explored QTNs and their genetic interactions for 16 agronomic traits using 18K rice lines (Wei et al., 2024). Ye et al. (2022) investigated the genetic changes that have occurred in major inbred rice cultivars over decades of genetic improvement in China (Ye et al., 2022). All these efforts have facilitated the integration of genomic research with practical applications in breeding. However, there remains a notable gap in research on the genetic basis of modern rice cultivars, particularly regarding the study of modern cultivars across different growing regions. This represents a critical challenge for future rice breeding and improvement efforts.

In this study, we used whole-genome resequencing to construct a comprehensive map of genomic variations based on 6044 accessions collected from the five major rice-growing regions of China (Supplementary Figure 1). Through population-scale genomic analyses, we explored the genetic diversity, population structure, breeding preferences, and selection pressures that underlie the phenotypes of this modern-cultivar diversity panel. To facilitate the use of large-scale diversity panels and associated data in rice breeding, we also established a publicly available database and analysis platform, *RiceAtlas* (https://www.cgris.net/RiceAtlas), that integrates genomic and phenotypic data from multiple rice research projects, including the 6044 accessions used in this study, the 3000 Rice Genomes Project (Wang et al., 2018), and a QTN

library comprising 404 accessions (Wei et al., 2021). To improve accessibility to the different types of information within this large data repository, we incorporated five main analysis functions into *RiceAtlas*: germplasm information, phenotype data, genome-wide association study (GWAS) results, genomic variation analysis, and a breeding design tool. As a proof-of-concept demonstration, we rapidly improved the grain shape of the Suigeng4 cultivar using the *RiceAtlas* breeding design function.

## Results

### Collection of 6044 rice accessions from five major growing regions in China

To investigate the genetic diversity of modern rice cultivars across the major rice-growing regions of China, we gathered 6044 accessions from five major growing regions (SC, CC, SW, NC, and NE) (Figure 1A, Supplementary Table 1), ensuring that they captured a broad range of geographic, genetic, and morphological variation. Among these accessions, 5164 were newly collected, and 880 were sourced from our previous collection (Cui et al., 2022; Han et al., 2022; Liu et al., 2022; Liu et al., 2023), yielding a total of 2706 *indica* and 3338 *japonica* accessions (Supplementary Table 1). Specifically, we included 1998 *japonica* accessions from the NE region, 397 *japonica* from the NC region, and 994 *indica* from the SC region. From the CC and SW regions where both *indica* and *japonica* are grown, we collected 1295 (CC-I) and 417 (SW-I) *indica* accessions and 478 (CC-J) and 465 (SW-J) *japonica* accessions (Figure 1B, Supplementary Table 2). This panel of genetic resources provides comprehensive representation of cultivars from the five major rice-growing regions in China.

### Phenotypic variation across different rice-growing regions

To characterize how agronomic traits of rice cultivars adapt to or reflect breeding preferences across different regions, we evaluated 11 agronomic traits for 3606 of the 6044 accessions at seven field sites across China. All 3606 accessions were grown and phenotyped at all seven locations, and phenotype data for all seven locations were used for the GWAS analysis (below). However, we initially characterized each cultivar using only the phenotype data recorded at the field site closest to its collection location in order to assess its performance under optimal growth conditions (Figure 1A, Supplementary Tables 3–6, Supplementary Figure 2). Accessions from the SW-I and SW-J groups had the latest heading date, i.e., flowering time (>120 days), followed by the NC group (119.63 ± 11.65 days) and the NE group (102.93 ± 7.26 days). The CC and SC groups had the earliest heading date (71.00–87.32 days) (Figure 1C, Supplementary Table 5). The SW-I and SC groups had the highest grain number per

panicle (>230 grains) (Figure 1D). The NE, NC, CC-I, CC-J, and SW-J groups had relatively high values of 1000-grain weight, with an average of ~25 g compared with ~21 g for the SW-I and SC groups (Figure 1E). The CC-J and CC-I groups had the highest yields (>26 g), whereas those of the NE, NC, SW-J, and SC groups were lower (average 21.96 g) (Figure 1F). Other traits also showed significant differences across the accession groups (Supplementary Figure 2). Rice cultivars from the different groups thus exhibited distinct regional phenotypic characteristics when grown under their optimal conditions, likely shaped by the combined influence of genetic, natural (e.g., temperature and day length) (Supplementary Figure 5), and human factors (e.g., cropping systems and dietary preferences).

**Genomic sequences, diversity, and population structure**

We resequenced the genomes of the 5164 newly collected accessions, obtaining 60.78 terabases (Tb) of sequencing data with an average read depth of 31.21× per accession (Supplementary Table 1). We aligned these clean reads, together with reads from 880 cultivars and four wild rice accessions published previously, to the *O. sativa* cv. Nipponbare IRGPS 1.0 reference genome. In total, we identified 5,694,922 single-nucleotide polymorphisms (SNPs) and 812,306 insertions and deletions (InDels) (Supplementary Table 7). Of these SNPs, 1,203,875 (21.14%) were located in exons, 874,006 (15.35%) in introns, 749,680 (13.16%) in the 2-kb regions upstream of transcription start sites, 599,123 (10.52%) in the downstream regions of translation stop sites, 2,033,682 (35.71%) in intergenic regions, and 234,556 (4.12%) in other regions (Supplementary Figures 6–7). Among these variants, 644,134 (11.80%) SNPs led to non-synonymous substitutions, and 46,989 (5.78%) indels caused frameshift mutations (Supplementary Table 8).

To investigate the genetic population structure and relationships among accessions from the five major rice-growing regions, we constructed a neighbor-joining (NJ) tree and performed population structure and Uniform Manifold Approximation and Projection (UMAP) analyses using 1,477,136 high-quality SNPs (MAF >0.01). The phylogenetic tree showed a clear distinction between *indica* and *japonica* rice

(Supplementary Figure 8A). However, accessions from different regional groups exhibited some overlap (Figure 2A), likely reflecting similarities in breeding objectives and cultivation environments that led to a degree of homogenization over the course of long-term breeding. To objectively assess the genetic characteristics of modern cultivars from different regions, we excluded 822 landraces, 14 accessions with unclear classifications, and the four wild rice accessions, leaving 5208 accessions for population genetic analysis. A UMAP analysis revealed that the accessions clustered into seven groups, roughly corresponding to their subspecies and geographic origins (Figure 2B). ADMIXTURE analysis at $K$=7 revealed distinct *indica* (SW-I, CC-I) and *japonica* (SW-J, CC-J) groups in the SW and CC accessions, consistent with the UMAP results (Supplementary Figures 8A and 8B).

The CC-I group had the highest nucleotide diversity ($\pi$) ($2.98 \times 10^{-3}$), whereas the CC-J group had the lowest ($1.07 \times 10^{-3}$) (Figure 2C, Supplementary Figure 8C). This suggests that the overall high genetic diversity of the CC region mainly originates from *indica* rice. The fixation index ($F_{ST}$) was lowest between the NC and CC-J groups (0.022), followed by the SC and CC-I groups (0.027) and the CC-I and SW-I groups (0.031). By contrast, the SW-I group was clearly genetically distinct from the NE and CC-J groups, with relatively high $F_{ST}$ values of 0.682 and 0.643, respectively (Figure 2C, Supplementary Figure 8D). To further explore population differentiation, we assessed linkage disequilibrium (LD) decay and found that it was more rapid in the SC, SW-I, and CC-I groups than in the NE, NC, CC-J, and SW-J groups (Figure 2D). Notably, the SW-J group displayed particularly rapid LD decay, consistent with its higher genetic diversity.

Analysis of allele accumulation showed that the SW-J group contained the largest number of private alleles (22,979 SNPs), followed by the NE group (1513 SNPs) and SW-I group (479 SNPs) (Supplementary Figure 9A). Doubleton sharing analysis revealed that the SC group shared a larger number of SNPs with the CC-I group (95% of total SNPs) than with any other groups (Supplementary Figure 9B), likely owing to the geographic proximity of these regions or to similar breeding goals and

environmental conditions.

**Regional variations in allelic combinations of heading-date genes**

Flowering time reflects major genetic and phenotypic differences among rice accessions from different regions (Huang et al., 2011). To explore potential allelic variations underlying the observed differences in heading date among the seven accession groups, we examined 47 allelic variants identified in 23 key genes associated with heading date (Supplementary Table 9). There were no significant differences in allelic variants of the key flowering regulators *Hd3a* (Takahashi et al., 2009) and *RFT1* (Peng et al., 2021) across groups, but the alternative allele of another key regulator, *Ehd1-2* (Li et al., 2022b), was detected only in the SW-J group (Figure 3A). We also examined the allele distributions of *Ghd7*, *DTH8*, and *Hd1*, which form complexes involved in photoperiod sensing and flowering regulation (Zong et al., 2021). The *Ghd7-5* loss-of-function (LOF) allele was present only in the NE group, whereas the CC-I, SW-I, and SC groups carried the alternative *Ghd7-1* and *Ghd7-4* alleles (Figure 3A, Supplementary Table 10). This distribution pattern aligns with previous findings that specific *Ghd7* alleles are linked to varietal adaptation (Xue et al., 2008).

We also examined alleles of *DTH8*, which functions as a flowering suppressor under long-day conditions and a flowering activator under short-day conditions (Dai et al., 2012; Wei et al., 2010; Yan et al., 2011). The CC-I, SW-I, and SC groups contained two distinct LOF alleles, *DTH8-3* and *DTH8-6*, whereas the NE, NC, CC-J, and SW-J groups predominantly contained the reference *DTH8* allele (Figure 3A, Supplementary Table 10). We next analyzed *Hd1*, whose reference alleles delay flowering under long-day conditions and promote flowering under short-day conditions. The *Hd1-7* LOF allele was found predominantly in the NE group, whereas the *Hd1-6* LOF allele was found in the CC-I, SW-I, and SC groups. By contrast, the NC, CC-J, and SW-J groups carried the same functional alleles as the reference genome, consistent with the photoperiod conditions in their corresponding regions (Figure 3A, Supplementary Table 10). These results suggest that heading-date alleles of different genes function in accordance with local light availability across different growing regions.

To investigate the distribution patterns of heading-date alleles, we identified 1163 unique combinations of the 47 allelic variants in key heading-date genes across all accession groups (Supplementary Tables 10–11). In the NE group, the top five allele combinations had a combined frequency of 32.95%; in the NC group, 61.50%; in the CC-J group, 79.10%; and in the SW-J group, 55.26% (Supplementary Tables 12–13). Notably, the combined frequency of the top five allele combinations was greater than 50% in the NC, CC-J, and SW-J groups but not in the NE group, possibly reflecting greater temperature variability in the NE region compared with the relatively stable conditions in the NC, SW, and CC regions (Supplementary Figure 5). In the CC-I, SW-I, and SC groups, the top five allele combinations had combined frequencies of 14.03%, 20.20%, and 20.96% (Supplementary Table 13). We next examined the heading dates of accessions carrying the top five allele combinations in each group. Whereas the NE, NC, CC-J, and SW-J groups showed minimal differences in heading date among the top allele combinations, the CC-I, SW-I, and SC groups exhibited more pronounced variation among the top allele combinations (Figure 3B). These findings reveal the most common allele combinations for heading-date genes in different groups of regional accessions and provide valuable genetic insights for molecular breeding, particularly for the development of cultivars adapted to diverse rice-growing environments.

On the basis of variations in heading-date genes, we developed a genomic selection (GS) model to accurately predict flowering time in different growing regions and guide future rice breeding. On the test dataset, the model demonstrated high robustness, with Pearson correlation coefficients of 0.84 for GZL, 0.86 for TH, 0.86 for HF, 0.88 for WH, 0.85 for HZ, 0.86 for KM, and 0.87 for NN (Supplementary Figure 10), confirming that data on heading-date allele combinations can be used to develop GS models for molecular breeding and improved regional adaptation of rice.

**Genetic selection preferences for agronomic traits across growing regions**

To investigate whether preferential selection has led to regional differences in the allele frequencies of genes associated with agronomic traits, we estimated the frequencies of favorable alleles for 152 genes linked to various traits, including yield components,

plant architecture, and aspects of biotic and abiotic stress tolerance (Supplementary Table 14). The mean frequencies of favorable alleles were 37.44%, 37.64%, 37.56%, and 36.75% in the NE, NC, CC-J, and SW-J groups, respectively, all of which consisted of *japonica* accessions. By contrast, the mean frequencies of favorable alleles were 46.24%, 45.30%, and 47.14% in the CC-I, SW-I, and SC groups, indicating that desirable alleles were more prevalent in *indica* cultivars (Figure 3C, Supplementary Tables 15–16).

The frequencies of favorable alleles for specific traits varied across groups. For instance, the favorable *Gn1a-1* genotype of the grain-number was detected at frequencies of 99.83%, 100.00%, and 100.00% in the CC-I, SW-I, and SC groups, respectively, but at frequencies of only 55.57%, 11.23%, 2.58% and 70.79% in the NE, NC, CC-J, and SW-J groups. Similar patterns of favorable allele frequency were observed in other grain number-related genes, including *LAX1-2*, *NOG1*, *GNP1*, *APO1-2*, and *BG2-1*. In addition, an allele of *OsMYB8* associated with early floret opening time was present at frequencies of >89% in the CC-I, SW-I, and SC groups but <2% in the NE, NC, CC-J, and SW-J groups. Among grain shape-related genes such as *GL3.2*, *GS5-2*, and *OsSPL12*, alleles associated with broad grains had combined average frequencies of 84.65%, 95.50%, 96.25%, and 83.07% in the NE, NC, CC-J, and SW-J groups but 20.21%, 29.31%, and 20.81% in the CC-I, SW-I, and SC groups (Figure 3C, Supplementary Table 16), consistent with the characteristically wider grains of *japonica* rice.

Favorable alleles involved in plant architecture and grain flavor have been subjected to widescale selection, likely contributing to the mean frequencies of 53.45%, 54.17%, 54.83%, and 52.13% for favorable alleles of plant architecture-related genes in the NE, NC, CC-J, and SW-J groups and mean frequencies of 40.94%, 41.89%, and 41.05% for these favorable alleles in the CC-I, SW-I, and SC groups. Such relatively high frequencies of favorable alleles were observed across all groups. In particular, the frequencies of favorable *IL13*, *SLR1*, *SBI/OsGA2ox4*, and *OsbHLH174* alleles exceeded 72% in all groups. Notably, the *D61-2*, *sd1-4*, *TAC1,* and *TIPS-11-9* favorable

alleles were present at high frequencies in the NE, NC, CC-J and SW-J groups (>99%), whereas *D2*, *APO1-1*, and *TIG1* frequencies were higher in the CC-I, SW-I, and SC groups (>61%) (Figure 3C, Supplementary Table 16).

Favorable alleles related to grain taste quality were present at combined average frequencies of 44.70% in *japonica* accessions and 44.36% in *indica* accessions. Among these genes, the favorable alleles *GBSSI-2*, *GBSSI-5*, and *GBSSI-6* had a combined average frequency of 91.58%, indicating positive selection for these alleles in all groups. Favorable alleles for genes involved in biotic and abiotic stress tolerance, fertilizer use efficiency, seed morphology, and other traits were more prevalent in *indica* cultivars (Figure 3C, Supplementary Table 16). Together, these results indicate that average favorable allele frequencies tended to be slightly higher in the CC-I, SW-I, and SC groups than in the NE, NC, CC-J and SW-J groups, with the exception of those related to plant architecture and some abiotic stress traits, which were higher in *japonica* accessions, and those related to grain flavor, which did not differ markedly among accession groups.

**Genomic selection signatures for different rice-growing regions**

Continuous artificial selection has driven directional improvements in the rice genome and corresponding phenotypic changes. To identify genomic signatures of selection in the five major rice-growing regions, we performed identity-by-descent (IBD) analysis (Figure 4A–C, Supplementary Figures 11–12, Supplementary Table 17) and identified a total of 1589 IBD segments. The NE group contained the highest number (404) and the SW-J group the lowest (266) (Supplementary Table 18). The annotation of these segments revealed that they contained 10,778 genes, of which 77 have been reported for the key agronomic trait QTGs (Supplementary Table 19). Notably, chromosomes 1, 2, and 3 exhibited higher densities of IBD segments (Supplementary Figure 12); GO enrichment analysis indicated that the genes in these segments were enriched in DNA binding, transmembrane transport and transporter activity functions (Supplementary Table 20), suggesting more intense selection pressures on these genomic regions during breeding.

Approximately 40.28% of the 1589 IBD segments were shared among two or more accession groups (Figure 4C), indicating convergent evolution or similar environmental selection pressures. Specifically, the IBD segment on chromosome 2 (~26 Mb) shared by the NC and NE groups was enriched in genes associated with cold and salt tolerance, reflecting the similar breeding goals in these two geographic regions (Supplementary Table 17). Some IBD segments were specific to individual groups: the NE group contained 176 (11.08%) such segments, the NC group 132 (8.31%), the SC group 141 (8.87%), the CC-I group 97 (6.10%), the CC-J group 161 (10.20%), the SW-J group 129 (8.17%) and the SW-I group 113 (6.99%).

To further reveal the association of these IBD segment with adaptive traits, we performed enrichment analysis of group-specific IBD segment. In the NE group, photoperiodism and seed development functions were significantly enriched, reflecting selection for yield optimization and disease resistance in a cold, short-season climate (Figure 4D). In the NC group, flower development and stress response functions were significantly enriched, indicating selection for reproductive resilience and stress tolerance in variable conditions. In the SC group, stress response and pest defense functions were significantly enriched, suggesting selection for stress tolerance and pest resistance in a humid environment. In the CC-I group, post-embryonic and metabolic functions were significantly enriched, suggesting selection for growth efficiency in productive conditions. In the CC-J group, embryo and stress response functions were significantly enriched, indicating selection for embryo vigor and broad adaptation in variable environments. In the SW-I group, abiotic stress and transcription functions were significantly enriched, reflecting selection for stress tolerance in diverse climates. In the SW-J group, stress and nitrogen regulation functions were significantly enriched, pointing to selection for resilience and nutrient efficiency in challenging conditions (Supplementary Figure 13). These findings highlight region-specific environmental adaptations.

At the gene level, heading date (*OsGI*, *Hd6*) were notably fixed in the NE group; biotic stress responses and fertilizer utilization genes (*OsCd1*, *SLG1*, *TOND1*) were

366 strongly selection in the SC group; and in the CC group, grain quality (*OsAAP6*,
367 *OsACS6*) and abiotic stress tolerance (*OsTPP7*) were the focus of selection in the CC
368 group. The NC group showed selection on the drought tolerance related gene *GH3-2*,
369 and the SW group had accumulated multiple stress-resistance genes (*TT3.1*, *HIS1*, *Bsr-*
370 *d1*). Fixed haplotypes were identified at the *GW5* locus (in the SW-I group) and the
371 *GW8* locus (in the SW-J group) (Figure 4E, Supplementary Table 19), highlighting
372 strong selection for grain shape and quality (Liu et al., 2017; Wang et al., 2012) . These
373 findings provide insight into how IBD segments have contributed to population-specific
374 adaptation and functional diversity in rice.

375 **GWAS of 53 phenotypes for key agronomic traits**

376 To dissect the genetic basis of important agronomic traits in rice and advance molecular
377 design breeding, we performed field experiments across five major rice-growing
378 regions in China and systematically evaluated 3606 rice accessions. We obtained 212
379 phenotypic datasets, each consisting of data for one of 53 phenotypes from one of 19
380 distinct locations measured in one of two years; not all phenotypes were measured in
381 all locations or years (Figure 5A–B, Supplementary Figures 2–4, Supplementary Tables
382 3–5). We then used these datasets to perform a large-scale GWAS for all 212 phenotypic
383 datasets. The 53 phenotypes for key agronomic traits could be divided into four
384 categories: abiotic stress (28 phenotypes), yield components (10 phenotypes), biotic
385 stress (14 phenotypes), and heading date (1 phenotype).

386 We identified a total of 3131 QTLs that were significantly associated with 53
387 phenotypes (Figure 5C, Supplementary Figures 14–19, Supplementary Table 21).
388 Among them, 450 QTLs showed significant associations with the same phenotype
389 across at least two phenotyping locations (Supplementary Table 22). We also identified
390 125 QTLs were shared among different phenotypes, suggesting the potential presence
391 of pleiotropic genes at these loci. Additionally, 2642 QTLs exhibited strong association
392 signals in only one location/year but no associations in others. For example, we
393 identified several significant GWAS signals on chromosome 11 associated with heading

394 date in Hefei but were not detected in other locations (Supplementary Figure 20). A
395 similar pattern was observed for other traits, such as plant height in Wuhan, where
396 significant GWAS signals were uniquely detected on chromosome 8 (Supplementary
397 Figure 22). These findings highlight the significant role of genotype-environments
398 interactions in shaping the phenotypic variation.

399 Grain shape is a fundamental trait that determines yield and quality, and
400 manipulation of grain shape can be essential for improving rice cultivars. We identified
401 a major peak on chromosome 3 in which the lead SNP (Chr03: 35,155,927) was
402 significantly associated with grain length ($P = 1.66\times10^{-7}$) (Figure 5D). Linkage
403 disequilibrium analysis of the peak region revealed that the lead SNP was located within
404 a ~60-kb block (from 35,151,384 to 35,214,566) that included 14 functional genes
405 (Supplementary Table 23). Interestingly, this locus overlapped with an IBD segment in
406 the SW-J group (Figure 5E). We investigated the function of these 14 genes and found
407 that *OsGL3.6* (*LOC_Os03g62060/Os03g0836800*), annotated as an indole-3-acetic
408 acid (IAA) amino acid hydrolase gene, was most likely to be the causal gene, as IAA-
409 related genes have previously been reported to regulate rice grain size (Ma et al., 2023a).

410 Haplotype analysis showed that *OsGL3.6* had three major haplotypes in 3547
411 accessions (Figure 5F). Hap1 was present at a frequency greater than 0.95 in the NE,
412 NC, CC-J, and SW-J groups, whereas Hap2 and Hap3 frequencies were higher in the
413 CC-I, SW-I, and SC groups (Figure 5G). Further analysis revealed that Hap1 and Hap2
414 were associated with a long-grain phenotype, whereas Hap3 conferred a short-grain
415 phenotype (Figure 5H–I) ($P<0.05$), providing further evidence that *OsGL3.6* is
416 involved in regulating grain length. To confirm the function of *OsGL3.6*, we knocked
417 out this gene in the *japonica* cultivar Zhonghua11 (ZH11) (Figure 5J, Supplementary
418 Table 24). Compared with wild-type plants (grain length, 6.52 mm), the two
419 independent knockout lines *osgl3.6-1* (6.32 mm, $P < 0.05$) and *osgl3.6-2* (6.35 mm, $P$
420 $< 0.05$) had significantly shorter grains (Figure 5K). Therefore, *OsGL3.6* represents a
421 promising target for regulation of rice grain shape in breeding programs. These analyses
422 demonstrate how large-scale GWAS and IBD approaches can facilitate rice research

and breeding.

## Accelerating rice breeding with *RiceAtlas*

By integrating genomic and phenotypic datasets for the 6044 accessions examined here, 3010 Asian cultivars from the 3K-Rice project (Wang et al., 2018), and 404 rice accessions reported by (Wei *et al.*, 2021), we constructed the comprehensive rice database *RiceAtlas* (https://www.cgris.net/RiceAtlas). *RiceAtlas* consists of five modules: Germplasm, Phenotype, GWAS, Variation, and Breeding (Figure 6A). It complements existing tools by integrating vast germplasm and genetic resources to facilitate various rice breeding strategies. It can be used to comprehensively assess region-specific ecological backgrounds, complementarity of allelic variations, and genetic similarity to obtain donor-parent recommendations for rice breeding design.

To demonstrate the breeding design function of *RiceAtlas*, we used it to successfully improve the grain shape of Suigeng4 (SG4) within two years. The SG4 cultivar has been widely planted in large areas of the NE region for the past 20 years. It features a short and round grain phenotype, with desirable flavor and quality profiles, and it is still cultivated in some parts of the NE region. However, breeders hope to develop a long-grain version of SG4 to meet market demands. To increase grain length in SG4, we used the breeding design system in *RiceAtlas* to guide our crossing strategies. As recommended by *RiceAtlas*, we selected Zhongkefa8 (ZKF8) as the donor to increase SG4 grain length. After a single backcross and subsequent genotyping of the progeny population, we obtained a target homozygous SG4 line (Figure 6B–C). Phenotyping of the introgression lines showed that grain length was significantly increased, while the flavor profile and ecological suitability of SG4 were retained (Figure 6D). Importantly, this entire process was completed within just two years, representing a substantial improvement in the rate and precision of breeding outcomes relative to traditional approaches. These results provide a proof-of-concept demonstration that *RiceAtlas* can serve as a key resource and a powerful, versatile tool for rice breeding design.

## Discussion

We generated large-scale genotype and phenotype data for thousands of modern cultivars from five major rice-growing regions, collectively covering 99% of China's annual rice cultivation area. Using these data, we characterized the genetic variation in modern Chinese rice cultivars and revealed genomic signatures of selection in cultivars from different regions. We identified numerous loci linked to key agronomic traits, including heading date, yield, and stress responses, which will be useful for advancing research in rice functional genomics. Leveraging these extensive data on rice phenotypes, population genomics, and GWAS cohorts, we developed the *RiceAtlas* platform to support rice research and breeding. *RiceAtlas* features an intuitive query interface and practical tools, including a precise and efficient system for the recommendation of parental lines to facilitate molecular breeding design and accelerate the breeding process.

China's vast geographic expanse and significant north-south latitude differences have resulted in distinct regional adaptations and selection preferences in modern rice breeding. Through an initial phenotypic analysis of local cultivars grown in their native rice-growing regions, we observed that heading date exhibits clear regional characteristics. Cultivars from the NC and SW regions have the longest heading dates. In the SW region, this is primarily due to high altitudes with low annual average temperatures, which slow rice growth (He and Tang, 2023). In the NC region, the longer heading date occurs because there are minimal constraints from subsequent crops and temperatures exceed 20°C until mid-October, conditions that are favorable for grain filling. Breeders in the NC region thus favor cultivars with a long growth duration in order to maximize yield and profits. In the NE group, heading dates average around 100 days. This reflects the high latitude, extended photoperiod, and low temperatures of the NE region. Moreover, rice in this region must be harvested before October 15 owing to a sharp temperature decline at the end of September ((Dong et al., 2023). By contrast, the CC and SC groups have shorter heading dates, typically between 70 and 80 days, primarily to accommodate subsequent crops or the double-cropping rice system (Xian

et al., 2023).

Across different growing regions, each subspecies exhibits similar selection trends in plant architecture, panicle type, and grain shape. Overall, *indica* varieties from the CC, SW, and SC regions are taller, with longer panicles, more grains per panicle, and longer and narrower grains than *japonica* varieties from the NC, NE, CC, and SW regions. These differences are consistent with the fundamental differentiation between the *indica* and *japonica* subspecies. Accessions from all regions had a similar tiller number of 7 to 10, consistent with the concept of ideal plant architecture in rice breeding (Wang et al., 2017). Notably, the SC region appears to prefer slender-grained varieties, as the SC group had narrower grain widths and higher grain length-to-width ratios compared with the SW-I and CC-I groups, consistent with the preference for slender *indica* rice grains in South China. The CC-I and CC-J groups had the highest 1000-grain weight and single-plant yield among all groups, highlighting the fact that modern varieties in the CC region are bred for high single-plant yields in order to achieve high yields per unit area (Xiao et al., 2021).

The SC, CC, and SW groups exhibited higher genetic diversity than the NC and NE groups, consistent with the well-established finding that *indica* rice generally exhibits greater genetic diversity than *japonica* rice (Campbell et al., 2020) Notably, both the SW-I and SW-J groups displayed high genetic diversity, supporting the notion that Southwest China serves as a major center of rice genetic diversity (Liu *et al.*, 2022).

For analyzing key heading-date allelic combinations, we integrated 23 major heading-date genes—more than previous studies focused on only the Ghd7–Hd1–DTH8 complex (Cai et al., 2021; Zhou et al., 2021)—thus enabling us to characterize the genetic basis of heading-date regulation in greater detail. Favorable alleles of genes associated with plant architecture and abiotic stress responses (e.g., *OsMYB2*, *OsCd1*, *OsCBL10*, *Sd1-4*, and *TAC1*) occurred at higher frequencies in the NE, NC, CC-J, and SW-J groups, suggesting strong selection for such traits in these regions. By contrast, the SW-I, CC-I, and SC regions exhibited more intense selection on genes associated with biotic stress resistance, yield, and nutrient use efficiency (e.g., *GW8*, *APO1*, *GNP1*,

509 *OsLG3*, and *TOND1*). Differences in the frequencies of favorable alleles across regions
510 suggest that many beneficial alleles have yet to be fully utilized and highlight the
511 potential for further enhancement of modern rice cultivars.

512 Through IBD analysis, we revealed the breeding preferences and genetic
513 characteristics of each region. The NE region had accumulated the largest number of
514 IBD segments, likely reflecting the sustained emphasis on early maturity and cold
515 tolerance over prolonged breeding cycles, consistent with previous reports (Zhang et
516 al., 2014). Genes present in IBD segments appeared to be associated with local stress
517 factors. For instance, in the NC region, the salt-alkali tolerance genes were under strong
518 selection, whereas the SC region exhibited selection for disease and pest resistance
519 genes. In the CC region, genes associated with stress tolerance and grain quality were
520 subject to intensive selection, and in the SW region, multiple stress-tolerance and
521 quality-related genes were strongly favored.

522 The marker density and sample size used in this study were sufficient for the
523 detection of common high-effect alleles in the population. We identified a total of 3131
524 QTLs associated with key agronomic traits, providing insight into the genetic
525 architecture and locus co-localization of various traits. Among the identified QTLs,
526 16.6% were detected consistently across multiple locations or years, whereas most were
527 observed in a single environment. This pattern highlights strong environmental
528 specificity or genotype–environment interactions, offering valuable insights for future
529 rice adaptive breeding programs. Of the 3131 identified QTLs, 96 overlapped with
530 previously reported loci of corresponding quantitative trait genes. Over a thousand
531 QTLs were newly detected, from which we successfully cloned a novel gene
532 (*LOC_Os03g62060*) associated with grain length in rice. The discovery of numerous
533 loci associated with diverse agronomic traits provides a foundation for further genetic
534 improvement of rice through marker-assistant selection or genomic selection.

535 To fully leverage genetic variation and phenotypic information for accelerated
536 breeding improvement, we integrated multiple datasets to construct the comprehensive
537 *RiceAtlas* database. Existing public databases, such as RiceVarMapv2.0 (Zhao et al.,

2021; Zhao et al., 2015), MBKBASE (Peng et al., 2020), and Rice SNP-Seek (Mansueto et al., 2017), focus primarily on multi-omics data for fundamental research queries (e.g., genetic variants and gene expression data). They place less emphasis on large population sizes and the integration of genetic, phenotypic, and environmental data, making them somewhat less useful as "one-stop" platforms for design breeding or targeted crop improvement. To address these issues, Wei et al. (2021) constructed the RiceNavi system, based on 348 QTNs for 404 rice accessions, to enable rapid and precise breeding design, demonstrating its use for improvement of rice through pyramiding of favorable variants.

*RiceAtlas* complements and expands upon these existing tools by integrating a larger number of accessions, a broader range of phenotypic data collected in multiple environments, and many newly identified QTNs, including environment-specific QTNs, offering a user-friendly, multifunctional platform that operates across multiple scales. By incorporating sequencing data from multiple studies and accounting for donor phenotypes, regional adaptability, and background genetic similarity, *RiceAtlas* can help breeders to aggregate advantageous alleles, facilitating rapid genetic improvement. In addition, the genetic resources available at *RiceAtlas* support the training of GS models. A GS model for heading-date prediction is already available, further expanding the utility of *RiceAtlas* in breeding programs. As more GS models are developed for additional traits, *RiceAtlas* aims to become a powerful, yet user-friendly, intelligent platform for rice design breeding.

Nonetheless, the present study has some limitations. Owing to the complexity of environmental conditions, our multi-site phenotyping and GWAS analyses in five major rice-growing regions may not fully capture local microclimates (e.g., variations in photoperiod, temperature gradients, and altitude). Consequently, the identification of critical QTLs and a comprehensive understanding of cultivar adaptations to light, temperature, and altitude remain partially constrained. Although data were collected in multiple environments, we have not yet performed an in-depth investigation of environmental interactions. Limited information on gene–gene and gene–environment

interactions means that *RiceAtlas* currently supports only relatively simple, single-trait breeding designs. Nonetheless, our findings provide a genomic overview of the genetic improvements observed in modern cultivated rice across China's five major rice-growing regions, together with a rich repository of genetic variation. This work lays a solid foundation for revealing the molecular basis of advantageous rice traits and for devising more accurate and efficient genome-based breeding strategies.

## Materials and Methods

### Plant materials

The diversity panel used in this study comprised 6044 accessions from 25 provinces, municipalities, and autonomous regions, covering five rice-growing regions in China (SC, CC, SW, NC, and NE). The panel was curated based on the agroecological distributions of the cultivars, their cultivation acreages, and prior analyses of their phenotypic traits, genetic diversity, and nucleotide variation (Cui et al., 2022; Han et al., 2022; Liu et al., 2022; Liu et al., 2023). Of the 6044 accessions, 5164 were newly collected for this research, and 880 were selected from accessions previously reported by our laboratory (Cui et al., 2022; Han et al., 2022; Liu et al., 2023). We also included four wild rice accessions used to root the phylogenetic tree. To maintain a focus on modern cultivars, we excluded 822 landraces and 14 misclassified accessions, resulting in a final set of 5208 cultivars used for analyses of diversity, causative variants, and artificial selection. To enhance the diversity of donor parents for breeding tool development, we used the full set of 6044 accessions in the Breeding Design module of the *RiceAtlas* platform.

### Phenotyping

We selected 3606 of the 6044 cultivars for phenotyping. To systematically evaluate variations in agronomic traits across rice-growing regions, we grew all 3606 cultivars at seven field sites in core rice-cultivation areas that represented the five major rice-growing regions. We evaluated key quantitative traits over two consecutive growing seasons (2022–2023), including heading date and yield components. The field sites were located in GZL (124°44′ E, 43°27′ N), Jilin Province, representing the NE region; TH (118°17′ E, 39°18′ N), Hebei Province, representing the NC region; KM (103°6' E, 25°20' N), Yunnan Province, representing the SW region; and NN (108°11′ E, 22°48′ N), Guangxi Zhuang Autonomous Region, representing the SC region. Because the CC region accounts for nearly half of China’s rice cultivation area, three field sites were

established in this region: HF (117°12′ E, 31°48′ N), Anhui Province; HZ (119°94′ E, 30°8′ N), Zhejiang Province; and WH (114°2′ E, 30°42′ N), Hubei Province.

Accessions were planted in four-row plots, each containing 16 plants, with 26.7 cm between rows and 10 cm between plants. The alleys between the plots were 50 cm wide. An augmented design was used, consisting of 73 blocks (60 m × 1.5 m), each containing 50 entries across a total of 200 rows. The 73 blocks were divided into five field sections, each physically separated from the others by ridges. Within each field section, the blocks were further separated by Additionally, a 1.0-meter buffer zone was established around each field section to minimize edge effects. Standardized field management practices were used across all experimental blocks to ensure phenotypic consistency. Ten plants (excluding border plants) were randomly selected from each plot for phenotyping. Measurements included heading date, plant height, panicle length, tiller number, grain per panicle, seed-setting rate, 1000-grain weight, grain length, grain width, grain length-to-width ratio, and yield per plant, following the standard evaluation system for rice (Han et al., 2006). Heading date, plant height, panicle length, and tiller number were measured directly in the field. Heading date was recorded as the number of days from sowing to the emergence of 50% of the inflorescences above the flag-leaf sheath. The remaining grain-related traits, including grain number per panicle, seed-set rate, 1000-grain weight, grain length, grain width, grain length-to-width ratio, and yield per plant, were measured in the laboratory after harvest.

In addition to evaluating basic agronomic traits, we performed resistance assessments to identify quantitative-trait genes associated with disease and pest resistance, as well as stress tolerance, for use in rice breeding. These resistance traits, combined with data on basic agronomic traits, were used for GWAS analyses and for breeding design in the *RiceAtlas* platform. The detailed methods used to assess resistance to biotic and abiotic stresses (e.g., leaf blast, neck blast, bacterial blight, brown planthopper, southern rice black-streaked dwarf virus, sheath blight, drought, salt, cold, high temperature, and sprouting) are provided in the Supplemental Note.

To ensure the accuracy and reliability of the phenotypic data, we manually reviewed the data to identify and correct inconsistencies, such as decimal point errors, during data entry. Trait assessments were performed over two consecutive years to obtain fully representative phenotypic data. The mean and standard deviation (SD) were calculated for each trait, and outliers more than three SDs from the mean were excluded. Phenotypic data that were unavailable due to environmental factors were treated as missing values. The verified and cleaned dataset, free from outliers and invalid entries, was used as input for phenotypic and GWAS analyses.

**DNA isolation and genome sequencing**

Genomic DNA (1.5 μg per sample) was isolated following standard protocols and used to prepare sequencing libraries with the MGIEasy FS DNA Prep kit (BGI, China). Unique index codes were assigned to each sample. DNA was sonicated to an average fragment size of ~350 base pairs (bp), then end-polished, A-tailed, and ligated to full-length adapters, followed by PCR amplification. The PCR products were purified using the AMPure XP bead system. The library size distribution was evaluated using an Agilent 2100 Bioanalyzer, and library concentrations were quantified by real-time PCR. Sequencing was performed on the DNBSEQ-T7 platform, generating approximately 60.78 Tb of clean sequence data for the 5164 newly collected accessions as 150-bp paired-end reads.

**Sequence quality checking and filtering**

To minimize sources of artificial bias, such as low-quality paired reads caused by base-calling errors, duplicate reads, and adaptor contamination, we applied the filtering criteria used in a previous study (Li et al., 2022a). The following reads were excluded: (i) reads that contained ≥10% unidentified nucleotides (N); (ii) reads in which more than 10 nucleotides aligned to the adaptor, permitting ≤10% mismatches; (iii) reads in which more than 50% of bases had a Phred quality below 5; and (iv) potential PCR duplicates generated during library construction.

656 **Sequence alignment, variant calling, and annotation**

657 The retained high-quality paired-end reads were mapped to the rice *O. sativa* cv.
658 Nipponbare IRGPS 1.0 reference genome (Kawahara et al., 2013) using Burrows–
659 Wheeler Aligner (BWA) software (Li and Durbin, 2009) with the command 'mem -t 4
660 -k 32 –M'. To reduce PCR-induced mismatches, duplicate reads were removed with
661 SAMtools v0.1.1. Genomic variants were identified in GVCF format using the
662 HaplotypeCaller module from the Genome Analysis Toolkit (GATK) (McKenna et al.,
663 2010). The GVCF files were merged, and a raw population genotype file containing
664 SNPs and InDels was created. The data were filtered using the following criteria:
665 individual read depth ≥4, genotype quality ≥40, number of genotypes at each position
666 = 2, minor allele frequency (MAF) ≥0.01, and missing data rate ≤0.2. This resulted in
667 the identification of 5,694,922 SNPs and 812,306 Indels. These variants were annotated
668 using ANNOVAR software (version 2013-05-20) (Wang et al., 2010), categorizing
669 them by genomic location (intergenic regions, upstream/downstream of transcription
670 start/stop sites, coding sequences, and introns).

671 **Phylogenetic tree and population structure**

672 We assessed population genetic structure using the Bayesian clustering program
673 fastStructure v.1.0 (Raj et al., 2014). K values from 2 to 14 were tested to determine the
674 optimal subpopulation size based on the cross-validation error at the inflection point.
675 Principal component analysis was performed with GCTA software (Yang et al., 2011),
676 which generated a genetic relationship matrix using the '–make-grm' command. The
677 top three principal components were then estimated using '–pca3'. VCFtools v0.1.15
678 (Danecek et al., 2011) was used to calculate nucleotide diversity and the fixation index
679 in 10-kb sliding windows using 5-kb steps, enabling us to quantify genomic
680 differentiation across different rice-growing regions.

681 **Linkage-disequilibrium analysis**

682 To evaluate the pattern of linkage disequilibrium (LD), we calculated the squared
683 correlation coefficient ($r^2$) between pairwise SNPs using PopLDdecay (Zhang et al.,

2019), with parameters set to '-MaxDist 1000kb'. Average $r^2$ values were computed for pairwise markers in 10-kb windows and then averaged across the genome.

**GWAS analyses**

GWAS analyses were performed separately for 212 datasets containing data for 53 phenotypes using EMMAX software (Kang et al., 2010) with all 5,694,922 high-quality SNPs and 812,306 high-quality Indels. A kinship matrix, derived from pairwise genetic similarities, was used as the variance–covariance matrix for random effects. To correct for population stratification, the top ten principal components (PC1–PC10) were used for GWAS with all accessions, the top thirty principal components (PC1–PC30) for *japonica* rice accessions, and the top thirty-five principal components (PC1–PC35) for *indica* rice accessions. The number of independent SNPs was estimated to be 1,477,136, and the genome-wide significance threshold was determined using a Bonferroni correction ($\alpha$ = 1). Candidate regions were then expanded to 100 kb centered on the GWAS signal peaks to identify candidate genes.

**Lead SNP calculation**

Genome-wide blocks were defined using PLINK v1.9 software (Purcell et al., 2007) with the parameters '--blocks --blocks-strong-lowci 0.70 --blocks-strong-highci 0.98', following the approach described by (Cervantes-Perez et al., 2023). Multiple SNPs within each block that exceeded the threshold were clustered, and the SNP with the lowest *P*-value in each cluster was identified as the lead SNP. Independent SNPs that exceeded the threshold but were not located within a block were retained.

**Identification of genotypes with favorable alleles**

On the basis of the lead SNP at each locus, the allele type (reference allele or alternative allele) that conferred better agronomic performance (for example, higher values for panicle length, tiller number, grain per panicle, grain length, grain width, seed-setting rate, 1000-grain weight, or yield per plant) was defined as the favorable allele. To minimize the influence of confounding factors, lead SNPs that were linked to the same

trait but exhibited different favorable genotypes in different locations or years were excluded from consideration. We used the R package *lme4* (Bates et al., 2015) to compute the phenotypic variance accounted for by each lead SNP.

**Selective sweep identification**

To detect potential selective sweeps between different rice-growing regions, we analyzed genetic differentiation between populations ($F_{ST}$) and diversity ($\pi$) within populations. Candidate outliers, indicative of selective sweeps, were identified as the top 5% of log2 ($\pi$ ratio) values and $F_{ST}$ values.

**Data preprocessing for Genomic selection**

We excluded 51 samples with missing heading-date phenotypic values from 3606 accessions and retained 3555 samples. These samples were randomly divided into a training set of 2844 samples and a test set of 711 samples in an 8:2 ratio. During data splitting process, we set a random seed to prevent the emergence of specific patterns or correlations between different subsets of the dataset, ensuring the representativeness of the training and testing sets.

Based on the 28 alleles associated with the heading-date as input features to the model to perform training. For machine learning, genotypic data should first be converted to numeric features. We use PLINK to encode SNP information as 0, 1, and 2, where 0 represents the homozygous genotype (AA) of the two major alleles, 1 represents the heterozygous genotype (AB) of one major and one minor allele, and 2 represents the homozygous genotype (BB) of the two minor alleles.

**Model training and evaluation**

LightGBM has been demonstrated by Yan et al. to be effective in genomic selection GS-assisted breeding (Yan et al., 2021). We employed LightGBM to construct GS model for predicting heading-date. To further optimize the LightGBM model, we

736 utilized a grid search to determine the optimal hyperparameters. The source code of
737 LightGBM is freely available on GitHub: https://github.com/jiekesen/Lightgbm.

738 Cross-validation is a commonly used technique for assessing the results of
739 statistical analysis. It can be used to objectively evaluate the predictive performance of
740 a model. In this study, we use 10-fold cross-validation to assess the predictive
741 performance of the model on the training set and the generalization ability of the model
742 using the testing set. The 10-fold cross-validation was repeated for 100 runs. The
743 Pearson correlation coefficient is used to assess the predictive performance of the model.
744 A coefficient closer to 1 indicates a higher predictive accuracy of the model.

745 **Gene Ontology analysis**

746 Gene Ontology (GO) annotations for rice genes were obtained from the Ensembl Plants
747 Genes database (https://plants.ensembl.org/biomart/martview/). GO enrichment
748 analysis was performed using agriGO v.2.0 (Tian et al., 2017) with significance
749 determined by Fisher's exact test. Enrichment results with more than five annotations
750 and a Bonferroni-corrected false discovery rate of <0.05 were visualized using the R
751 package ClusterProfiler v.3.10.0 (Yu et al., 2012).

752 **Pairwise identity-by-descent detection**

753 All SNPs were used to identify pairwise shared haplotypes across different groups using
754 IBD analysis as described previously (Bosse et al., 2014); with minor modifications.
755 The approach involves two main steps: identifying pairwise IBD regions and
756 calculating shared haplotype frequencies. First, all individuals were phased using the
757 fastPhase function in Beagle v5.4 (Browning and Browning, 2007). Pairwise shared
758 haplotypes were extracted using the Beagle RefinedIBD function (Browning and
759 Browning, 2013). Second, to characterize the frequency of shared haplotypes along
760 each chromosome, the genome was divided into 50-kb bins, and the number of recorded
761 IBD segments between different groups was quantified for each bin. The bins were then
762 ranked based on the number of IBD segments, with the top 20% identified as candidate
763 regions. To further enhance the confidence of the analysis, genetic diversity ($\pi$) was

764 introduced to correct for potential false positives in high-frequency IBD segments. The
765 top 20% of bins with the lowest genetic diversity were selected and compared with the
766 top 20% of bins ranked by IBD segment counts, and their intersection was defined as
767 the set of candidate bin regions.

768 **Plasmid construction**

769 Our GWAS analysis identified *OsGL3.6* as a high-confidence candidate gene associated
770 with grain length. This locus, located on chromosome 3 (Chr03: 35,155,927) within a
771 large LD block (60 kb), exhibited a strong association signal in Manhattan plots. We
772 therefore performed gene editing of *OsGL3.6* using the CRISPR/Cas9 method.
773 Specifically, *OsGL3.6* targeting sequences were amplified and inserted into the
774 pYLgRNA-OsU6 vector as described by (Ying et al., 2018). These constructs were
775 confirmed by DNA sequencing and introduced into *Agrobacterium tumefaciens* strain
776 EHA105 for *Agrobacterium*-mediated transformation into the *O. sativa* ssp. *japonica*
777 cultivar 'ZH11'. Homozygous $T_2$ seeds from all transgenic plants were used for
778 subsequent analyses. All primers used are listed in Supplementary Table 23.

779 **Breeding tools**

780 To enhance specific traits in a given sample or breeding line, the selection of appropriate
781 donor parents is essential. By analyzing a sample's allelic variants and integrating allele
782 effects, the *RiceAtlas* breeding design module recommends donor parents that
783 complement the sample's unfavorable alleles, align with desired phenotypic traits, and
784 exhibit the highest genetic similarity to the sample. These selected donor parents can
785 accelerate the fixation of segregating loci in the offspring population, thus expediting
786 the breeding process.

787 In the breeding design module of the *RiceAtlas* platform, the recommendation of
788 donor parents involves four steps: (1) Upload Genome Resequencing Data. Users
789 upload the genome resequencing data for the target sample, which the system
790 automatically standardizes. Using published quantitative trait genes (QTGs) and trait-
791 associated loci identified in this study, the system analyzes genotypes associated with

key agronomic traits and identifies QTNs that can be replaced with favorable alleles. (2) Select Target Traits for Improvement. Users select traits for improvement; these are categorized as core traits—supported by datasets for 53 phenotypes—or extended traits, which lack phenotypic data support. For core traits, users can specify initial phenotypic values as thresholds for donor-parent recommendations. (3) Calculate Improvable QTNs. On the basis of the selected target traits, the system integrates reported QTNs and newly identified loci to determine which QTNs in the sample can be improved, presenting such loci in a list format. (4) Recommend donor parents. Donor parents are recommended from three sources, totaling 9458 accessions: 6K-Rice accessions categorized by rice-growing regions, the 3K-Rice panel of 3010 Asian cultivated rice accessions, and the *RiceNavi* database resources, which include 404 diverse accessions. When the user specifies the donor resources, the system automatically recommends suitable donor parents based on the results of Step 3 and displays detailed information in a list format. The system also calculates genetic similarity between the recommended donor parents and the target sample using a whole-genome fingerprint map of 924 SNPs (Ma et al., 2023b). The built-in sorting function enables users to sort by genetic similarity and phenotypic value, facilitating the efficient selection of ideal donor parents.

## SUPPLEMENTAL INFORMATION

Supplemental information is available at Molecular Plant Online.

## FUNDING

This work was supported by the National Key Research and Development Program of China (2021YFD1200500), the Biological Breeding-National Science and Technology Major Project (2022ZD04017), the Biological Breeding-Major Projects (2023ZD04076), the National Natural Science Foundation of China (32371996), and the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences.

## AUTHOR CONTRIBUTIONS

Q.H., W.F., X.D.M. and L.Z.H. conceived and designed the research. X.D.M., H.W., Q.H., S.Y., X.J. and C.Q.Z. conducted the data analysis and drafted the manuscript. K.N.Z., Q.Z., M.M.L., Y.L.Y., D.T.L., P.S., C.F.T., L.Y.G., J.C.S., Z.Y.J., X.J.S., Y.L.Z., P.Z., D.C. and B.H. carried out field planting and phenotypic surveys. All authors read and approved the final manuscript.

## ACKNOWLEDGMENTS

We thank Dr. Wenbin Zhou and Dr. Wensheng Wang of the Institute of Crop Sciences (CAAS) for critical comments and advice; Prof. Yongming Chen of Peking University of China for helpful discussions; Ph.D Pengfei Liang, Ms. Jiaxin Zhu of the Institute of Crop Sciences (CAAS) and Mr. Jingpeng Hong of Henan Agricultural University for their assistance with the photo processing and technical support in server and network matters. We are grateful to the National Medium-Term Gene Bank, Institute of Crop Sciences and China National Rice Research Institute, Chinese Academy of Agricultural Sciences, for providing the rice seeds.

## Data availability

The raw genome sequencing data were deposited in the Genome Sequence Archive (https://bigd.big.ac.cn/gsa) under the accession code PRJCA034255 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA034255). Download links for the imputed SNP data can be found in the "About/Download" section of https://www.cgris.net/RiceAtlas. The phenotype dataset of the heading date has been included in the Supplementary Table 5, and other phenotypic datasets used in GWAS and GS studies have been deposited in the Science Data Bank public database (https://doi.org/10.57760/sciencedb.agriculture.00211). At present, readers may access the provided link to review the data introduction and associated metadata. Upon completion of the one-year embargo period, all phenotypic data generated in this study will be publicly available and freely accessible through this link.

## References


**Bates, D., Mächler, M., Bolker, B., and Walker, S.** (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software **67**:1 - 48. 10.18637/jss.v067.i01.

**Bosse, M., Megens, H.J., Frantz, L.A., Madsen, O., Larson, G., Paudel, Y., Duijvesteijn, N., Harlizius, B., Hagemeijer, Y., Crooijmans, R.P., et al.** (2014). Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nat Commun **5**:4392. 10.1038/ncomms5392.

**Browning, B.L., and Browning, S.R.** (2007). Documentation for BEAGLE 2.1.

**Browning, B.L., and Browning, S.R.** (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics **194**:459-471. 10.1534/genetics.113.150029.

**Cai, M., Zhu, S., Wu, M., Zheng, X., Wang, J., Zhou, L., Zheng, T., Cui, S., Zhou, S., Li, C., et al.** (2021). DHD4, a CONSTANS-like family transcription factor, delays heading date by affecting the formation of the FAC complex in rice. Mol Plant **14**:330-343. 10.1016/j.molp.2020.11.013.

**Campbell, M.T., Du, Q., Liu, K., Sharma, S., Zhang, C., and Walia, H.** (2020). Characterization of the transcriptional divergence between the subspecies of cultivated rice (Oryza sativa). BMC Genomics **21**:394. 10.1186/s12864-020-06786-6.

**Cervantes-Perez, S.A., Thibivilliers, S., Laffont, C., Farmer, A.D., Frugier, F., and Libault, M.** (2023). Cell-specific pathways recruited for symbiotic nodulation in the Medicago truncatula legume. Mol Plant **16**:481-483. 10.1016/j.molp.2023.01.002.

**Cui, D., Zhou, H., Ma, X., Lin, Z., Sun, L., Han, B., Li, M., Sun, J., Liu, J., Jin, G., et al.** (2022). Genomic insights on the contribution of introgressions from Xian/Indica to the genetic improvement of Geng/Japonica rice cultivars. Plant Commun **3**:100325. 10.1016/j.xplc.2022.100325.

**Dai, X., Ding, Y., Tan, L., Fu, Y., Liu, F., Zhu, Z., Sun, X., Sun, X., Gu, P., Cai, H., et al.** (2012). LHD1, an allele of DTH8/Ghd8, controls late heading date in common wild rice (Oryza rufipogon). J Integr Plant Biol **54**:790-799. 10.1111/j.1744-7909.2012.01166.x.

**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. 10.1093/bioinformatics/btr330.

**Ding, Y.** (1961). Rice Cultivation in China (Beijing: China Agricul. Press).

**Dong, X., Zhang, T., Yang, X., and Li, T.** (2023). Breeding priorities for rice adaptation to climate change in Northeast China. Climatic Change **176**:75. 10.1007/s10584-023-03556-7.

**Han, B., Cui, D., Ma, X., Cao, G., Zhang, H., Koh, H.J., and Han, L.** (2022). Evidence for evolution and selection of drought-resistant genes based on high-throughput resequencing in weedy rice. Journal of experimental botany **73**:1949-1962. 10.1093/jxb/erab515.

**Han, L., Wei, X., Cao, G., Yu, H., and Zhang, Y.** (2006). Descriptors and data standard for rice (Oryza sativa L.). China Agriculture Press **4**:70-71.

**He, Z.-W., and Tang, B.-H.** (2023). Spatiotemporal change patterns and driving factors of land surface temperature in the Yunnan-Kweichow Plateau from 2000 to 2020. Science of The Total Environment **896**:165288. https://doi.org/10.1016/j.scitotenv.2023.165288.

**Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C., et al.** (2011). Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nature genetics **44**:32-39. 10.1038/ng.1018.

**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. Nat Genet **42**:348-354. 10.1038/ng.548.

**Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al.** (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) **6**:4. 10.1186/1939-8433-6-4.

**Li, C., Guan, H., Jing, X., Li, Y., Wang, B., Li, Y., Liu, X., Zhang, D., Liu, C., Xie, X., et al.** (2022a). Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding. Nat Plants **8**:750-763. 10.1038/s41477-022-01190-2.

**Li, H., and Durbin, R.** (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754-1760. 10.1093/bioinformatics/btp324.

**Li, X., Tian, X., He, M., Liu, X., Li, Z., Tang, J., Mei, E., Xu, M., Liu, Y., Wang, Z., et al.** (2022b). bZIP71 delays flowering by suppressing Ehd1 expression in rice. J Integr Plant Biol **64**:1352-1363. 10.1111/jipb.13275.

**Liu, C., Cui, D., Jiao, A., Ma, X., Li, X., Han, B., Chen, H., Ruan, R., Wang, Y., and Han, L.** (2022). Kam Sweet Rice (Oryza sativa L.) Is a Special Ecotypic Rice in Southeast Guizhou, China as Revealed by Genetic Diversity Analysis. Frontiers in plant science **13**:830556. 10.3389/fpls.2022.830556.

**Liu, C., Wang, T., Chen, H., Ma, X., Jiao, C., Cui, D., Han, B., Li, X., Jiao, A., Ruan, R., et al.** (2023). Genomic footprints of Kam Sweet Rice domestication indicate possible migration routes of the Dong people in China and provide resources for future rice breeding. Molecular plant **16**:415-431. 10.1016/j.molp.2022.12.020.

**Liu, J., Chen, J., Zheng, X., Wu, F., Lin, Q., Heng, Y., Tian, P., Cheng, Z., Yu, X., Zhou, K., et al.** (2017). GW5 acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. Nature Plants **3**:17043. 10.1038/nplants.2017.43.

**Lv, Z., Zhu, Y., Liu, X., Ye, H., Tian, Y., and Li, F.** (2018). Climate change impacts on regional rice production in China. Climatic Change **147**:523-537. 10.1007/s10584-018-2151-0.

**Ma, M., Shen, S.Y., Bai, C., Wang, W.Q., Feng, X.H., Ying, J.Z., and Song, X.J.** (2023a). Control of grain size in rice by TGW3 phosphorylation of OsIAA10 through potentiation of OsIAA10-OsARF4-mediated auxin signaling. Cell Rep **42**:112187. 10.1016/j.celrep.2023.112187.

**Ma, X., Cui, D., Han, B., Jiao, c., and Han, L.** (2023b). Identification and Evaluation Method for Genome-wide DNA Fingerprinting of Rice Germplasm. Journal of Plant Genetic Resources **24**:1106-1113.

**Mansueto, L., Fuentes, R.R., Borja, F.N., Detras, J., Abriol-Santos, J.M., Chebotarov, D., Sanciangco, M., Palis, K., Copetti, D., Poliakov, A., et al.** (2017). Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res **45**:D1075-D1081. 10.1093/nar/gkw1135.

**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 Res **20**:1297-1303. 10.1101/gr.107524.110.

**Nie, L., and Peng, S.** (2017). Rice Production in China. In Rice Production Worldwide, B.S. Chauhan and K. Jabran and G. Mahajan, eds. (Springer International Publishing: Cham), pp. 33-52. 10.1007/978-3-319-47516-5_2.

**Peng, H., Wang, K., Chen, Z., Cao, Y., Gao, Q., Li, Y., Li, X., Lu, H., Du, H., Lu, M., et al.** (2020). MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. Nucleic Acids

Res **48**:D1085-D1092. 10.1093/nar/gkz921.

**Peng, Q., Zhu, C., Liu, T., Zhang, S., Feng, S., and Wu, C.** (2021). Phosphorylation of OsFD1 by OsCIPK3 promotes the formation of RFT1-containing florigen activation complex for long-day flowering in rice. Mol Plant **14**:1135-1148. 10.1016/j.molp.2021.04.003.

**Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al.** (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**:559-575. 10.1086/519795.

**Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., et al.** (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell **184**:3542-3558 e3516. 10.1016/j.cell.2021.04.046.

**Raj, A., Stephens, M., and Pritchard, J.K.** (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics **197**:573-589. 10.1534/genetics.114.164350.

**Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.C., Iwamoto, M., Abe, T., et al.** (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol **54**:e6. 10.1093/pcp/pcs183.

**Sasaki, T., and Burr, B.** (2000). International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. Curr Opin Plant Biol **3**:138-141. 10.1016/s1369-5266(99)00047-3.

**Saud, S., Wang, D., Fahad, S., Alharby, H.F., Bamagoos, A.A., Mjrashi, A., Alabdallah, N.M., AlZahrani, S.S., AbdElgawad, H., Adnan, M., et al.** (2022). Comprehensive Impacts of Climate Change on Rice Production and Adaptive Strategies in China. Front Microbiol **13**:926059. 10.3389/fmicb.2022.926059.

**Shang, L., He, W., Wang, T., Yang, Y., Xu, Q., Zhao, X., Yang, L., Zhang, H., Li, X., Lv, Y., et al.** (2023). A complete assembly of the rice Nipponbare reference genome. Mol Plant **16**:1232-1236. 10.1016/j.molp.2023.08.003.

**Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H., Hu, M., Zhao, F., Zhang, C., et al.** (2022). A super pan-genomic landscape of rice. Cell Res **32**:878-896. 10.1038/s41422-022-00685-z.

**Takahashi, Y., Teshima, K.M., Yokoi, S., Innan, H., and Shimamoto, K.** (2009). Variations in Hd1 proteins, Hd3a promoters, and Ehd1 expression levels contribute to diversity of flowering time in cultivated rice. Proc Natl Acad Sci U S A **106**:4555-4560. 10.1073/pnas.0812092106.

**Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., and Su, Z.** (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res **45**:W122-W129. 10.1093/nar/gkx382.

**Wang, J., Yu, H., Xiong, G., Lu, Z., Jiao, Y., Meng, X., Liu, G., Chen, X., Wang, Y., and Li, J.** (2017). Tissue-Specific Ubiquitination by IPA1 INTERACTING PROTEIN1 Modulates IPA1 Protein Levels to Regulate Plant Architecture in Rice. The Plant cell **29**:697-707. 10.1105/tpc.16.00879.

**Wang, K., Li, M., and Hakonarson, H.** (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res **38**:e164. 10.1093/nar/gkq603.

**Wang, S., Wu, K., Yuan, Q., Liu, X., Liu, Z., Lin, X., Zeng, R., Zhu, H., Dong, G., Qian, Q., et al.** (2012). Control of grain size, shape and quality by OsSPL16 in rice. Nature genetics **44**:950-954. 10.1038/ng.2327.

**Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., et al.** (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice.

Nature **557**:43-49. 10.1038/s41586-018-0063-9.

**Wei, X., Xu, J., Guo, H., Jiang, L., Chen, S., Yu, C., Zhou, Z., Hu, P., Zhai, H., and Wan, J.** (2010). DTH8 suppresses flowering in rice, influencing plant height and yield potential simultaneously. Plant Physiol **153**:1747-1758. 10.1104/pp.110.156943.

**Wei, X., Qiu, J., Yong, K., Fan, 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. 10.1038/s41588-020-00769-9.

**Wei, X., Chen, M., Zhang, Q., Gong, J., Liu, J., Yong, K., Wang, Q., Fan, J., Chen, S., Hua, H., et al.** (2024). Genomic investigation of 18,421 lines reveals the genetic architecture of rice. Science **385**:eadm8762. 10.1126/science.adm8762.

**Xian, Y., Cai, G., Lin, J., Chen, Y., and Wang, X.** (2023). Comparison of crop productivity, economic benefit and environmental footprints among diversified multi-cropping systems in South China. Science of The Total Environment **874**:162407. https://doi.org/10.1016/j.scitotenv.2023.162407.

**Xiao, N., Pan, C., Li, Y., Wu, Y., Cai, Y., Lu, Y., Wang, R., Yu, L., Shi, W., Kang, H., et al.** (2021). Genomic insight into balancing high yield, good quality, and blast resistance of japonica rice. Genome Biol **22**:283. 10.1186/s13059-021-02488-8.

**Xue, W., Xing, Y., Weng, X., Zhao, Y., Tang, W., Wang, L., Zhou, H., Yu, S., Xu, C., Li, X., et al.** (2008). Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. Nat Genet **40**:761-767. 10.1038/ng.143.

**Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., and Wang, X.** (2021). LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol **22**:271. 10.1186/s13059-021-02492-y.

**Yan, W.H., Wang, P., Chen, H.X., Zhou, H.J., Li, Q.P., Wang, C.R., Ding, Z.H., Zhang, Y.S., Yu, S.B., Xing, Y.Z., et al.** (2011). A major QTL, Ghd8, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. Mol Plant **4**:319-330. 10.1093/mp/ssq070.

**Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M.** (2011). GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet **88**:76-82. 10.1016/j.ajhg.2010.11.011.

**Ye, J., Zhang, M., Yuan, X., Hu, D., Zhang, Y., Xu, S., Li, Z., Li, R., Liu, J., Sun, Y., et al.** (2022). Genomic insight into genetic changes and shaping of major inbred rice cultivars in China. New Phytologist **236**:2311-2326. 10.1111/nph.18500.

**Ying, J.Z., Ma, M., Bai, C., Huang, X.H., Liu, J.L., Fan, Y.Y., and Song, X.J.** (2018). TGW3, a Major QTL that Negatively Modulates Grain Length and Weight in Rice. Molecular plant **11**:750-753. 10.1016/j.molp.2018.03.007.

**Yu, G., Wang, L.G., Han, Y., and He, Q.Y.** (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS **16**:284-287. 10.1089/omi.2011.0118.

**Zhang, C., Dong, S.S., Xu, J.Y., He, W.M., and Yang, T.L.** (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics **35**:1786-1788. 10.1093/bioinformatics/bty875.

**Zhang, F., Xue, H., Dong, X., Li, M., Zheng, X., Li, Z., Xu, J., Wang, W., and Wei, C.** (2022). Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. Genome Res **32**:853-863. 10.1101/gr.276015.121.

**Zhang, Q., Chen, Q., Wang, S., Hong, Y., and Wang, Z.** (2014). Rice and cold stress: methods for its evaluation and summary of cold tolerance-related quantitative trait loci. Rice (N Y) **7**:24. 10.1186/s12284-014-0024-3.

**Zhao, H., Yao, W., Ouyang, Y., Yang, W., Wang, G., Lian, X., Xing, Y., Chen, L., and Xie, W.** (2015). RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res **43**:D1018-1022. 10.1093/nar/gku894.

**Zhao, H., Li, J., Yang, L., Qin, G., Xia, C., Xu, X., Su, Y., Liu, Y., Ming, L., Chen, L.L., et al.** (2021). An inferred functional impact map of genetic variants in rice. Mol Plant **14**:1584-1599. 10.1016/j.molp.2021.06.025.

**Zhou, S., Zhu, S., Cui, S., Hou, H., Wu, H., Hao, B., Cai, L., Xu, Z., Liu, L., Jiang, L., et al.** (2021). Transcriptional and post-transcriptional regulation of heading date in rice. New Phytol **230**:943-956. 10.1111/nph.17158.

**Zong, W., Ren, D., Huang, M., Sun, K., Feng, J., Zhao, J., Xiao, D., Xie, W., Liu, S., Zhang, H., et al.** (2021). Strong photoperiod sensitivity is controlled by cooperation and competition among Hd1, Ghd7 and DTH8 in rice heading. The New phytologist **229**:1635-1649. 10.1111/nph.16946.

**Figure legends:**

**Figure 1. Geographic distribution and phenotypic variation of 6044 rice accessions across five major rice-growing regions in China.**

(A) Geographic distribution of 6044 rice accessions, with different colors representing the six rice-growing regions: NE, NC, CC, SW, SC, and NW. Field sites used for phenotypic evaluation are indicated by red symbols and included Gongzhuling (GZL) for NE; Tanghai (TH) for NC; Hefei (HF), Wuhan (WH), and Hangzhou (HZ) for CC; Kunming (KM) for SW; and Nanning (NN) for SC. Numbers represent the number of accessions collected from each Province in China. No accessions were collected from the NW growing region.

(B) Numbers of *indica* and *japonica* accessions collected from five growing regions. The NE and NC regions grow only *japonica* rice, the SC region grows only *indica* rice, and the SW and CC regions grow both *indica* and *japonica* varieties.

(C – F) Phenotypic distributions of (C) heading date, (D) grain per panicle, (E) 1000-grain weight, and (F) yield per plant for 3606 local accessions grouped by their collection region and phenotyped at field site closest to their collection location. The upper and lower boundaries of each box represent the 25th and 75th percentiles, the horizontal line indicates the median, whiskers represent 1.5× the interquartile range, and dots outside the whiskers are outliers. Different letters indicate significant differences ($P < 0.05$, Least Significant Difference). Colors represent the seven accession groups. *Indica* and *japonica* varieties from the CC and SW regions were analyzed separately and are denoted as CC-I/SW-I and CC-J/SW-J.

**Figure 2. Genetic diversity and population differentiation among the rice accessions analyzed in this study.**

(A) Phylogenetic tree of 6044 *O. sativa* accessions and four wild rice accessions constructed using whole-genome SNPs. The four wild rice accessions were used to root the phylogenetic tree.

(B) Uniform Manifold Approximation and Projection (UMAP) plots showing the first two components for 5208 accessions from seven regional groups.

(C) Nucleotide diversity ($\pi$) and population divergence ($F_{ST}$) between different groups of accessions. Values of $\pi$ are displayed as a histogram, and values of $F_{ST}$ are shown as a heat map.

(D) Genome-wide average linkage disequilibrium decay for different groups of accessions.

**Figure 3. Causative variants associated with heading date and other agronomic trait QTGs across seven groups.**

(A) Combinations of 47 alleles from 23 QTGs associated with heading date were compared across seven accession groups; only the 28 alleles from 19 QTGs that showed allelic variation among the groups are displayed. ref, homozygous reference allele; alt, homozygous alternative allele; het, heterozygous; del, deletion.

(B) Heading dates of accessions carrying the top five allele combinations of heading date QTGs for each of the seven accession groups recorded at the HF field site in 2023. The x-axis labels (beginning with C) indicate the top five allele combinations for each group of accessions, followed by percentages indicating the prevalence of each combination in that group.

(C) Favorable allele frequencies for 233 alleles of 152 QTGs associated with key agronomic traits were compared across the seven accession groups; only the 116 alleles of 96 QTGs that showed allelic variation among the groups are displayed. favor, favorable allele frequency; infer, inferior allele frequency.

**Figure 4. Patterns of artificial selection in seven groups of regional accessions.**

(A) Analysis of genetic diversity and IBD along the 12 rice chromosomes in different accession groups. The colored line graphs show genetic diversity ($\pi$), and the heatmaps show IBD frequency, with a darker red color indicating regions of higher IBD frequency. Dashed lines indicate the physical locations of 77 known functional QTGs, with

different colors representing specific categories of agronomic traits. Overlap between high-frequency IBD regions and regions of low genetic diversity suggests that these regions have undergone strong selection during breeding programs in the different accession groups.

(B) Analysis of genetic diversity and IBD along chromosome 12 for the different accession groups.

(C) A pie chart illustrates the proportion of total IBD segments contributed by each accession group, with the "Shared" segment indicating that 40.28% of the IBD segments were shared among two or more groups. The upset plot provides a detailed representation of the overlap in IBD segments among groups; each column represents a set of IBD segments contained in one or more groups, as indicated by the connected dots below.

(D) GO enrichment analysis of genes within IBD segments in the NE group. The intensity of the circle color indicates the significance of enrichment (*P*-value, calculated using a two-sided Fisher's exact test), with darker colors indicating higher significance. The size of the circle reflects the frequency of the GO term among the annotated genes. The spatial arrangement of terms in the semantic space does not have a specific meaning but is designed to visually separate the different GO terms for clarity.

(E) Haplotype display for *GW5* and *GW8* in different accession groups. The analyzed intervals include the gene coding region and the 3-kb regions upstream and downstream of the gene.

**Figure 5. Large-scale GWAS and identification of the novel gene *OsGL3.6* through integration of GWAS results with selective sweep and IBD analyses.**

(A) Phenotype data were collected for all 3606 accessions planted in 19 geographic locations over one or two years. Numbers in parentheses indicate the number of years and the number of traits evaluated at each location.

(B) Phenotypic variation in 53 phenotypic traits across 19 geographic locations. Different letters in the heatmap indicate significant differences ($P < 0.05$) determined

by two-way ANOVA followed by Duncan's multiple comparison test. Heatmap colors represent scaled phenotype values. Phenotypes 1 to 53 are described in Supplementary Table 4.

(C) Combined Manhattan plot from separate GWAS analyses of the 212 datasets; the phenotypes were classified into four categories (abiotic stress, biotic stress, yield components, and heading date), each represented by a different color. The horizontal dashed lines indicate the genome-wide significance thresholds for GWAS ($10^{-6.2}$).

(D) Local Manhattan plot from the GWAS analysis for grain length on chromosome 3. The red dashed line indicates the Bonferroni-corrected significance threshold ($\alpha = 1$), and the arrow highlights a significant SNP within the *qGL3.6* region (chromosome 3: 35,155,927).

(E) The π-ratio (upper plot) and locations of high-frequency IBD windows (lower plot) in the *qGL3.6* region for different accession groups. The *qGL3.6* locus is present in a region that contains multiple π-ratio peaks and high-frequency IBD windows, particularly in the SW-J and CC-J accessions, as highlighted by the dashed line.

(F) *OsGL3.6* haplotypes spanning the 1.5-kb promoter region and the coding sequence (excluding synonymous SNPs). Only SNPs supported by at least ten samples were included in the analysis.

(G) *OsGL3.6* haplotype frequency across different groups.

(H and I) Grain lengths of the three *OsGL3.6* haplotypes in *indica* and *japonica* rice (H) and in various rice accession groups (I). Letters above the boxes indicate significant differences within subspecies (H) or among groups (I) ($P < 0.05$, Bonferroni correction).

(J) Functional validation of *OsGL3.6* using CRISPR-Cas9 gene editing. Information on the target sites and protospacer adjacent motif (PAM) sequences is shown at the top. Mature grains from the *osgl3.6* knockout mutants are shown below. Scale bar, 3 mm.

(K) Grain lengths of ZH11 and the *osgl3.6* mutants ($n = 10$). Bars represent the mean ± SD, and *P*-values were calculated using a two-tailed Student's *t*-test.

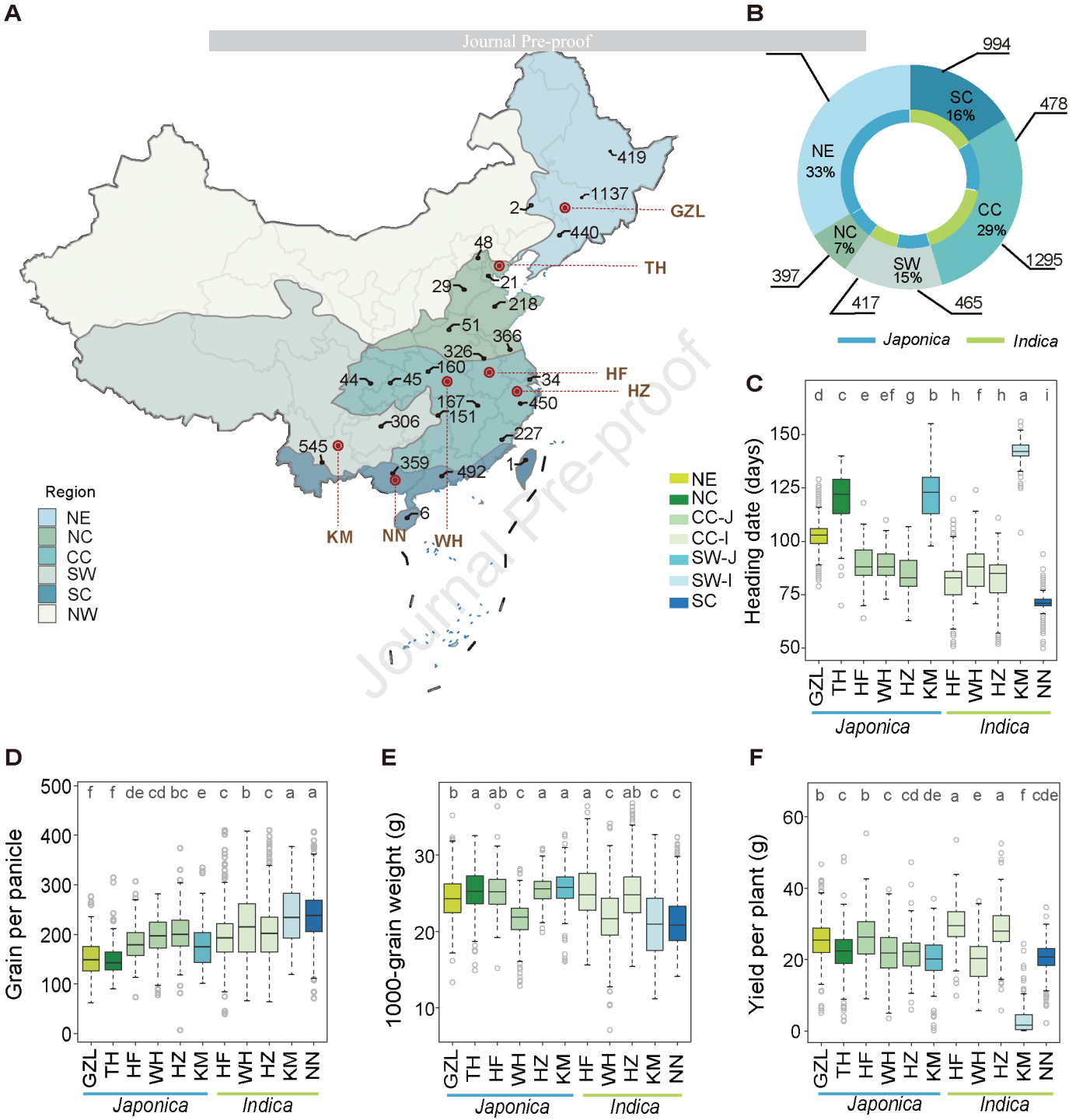
**Figure 6. Development of the *RiceAtlas* rice breeding database and improvement of the SG4 cultivar using the *RiceAtlas* breeding design tool.**

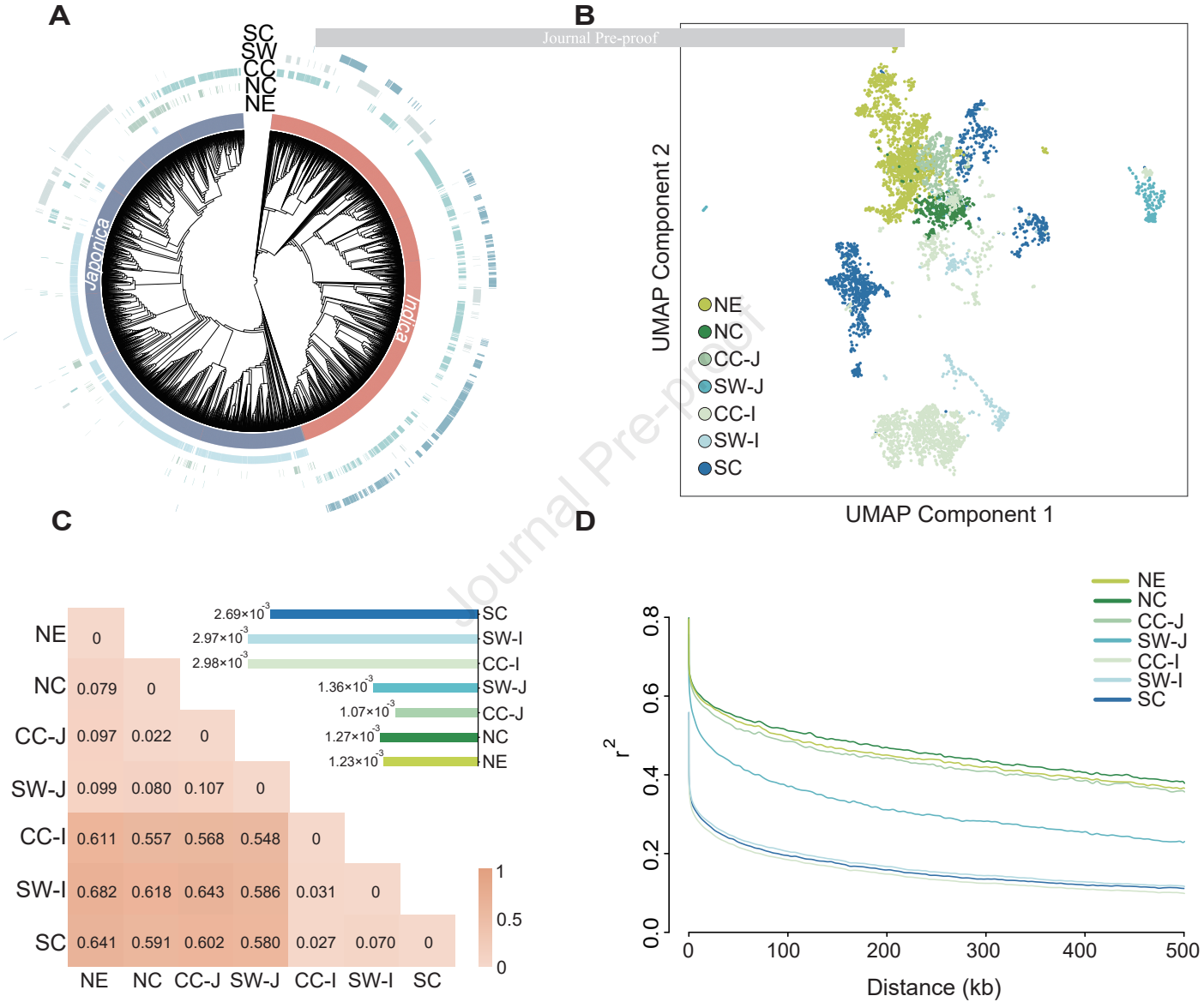
(A) *RiceAtlas* integrates data from the 6044 accessions analyzed here with published QTNs and data from 3414 additional rice genotypes, creating a comprehensive allele and germplasm library. It provides five analytical functions: Germplasm, Phenotype, GWAS, Variation, and Breeding.

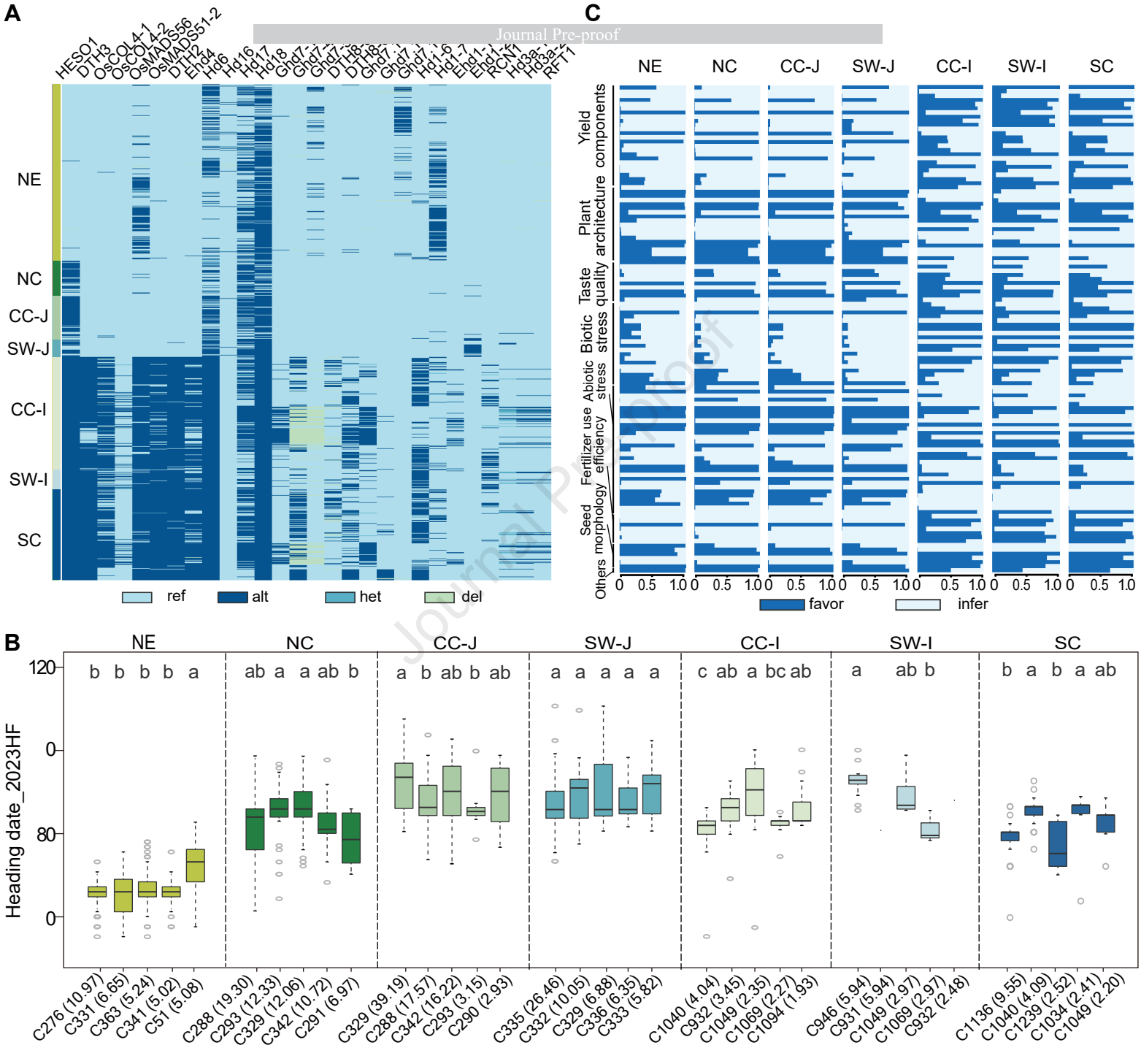
(B) To improve grain length while preserving the desirable traits of SG4, we used the *RiceAtlas* Breeding module to compare the SG4 genotype with accessions in the germplasm library. ZKF8 was identified as the optimal donor parent, as it contained long-grain alleles for two QTGs (*GS3* and *GW5*) and exhibited high genetic similarity to SG4. ZKF8 is also well-suited for cultivation in the same rice-growing region as SG4. SG4 has the GG genotype at the causative site of the *GS3* gene (Chr.3: 16,733,441) and the AA genotype at the causative site of the *GW5-1* gene (Chr.5: 5,365,256), both of which are associated with a short-grain effect. By contrast, ZKF8 has the TT and GG genotypes at these sites, which are associated with a long-grain effect.

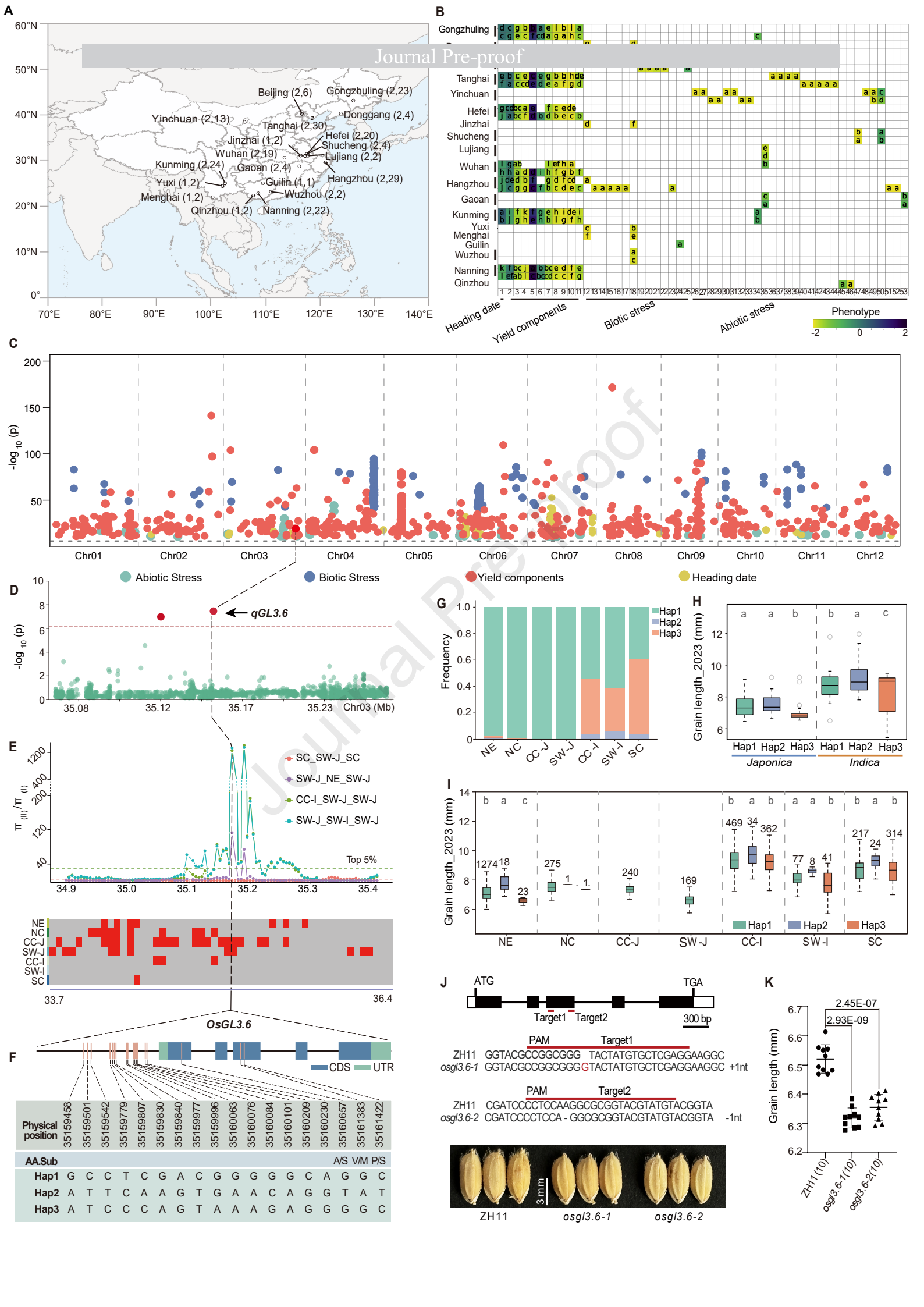
(C) ZKF8 was crossed with SG4, and the progenies were backcrossed to SG4. The $BC_1F_1$ generation was genotyped by resequencing, and individuals with heterozygous genomic segments covering the two target genes were manually chosen as backcrossing parents. Because of the high genetic similarity between ZKF8 and SG4, a $BC_1F_3$ individual with homologous donor alleles (red) only at the segment covering the two QTGs was selected as the improved SG4 line.

(D) Grain lengths of SG4, ZKF8, and the improved SG4 line grown in Sanya. Error bars represent standard deviations. *P*-values were calculated using a two-tailed Student's *t*-test. Scale bar, 5 mm.









A

**Germplasm**

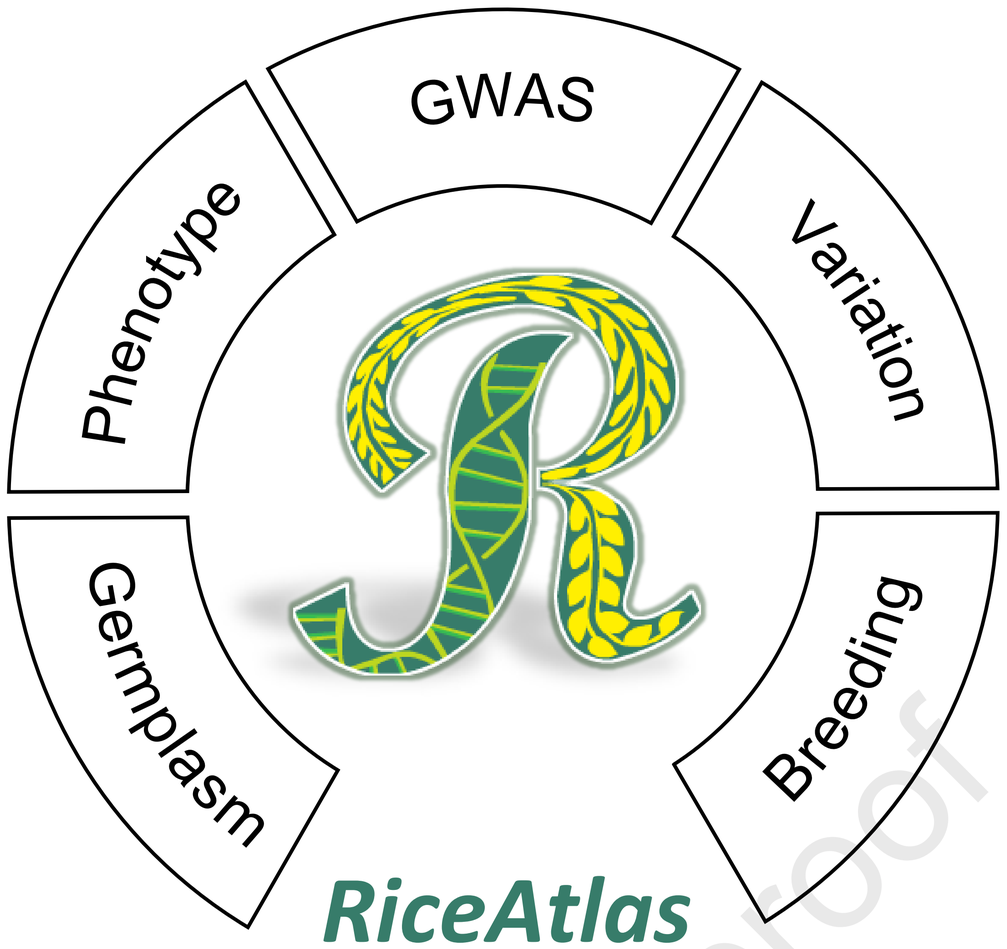
6044 accessions
(2706 *indica*, 3338 *japonica* )

**Resequencing**

21.1 million SNPs
4.0 million InDels

**Phenotype**

53 phenotypes,
19 geographic locations,
212 datasets




**GWAS**

3131 QTLs and
5187 mix-GWAS Peak-SNP

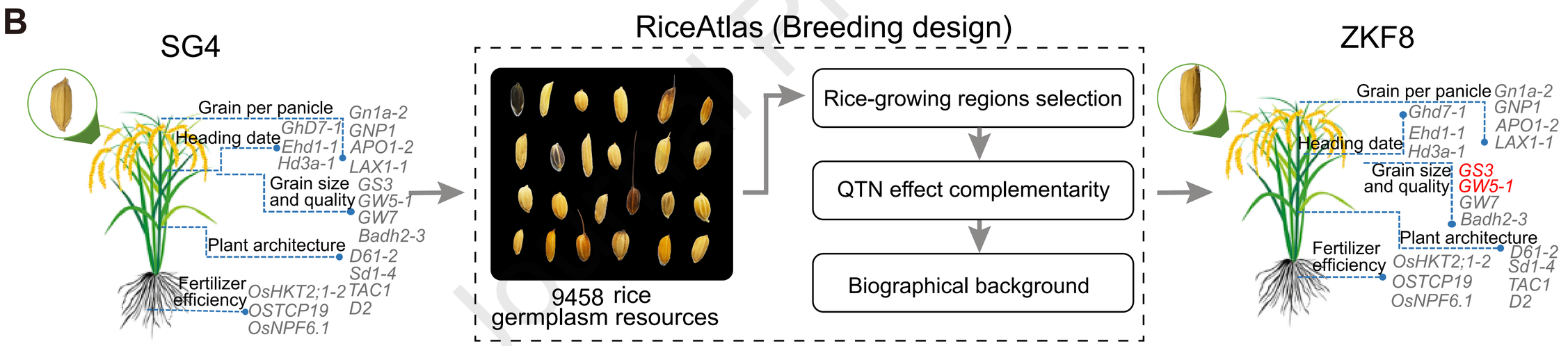
**Favorable alleles (QTNs)**

53 phenotypes totaling 24497
advantaged SNPs

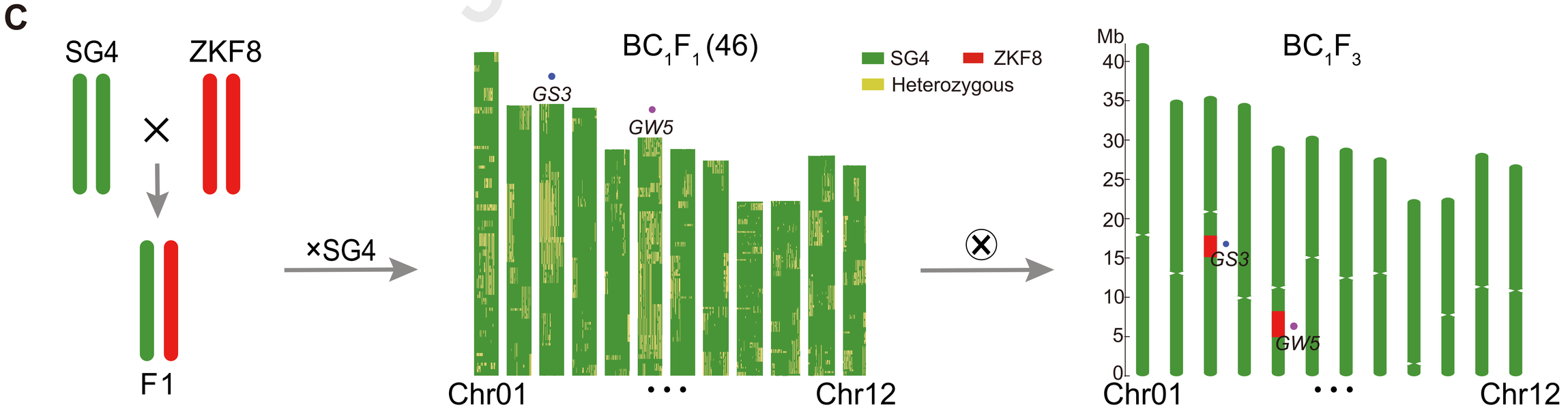
**Breeding design**

6044, 3K and 404 accessions
totaling 67.1 million SNPs

B




C




D