# RESEARCH ARTICLE

[1]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China; [2]Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin 150081, China; [3]Key Laboratory of Crop Biotechnology Breeding of the Ministry of Agriculture and Rural Affairs, Beidahuang Kenfeng Seed Co., Ltd, Harbin 150090, China; [4]Institute of Botany, Chinese Academy of Sciences, Beijing 100193, China; [5]Guangdong Laboratory for Lingnan Modern Agriculture and State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Agriculture, South China Agricultural University, Guangzhou 510642, China; [6]University of Chinese Academy of Sciences, Beijing 100049, China; [7]State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China and [8]Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China

## AGRICULTURAL SCIENCES

# Genomic decoding of breeding history to guide breeding-by-design in rice

Zhuo Chen [1],[†], Qingyun Bu [2],[†], Guifu Liu[1],[†], Maoqing Wang[3], Hongru Wang [1], Huazhao Liu[2,4], Xiufeng Li [2], Hong Li[5], Jun Fang[2], Yan Liang[1], Zhenfeng Teng[1,6], Sai Kang[1], Hong Yu[1], Zhukuan Cheng [1], Yongbiao Xue [7], Chengzhi Liang[1], Jiuyou Tang [1,*], Jiayang Li[1,8,*] and Chengcai Chu [1,5,6,8,*]

*Corresponding authors. E-mails: ccchu@genetics.ac.cn; jyli@genetics.ac.cn; jytang@genetics.ac.cn
[†]Equally contributed to this work.

## ABSTRACT

Deciphering the intrinsic molecular logic of empirical crop breeding from a genomic perspective is a decisive prerequisite for breeding-by-design (BbD), but remains not well established. Here, we decoded the historical features of past rice breeding by phenotyping and haplotyping 546 accessions covering the majority of cultivars bred in the history of Northeast China (NEC). We revealed that three groups founded the genetic diversities in NEC rice with distinct evolution patterns and traced and verified the breeding footprints to known or genome-wide association study (GWAS)-detected quantitative trait loci (QTLs), or introgressions from *indica* sub-species with chronological changes in allele frequencies. Then we summarized a rice breeding trend/principle in NEC, and combined with the successful example in breeding and application of Zhongkefa5 to demonstrate the guiding value of our conclusion for BbD in practice. Our study provides a paradigm for decoding the breeding history of a specific crop to guide BbD, which may have implications in different crop breeding.

**Keywords:** rice, breeding-by-design, breeding history, genome-wide association study, genomic decoding

## INTRODUCTION

With the continuous progress of functional genomic research, the concept of breeding-by-design (BbD) was proposed and put into practice to breed cultivars in a rapid, precise and systematic manner [1–4]. Generally, the design scheme will be established based on the current knowledge of the functional dissection of agronomic traits [5], which is essentially a process for purposefully combining a series of favorable allelic variations that are usually dispersed in different genetic resources, or designed allelic variations that not present in natural populations, for genes of agronomic importance. However, breeding based solely on the functional information of multiple genes themselves is impractical. There are at least two implicit prerequisites for the successful implementation of this process, namely the selection of backbone parents and the matching of characteristics of target cultivars with the environment and

planting conditions in the planting region and the preferences of consumers. The former is the starting platform of design breeding and in most cases includes major products of past breeding, and the latter determines the promotion and application of the final bred cultivar. The setting of both is actually inseparable from a comprehensive understanding of the intrinsic logic of past breeding practice. How to interpret the intrinsic rules from the functional genomics perspective and integrate them into BbD practice is one of the key but not well-elucidated fields.

Breeding objectives for crops always have historical and regional features that largely depend on new cultivation techniques, market demand and commercial value for the end products of crops, and the fluctuating environments including climatic changes and pathogen/insect evolvement [6]. For example, the popularization of cultivation techniques has

led to the selection of varieties adapted to direct seeding [7] and the improvement in the timeliness of harvesting and storage has spawned new requirements for low grain moisture content at harvest [8]. In addition, with the improvement in people's quality of life, food smell and taste are also changing and required [4,9]. All these have imperceptibly led breeders to flexibly tune their breeding directions. Therefore, the analyses from the perspective of changes in the times will be helpful to better illustrate the past breeding logic for a specific region.

Rice is grown in almost every province of China. Among them, rice breeding in Northeast China (NEC) covered by Heilongjiang (HLJ), Jilin (JL) and Liaoning (LN) provinces has attracted more attention due to the following facts. NEC is the northern limits (38.7–53.5°N, latitude) of rice cultivation in China and one of the northern limits of rice cultivation around the world. Owing to its unique natural and geographical conditions, NEC is an important region producing *japonica* rice with high grain quality in China. As the excellent eating quality of rice caters to the growing demands of consumers, the rice planting area and grain production in NEC have also increased substantially over the past few decades [10]. Notably, HLJ has been in a stable position as the largest rice production province in China (National Bureau of Statistics of China, NBSC, http://www.stats.gov.cn/enGliSH/). Since the planting area has reached the upper limit, strengthening and accelerating the breeding of new elite varieties is the best way to maintain the dominant position of the region in the supply of high-quality rice. Moreover, while rice cultivation in South and Southeast China can be traced to thousands of years ago, large-scale rice cultivation was introduced to NEC by Korean migrants only in the nineteenth century [11,12]. The short breeding history makes it possible to establish a basic decoding model of breeding logic that considers the chronological characteristics of breeding.

Based on these, we collected and resequenced a panel of 546 rice cultivars bred in NEC or imported from Korea and Japan (Supplementary Table S1). By integrating phenotypic, genomic and historical clues, we disclosed the landscape of genetic variations in NEC rice collection, providing informative insights regarding agronomical traits, pedigrees, introgressions and quantitative trait loci (QTLs). Finally, we summarized a trend/principle of rice breeding in NEC and demonstrated its potential guiding value in rice BbD in combining with the successful breeding of Zhongkefa5 (ZKF5).

# RESULTS

## Chronological features of rice cultivars in NEC

Modern rice breeding in NEC began with the introduction of few rice varieties from Japan and Korea in the 1940s, which subsequently became the backbone parents for local pedigree breeding and hybrid breeding [13,14]. To explore the patterns of genetic diversity within the rice cultivars in NEC, we collected and resequenced 546 cultivars (Supplementary Table S1) from the three provinces in NEC as well as cultivars from Korea and Japan (Fig. 1A), which represent the vast majority of introduced or bred cultivars in NEC from the 1940s to the 2010s. HLJ is the largest rice production province in China, accounting for ∼72% of the total rice production in NEC (NBSC, http://www.stats.gov.cn/enGliSH/). Accordingly, over half (309/546) of the cultivars in this collection were from HLJ. To more intuitively show the chronological features of the phenotypic and genomic changes of rice cultivars in HLJ, we further divided the HLJ cultivars into four historical groups (Fig. 1A), namely HLJ-I (∼1980), HLJ-II (1980–2000), HLJ-III (2000–10) and HLJ-IV (2010∼), by considering the representativeness determined by population size and the fact that the number of newly bred rice varieties exploded after 2000.

Phenotype data for 22 agronomic traits were measured in Experimental Station of Chinese Academy of Sciences (45.7°N, 126.6°E) in Harbin city of HLJ, during the natural growing season (Supplementary Table S2). Cultivars from different regions showed divergent phenotype, in correspondence with the growing conditions and preferences of the breeders. Cultivars from JL (40.9–46.3°N, latitude) and LN (38.7–43.4°N, latitude) had a later heading date and a higher grain number per panicle, which enable them to obtain higher yields in lower latitude regions with relatively abundant light and accumulated temperature compared with that in HLJ (43.4–53.5°N, latitude) (Fig. 1B). Indeed, cultivars from HLJ show generally earlier heading to harvest a reasonable yield before cold winter in higher latitudes (Fig. 1B). However, it is intriguing to observe a trend of delayed heading date in HLJ cultivars after 1980, which may benefit from warming climate [15] and the promotion of greenhouse seedling technology [16], making it possible to obtain higher yield by prolonging the growth period (Fig. 1C). Delayed flowering increased the risk of chilling damage during reproductive growth, also driving the improved chilling tolerance of cultivars (Fig. 1C). We also observed a trend for increasing grain length/width ratio (GLWR) and panicle blast
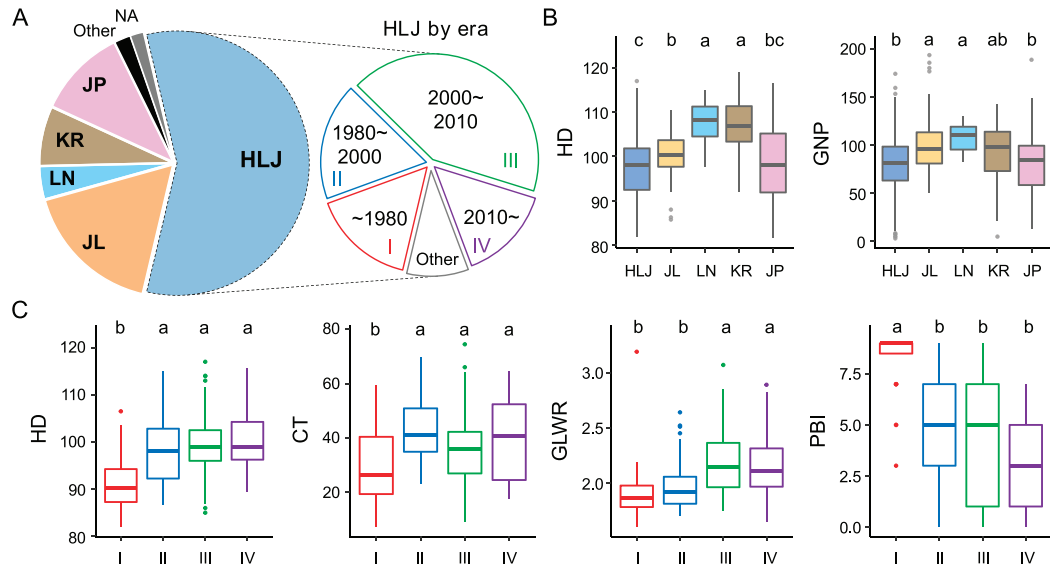
**Figure 1.** Geographical distribution and phenotypic divergence of rice in NEC. (A) Geographic and chronological distribution of cultivars in this study. HLJ, Heilongjiang province; JL, Jilin province; LN, Liaoning province; KR, Korea; JP, Japan. (B) Phenotypic variation of rice heading date (HD) and grain number per panicle (GNP) among different regions. (C) Phenotypic variation of rice in HLJ bred in different eras. CT, seed-setting rate after cold treatment; GLWR, grain length/width ratio; PBI, panicle blast index. Multiple comparison was performed by using Tukey's honestly significant difference (HSD) test.

resistance over time in cultivars from HLJ, which may have resulted from the growing preference of consumers for slender grain rice and increasing threat of rice blast disease (Fig. 1C).

## NEC rice cultivars can be traced to three subgroups

To dissect genomic changes of rice cultivars in NEC, all cultivars collected were sequenced at least five times (Supplementary Table S1). While it has been clear that most rice cultivars in NEC belong to the temperate-*japonica* subpopulation [17], we sought to gain a more comprehensive knowledge of the population structure and genetic make-up of this rice collection with genome-wide markers. First, we performed genome-wide subpopulation ancestry inference using single nucleotide polymorphism (SNP) markers [17] derived from the 3K-RG project [18]. As inferred by 3K-RG markers, all NEC rice cultivars were inferred as from temperate-*japonica* (Supplementary Table S3). Population structure inference was performed by using Admixture [19] software, identical-by-state (IBS) analysis and principal component analysis (PCA) (Supplementary Table S3). According to the inference of Admixture, the cross-validation error continued to decrease as the number of assigned ancestries (K) increased (Supplementary Fig. S1A); we therefore took the combination of Admixture and IBS analysis to determine a proper grouping. When assigning two ancestries (K = 2), the population was sepa-

rated into two subgroups: K2G1 (Group 1 when assigning two ancestries) and K2G2 (Fig. 2A). When K = 3, the former K2G2 group was further divided into two groups: K3G2 and K3G3. When K = 4, a new group K4G4 that consists of cultivars from both K3G2 and K3G3 emerged (Supplementary Fig. S1B), but the mean dissimilarity among cultivars within K4G4 exceeded that between K4G4 and K4G3, indicating an unproper grouping (Supplementary Fig. S1C). Therefore, we found that it is reasonable to separate the population into three subgroups (Fig. 2B and Supplementary Table S3). We reperformed the Admixture analysis 100 times with subsets of SNP markers and found that all the replicates supported the original grouping with the total data set when K = 3. The grouping became highly unstable with K > 3 (Supplementary Fig. S1D). When K = 3, 56 samples were assigned to different subgroups in >25 replicates and were considered to be unreliably grouped. These 56 samples were not included in analysis based on the three subgroups. A neighbor-joining tree constructed using the same SNP data set also supported the divergence of the inferred subgroups (Supplementary Fig. S1E). PCA showed that the population structure was weak, with the top 10 principal components (PCs) explaining only 23.49% of the total variation (Supplementary Table S3). Among the top three PCs, PC1 and PC3 supported the clustering of the three subgroups inferred by using Admixture (Fig. 2C), while PC2 was highly correlated with the proportion of *indica* components in the cultivars ($r^2 = 0.545$,
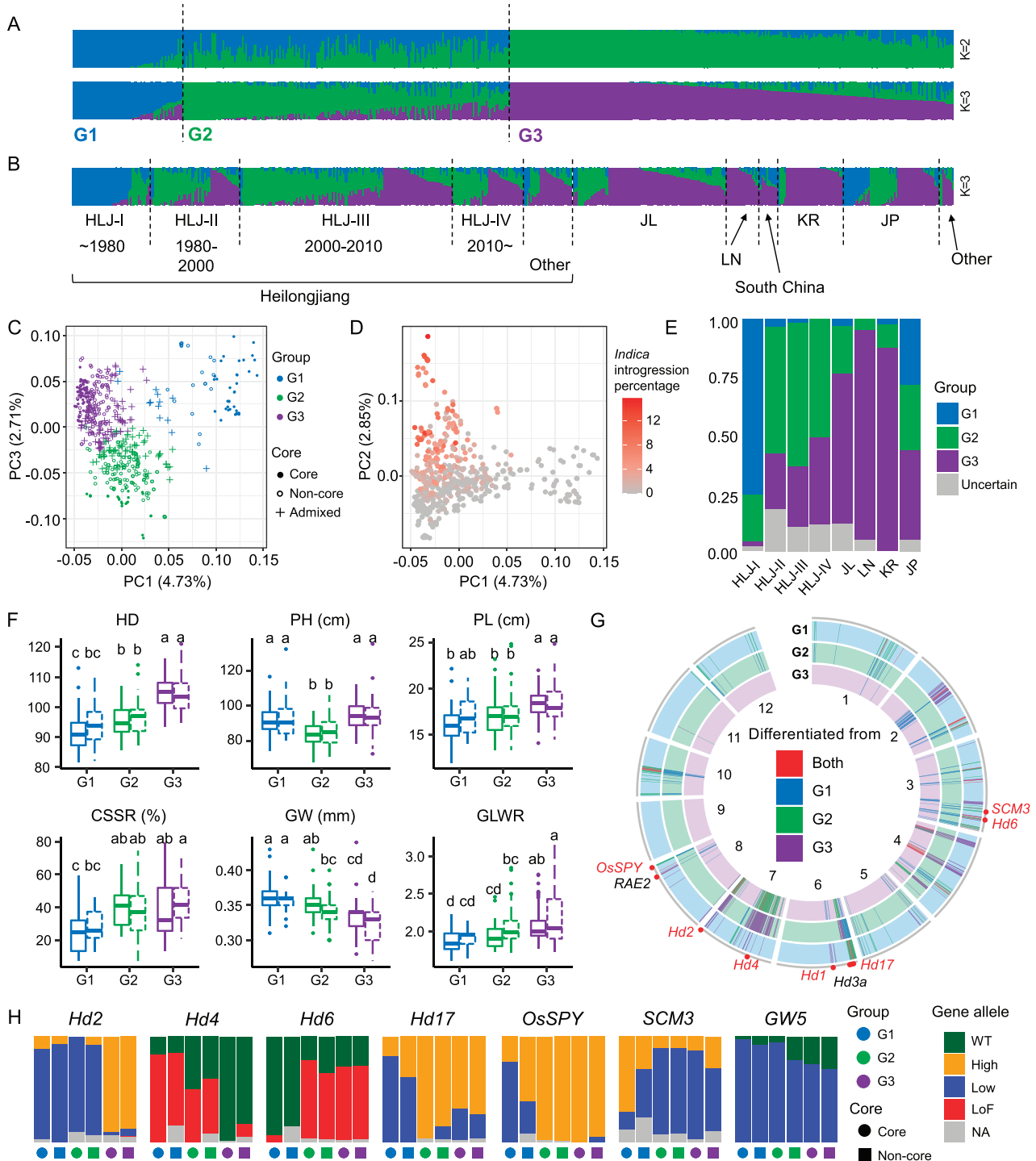
**Figure 2.** Three rice subgroups founded the genetic diversities in NEC rice with distinct evolution patterns. (A) and (B) Subgroups inferred by using Admixture software assuming two (K = 2) and three ancestries (K = 3) of 546 cultivars using genome-wide SNP markers. Samples are sorted by ancestry scores (A) and by geographical origin when K = 3 (B). HLJ, Heilongjiang province; JL, Jilin province; LN, Liaoning province; KR, Korea; JP, Japan. (C) and (D) Principal component analysis (PCA) of 546 cultivars using SNP markers; samples are colored by the subgroups inferred by using Admixture assuming three ancestries (G1, G2 and G3) (C) and by the percentage of introgressions from *indica* sub-species (D). (E) Proportion of cultivars from the three subgroups (G1, G2 and G3) in different regions and eras. (F) Phenotypic variation of four traits among the three subgroups (G1, G2 and G3) inferred by using Admixture. Solid line for core cultivars and dashed line for non-core cultivars. HD, heading date; PH, plant height; PL, panicle length; GW, grain width; GLWR, grain length–width ratio; CSSR, seed-setting rate after cold treatment. Multiple comparison performed by using Tukey's HSD test. (G) Genome distribution of highly differentiated ($F_{ST}$ top 5%) regions between each of the two subgroups. (H) Allele frequency of important differentiated QTLs among subgroups. In each group, core cultivars (ancestry score > 0.99) and non-core cultivars (ancestry score ≤ 0.99) are tabulated separately. Gene alleles are defined as follows: WT, functional wild-type; High, higher protein function or expression level; Low, lower protein function or expression level; LoF, loss-of-protein function.

$P$-value $= 8.57 \times 10^{-95}$), indicating a substantial impact of *indica* introgression on the genetic structure of NEC *japonica* cultivars (Fig. 2D).

Known backbone cultivars were found within the core cultivars (ancestry score > 0.99) of the three subgroups (Supplementary Table S3). For example, the core cultivars of K3G1 included a well-known Japanese cultivar, Ishikari-Shiroge, which was developed in the high-latitude region Hokkaido. Over 70% of the cultivars from HLJ before the 1980s were clustered within K3G1 (Fig. 2E), consistently with the record that Ishikari-Shiroge was introduced to HLJ in the 1940s and since then has served as the backbone breeding material for decades [13]. Core cultivars of K3G2 included backbone cultivars from Japan, such as Yukara and KuiKu131 (Kongyu131 in China, KY131). K3G3 included backbone cultivars such as Fujihikari, Akihikari and Fukei138. These results collectively indicate the fundamental contribution of *japonica* backbone pedigrees to the genetic structure of NEC rice.

The distribution of the three subgroups was uneven in different regions and eras (Fig. 2E). In HLJ, K3G1 was dominant in cultivars before the 1980s, but was replaced by K3G2 thereafter. K3G3 was dominant in JL and LN, and with increasing proportions in HLJ over time. The phenotypic differentiation of the three major subgroups corresponded to their geographical and chronological distribution (Fig. 2F). Group K3G1 was characterized by extremely early flowering that may be the key to their successful adaptation to high-latitude HLJ in the early years. In contrast, with the improvement in the cultivation experience and farming techniques, cultivars in K3G2 showed relatively later flowering than K3G1, but lower plant height and stronger cold tolerance (Fig. 2F). K3G3 showed even later flowering than K3G2 and with higher plant and longer panicle (Fig. 2F). Also, the grain width decreased from K3G1 to K3G3 and the GLWR increased accordingly (Fig. 2F).

## Breeding traces of NEC rice cultivars are linked to major QTLs

The phenotypic differentiation between the subgroups should be caused by genetic variation. We performed genome-wide scan for genetic differentiation ($F_{ST}$) in a pairwise manner among the three subgroups (Fig. 2G and Supplementary Fig. S2). The top differentiated regions (top 5% in each comparison) harbored several known QTLs for heading, awn and plant architecture (Fig. 2G and Supplementary Table S4). Among them, *Hd2* and *Hd4* were reported to be the major determinants of the heading date variation in NEC [20]. The weak al-

lele (allele with low effect) of *Hd2* confers earlier flowering [21,22] and was nearly fixed in K3G1 and K3G2, while the strong allele was dominant in K3G3 (Fig. 2H). A similar pattern was also found on *Hd4* as its loss-of-function (LoF) allele [23] presents mainly in K3G1 (Fig. 2H). The weak allele of *OsSPY* confers higher plant height [24] and was present mainly in K3G1 but was absent in K3G2, which may explain the difference in plant height between these two populations (Fig. 2H). The strong allele of *SCM3/OsTB1* on chromosome 3 was reported to confer strong culm and bigger panicle but lower tiller number [25], and was present mainly in K3G1 (Fig. 2H). Interestingly, while they are located on different chromosomes, the weak culm and tiller promoting allele of *SCM3* was linked to the short culm allele of *OsSPY* in K3G1 (Chi-square test Phi = 0.56; $P$-value $= 9.44 \times 10^{-5}$). This linkage indicates a potential co-selection for lodging resistance and tillering in *japonica* rice, just like the relationship between *SD1* and *HTD1* [26] in *indica* rice for the Green Revolution. Some other QTLs that did not fall within the top differentiated regions also showed diverged allele distribution among the subgroups. The weak allele of *GW5* [27] in *japonica* rice cultivars contributed to wider grain and the higher ratio of wild-type allele in G2 and G3 (Fig. 2H) may account for their slender grain compared with G1 (Fig. 2F).

The effect of a QTL may rely on a certain genetic background and/or experimental conditions; we thus tested the phenotypic effects of these QTLs separately in each of the three subgroups. We genotyped 111 QTLs with functionally verified natural variants or linkage markers (Supplementary Table S5). This analysis yielded a list of QTLs with significant effect on their related traits in NEC (Supplementary Table S6). For example, *Hd2* and *Hd4*, the major contributors of the heading date variation, also showed impacts on the panicle traits. Grain shape was controlled by *GW5* [27], *GS3* [28] and *GW6* [29]. *Pi9* [30] and *Pib* [31] were effective to blast resistance. Multiple QTLs including *SCM2* [32], *SCM3* [25], *DEP1* [33], *LAX1* [34] and *OsSPY* [24] collectively influenced panicle architecture. Some known QTLs were further validated and novel QTLs were detected by using genome-wide association study (GWAS) (Supplementary Fig. S3 and Supplementary Table S7). The known QTLs validated by using GWAS include *Hd2* for heading, *GW5* for grain shape and the Reimei-type semidwarf allele of *SD1* [35] (Fig. 3A and B, and Supplementary Fig. S4). A novel QTL for cold tolerance (seed-setting rate after cold treatment at booting stage) was found on chromosome 5, named as *qC-SSR5* (Fig. 3C and E). The cold-tolerant allele of this QTL was dominant (>90%) in K3G2 and was also
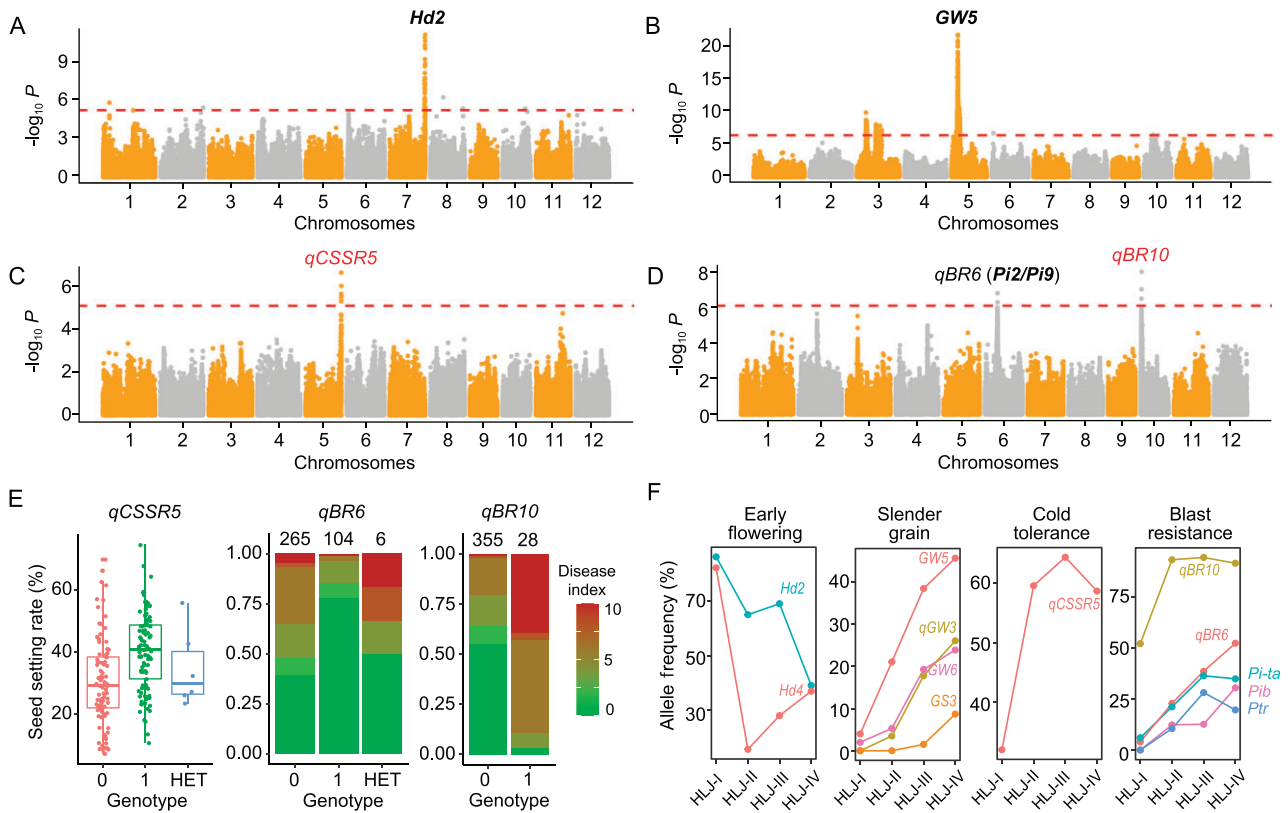
**Figure 3.** Known and novel QTLs detected by using genome-wide association studies with chronological changes in allele frequencies. (A)–(D) Manhattan plots of genome-wide association studies on heading date (A), grain width/length ratio (B), seed-setting rate after cold treatment (C) and leaf blast index in 2010 (D). Known QTLs are marked in bold and novel QTLs except *qBR6*, which may be allelic to *Pi2/Pi9*, are marked in red. (E) Phenotype distribution of different alleles on novel QTLs for cold tolerance and blast resistance. On each QTL, genotype 0 for reference genome allele (Nipponbare) and genotype 1 for alternative allele. HET, heterozygous genotypes. (F) Allele frequency of known and novel QTLs that showed chronological trends of improvement in Heilongjiang province.

present with a relatively high ratio in the other two subgroups. Two QTLs for rice blast resistance were detected including a novel QTL *qBR10* on chromosome 10 and *qBR6* on chromosome 6, which is allelic to the well-known *Pi2/Pi9/Piz-t/Pigm* [30,36,37] locus (Fig. 3D and E). The susceptible allele of *qBR10* was present mainly in early cultivars and is nearly eliminated in modern cultivars.

By combining the known QTLs with significant phenotypic effect or geographical differentiation signatures, as well as those novel QTLs detected by using GWAS, we assembled a list of applicable QTLs for breeding in NEC (Fig. 5B). The weak allele of *Hd2* leads to earlier heading and presents with high frequency in HLJ but with low frequencies in lower latitudes such as JL, LN and KR. The dense and erect allele of *DEP1* presents with high frequency in LN, which reflects local breeding preference. The Reimei semi-dwarf allele of *SD1* is dominant in Korean cultivars (Supplementary Table S5D). Corresponding to the chronological evolution of agronomic traits in HLJ (Fig. 1C), we detected changes in allele frequencies of QTLs associated with them (Fig. 3F), in-

cluding a remarkable reduction in the early heading allele of *Hd2*, increasing frequencies of multiple alleles (*GW5*, *GS3*, *GW6*, *qGW3*) conferring slender grain, cold tolerance (*qCSSR5*) and blast disease resistance (*qBR6* and *qBR10*). These QTLs partially explained the genetic basis for increased yield, improved quality and strengthened adaptation to biotic and abiotic stresses in HLJ rice breeding over decades. Such findings will be useful for the choice of basic breeding materials and also marker-assisted selection during BbD in the future.

## Introgressions from *indica* sub-species mainly contribute to yield and resistance traits

In NEC, almost all rice cultivars belong to temperate-*japonica* with limited genetic background. To further improve the yield and resistance traits, rice breeders broadened the genetic background by introducing *indica* genetic components [38]. These efforts changed the genetic components of NEC rice as clearly reflected in the
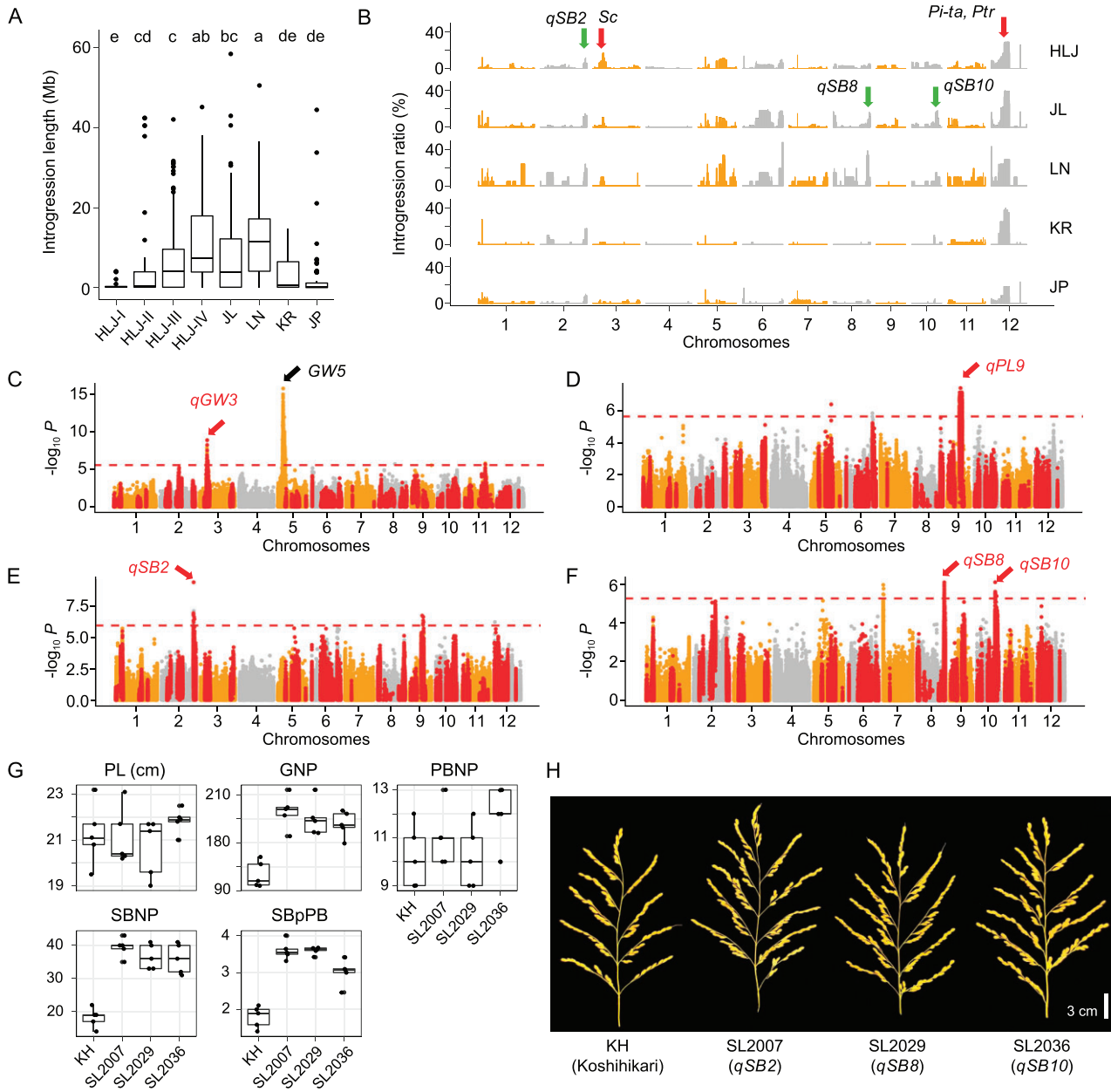
**Figure 4.** Distribution of *indica*-derived introgressions and their impact on agronomical traits. (A) Percentage of introgressions from *indica* sub-species in cultivars from different regions and eras. Multiple comparison performed by using Tukey's HSD test. (B) Frequency and positional distribution of *indica* introgressions in cultivars from different regions. (C)–(F) Manhattan plots of genome-wide association studies on grain width (C), panicle length (D), secondary branching number per panicle (E) and secondary branching number per primary branching (F). SNPs correlated with *indica* introgressions are highlighted in red. (G) and (H) Panicle phenotypes of *japonica* cultivar Koshihikari (KH) and introgression lines carrying QTLs from *indica* cultivar IR64. PL, panicle length; GNP, grain number per panicle; PBNP, primary branching number per panicle; SBNP, secondary branching number per panicle; SBpPB, secondary branching number per primary branching.

population structure analysis (Fig. 2D). To visually demonstrate the genetic impact of this process, we constructed a detailed map of *indica* introgressions and found that *indica* introgressions were present on all 12 chromosomes of NEC cultivars (Fig. 4B). The percentage of *indica* introgression was low in

the cultivars from Japan but significantly higher in NEC cultivars (Fig. 4A). In NEC, cultivars from LN showed the highest level of *indica* introgressions, with an average of 80.1 Mb per cultivar. In HLJ province, the average content of *indica* introgressions in a cultivar increased significantly over time,

from 8.2 Mb before 1980 to 76.6 Mb after 2010, which reflects its preference in breeding (Fig. 4A and Supplementary Fig. S5).

In the breeding process, *indica* introgressions with beneficial effect are more likely to be inherited. Accordingly, we focused on the introgression segments with high frequency in NEC (Fig. 4B and Supplementary Table S8A). A large *indica* introgression region with the highest frequency is located on chromosome 12, which introduced the resistant allele of one or both of the two adjacent rice blast resistance loci, *Pi-ta* [39] and *Ptr* [40], which significantly improved blast resistance in *japonica* rice [41]. Other highly frequent introgressions on chromosomes 2, 3, 5 and 6 may also introduce beneficial traits from *indica* (Fig. 4B). In HLJ, some introgressions showed increasing frequency since the 1980s. These included the one harboring *Pi-ta* and *Ptr* on chromosome 12, which was increased from 6% before 1980 to 37% after 2010, and the one harboring hybrid male sterility locus *Sc-I* [42] on chromosome 3, which was increased from 0% before 1980 to 39% after 2010 (Supplementary Fig. S5).

In K3G2 and K3G3, the proportion of *indica* introgression in cultivars showed significantly positive correlation (Supplementary Table S8B) with panicle traits, including panicle length, primary and/or secondary branching number per panicle and grain number per panicle. Using GWAS, we detected QTLs for several yield traits that were associated with *indica* introgressions. These include several QTLs on which *indica*-derived alleles positively contribute to grain width (Fig. 4C), panicle length (Fig. 4D) and branching number (Fig. 4E and F), consistently with the correlation analysis and previous studies [43]. The effects of three introgression-related QTLs on Chr2, Chr8 and Chr10 for panicle traits, named as *qSB2*, *qSB8* and *qSB10*, respectively, were further verified by using chromosome segment substitution lines (CSSLs) of Koshihikari (temperate-*japonica*) background carrying IR64 (*indica*) introgressions. Introgression of *indica* segments carrying *qSB2*, *qSB8* and *qSB10*, respectively, increased the secondary branching number per panicle by 100% and grain number per panicle by 50% (Fig. 4G and H, Supplementary Fig. S6 and Supplementary Table S9). These results proved the substantial contribution of *indica* introgressions to the improvement of yield traits in NEC rice breeding (Fig. 5A).

## Decoding past breeding practice has guiding value in BbD

Based on our above analyses, to achieve a coordinated improvement in grain yield, quality and

stress tolerance in NEC rice breeding, the following principle or trend will be expected. (i) Given that the group K3G2 that dominates HLJ is being replaced by K3G3 over time (Fig. 2E), it may be better to choose elite varieties of K3G3 as the starting platform of design breeding. (ii) Increase yield or rice blast resistance by pyramiding more *indica*-derived segments or QTLs including *SCM2*, *qSB2*, *qSB8*, *qSB10*, *Ptr*, *Pi-ta* and unknown loci to be identified. (iii) Improve grain quality to cater to the demand of consumers by introducing multiple grain shape QTLs, such as the weak/null allele of *GW5*, *GS3* or other alleles that increase grain length. (iv) Use other gene alleles in our list (Fig. 5B) rationally and flexibly to breed cultivars that can better adapt to the environment and cultivation conditions in NEC, such as using the *qCSSR5* locus to improve cold tolerance, using the *SCM3* and *OsSPY* linkage to tackle lodging at the late maturity stage caused by typhoons that more frequently occurred in NEC in recent years [44], etc.

In the practice of rice production, since 1990, KY131 from K3G2 has been in an absolutely dominant position in the NEC rice planting area [45]. From 2013 to the present, KY131 was replaced by Longgeng31 (LG31) with enhanced blast resistance from the same subgroup [46]. The success of KY131 and LG31 largely benefits from the recognition of rice growers for their stable yield, early maturity, cold tolerance, lodging resistance and acceptable quality [45,46]. These features may be attributed to the pyramid of a number of excellent natural alleles as we listed based on genome analyses (Fig. 6C). From the perspective of consumers, high-quality rice has much more commercial value and the most typical example is Daohuaxiang2 (DHX2, Fig. 6A), which has all indicators at the highest level of rice quality standards fixed by the Ministry of Agriculture and Rural Affairs of China [47] and is always in shortage. Therefore, breeding rice cultivars with high stable yield and high quality is still the goal and top priority.

Recently, Jiayang Li's team and their collaborators bred a new elite variety with both high yield and good quality, namely Zhongkefa5 (ZKF5, Fig. 6A). After decoding the breeding history of ZKF5 from a genomic perspective (Fig. 6B and C, and Supplementary Table S10), we demonstrated that it essentially provided practical validation for the breeding trend sorted out based on our analysis. ZKF5 was selected from a multi-parental cross population using Jigeng88 (JG88), an elite cultivar from K3G3, as the backbone cultivar to hybridize successively with KY131 from K3G2, and Nanfangchangligeng (CLG), a *japonica* cultivar from South China rice planting areas carrying a
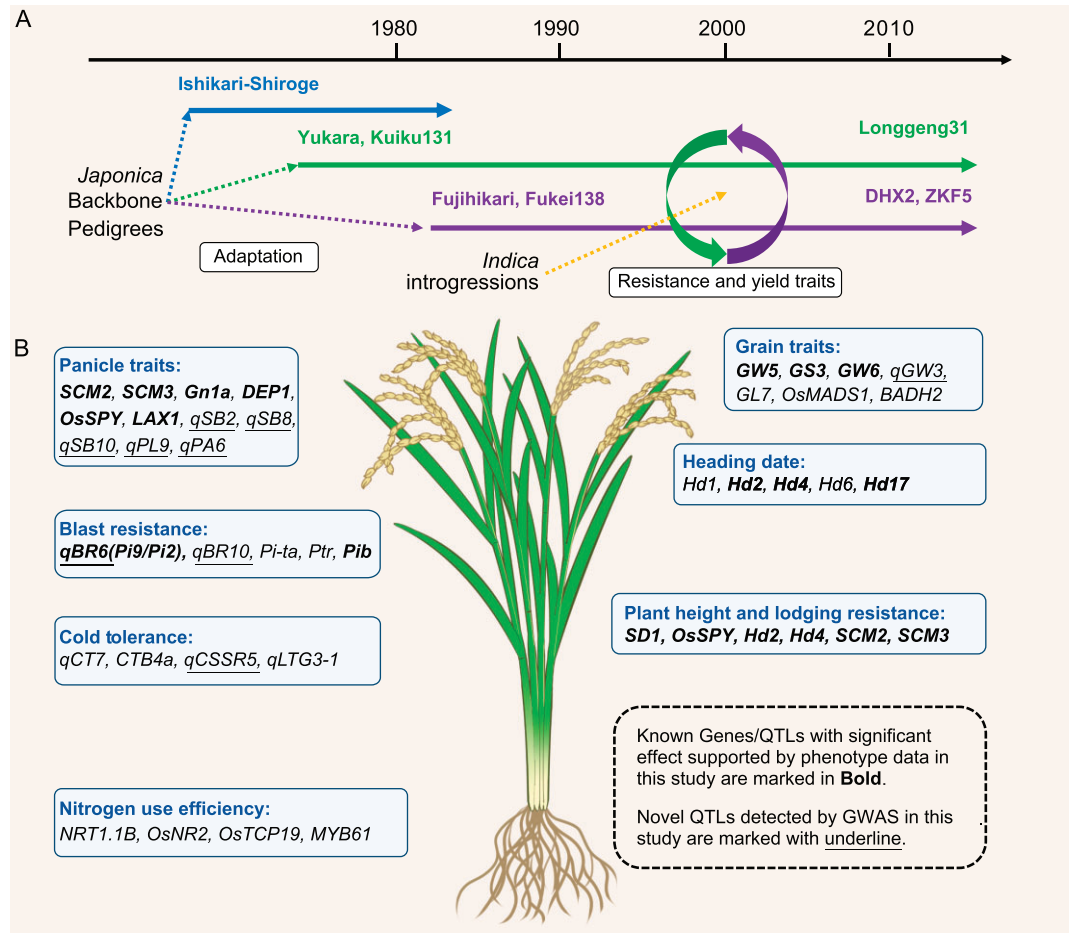
**Figure 5.** Inferred contribution of pedigrees, *indica* introgressions and applicable QTLs in NEC rice breeding. (A) An inferred model of rice breeding history in Heilongjiang province of NEC. Three major *japonica* subgroups with representative backbone cultivars and introgressions from *indica* rice that collectively contributed to genetic and phenotypic diversity. (B) Applicable QTLs collected from multiple methods, regulating multiple agronomically important traits. Known genes/QTLs with significant effect supported by phenotype data in this study are marked in bold. Novel QTLs detected by using GWAS in this study are marked with underline.

multi-copy *GL7* allele that increases grain length but has not yet been detected in NEC varieties, *Pib* allele that confers rice blast resistance and several *indica* introgression regions that are expected to improve yield (Supplementary Table S5 and Supplementary Fig. S7). In terms of appearance quality, genomic characterization confirmed that ZKF5 has a genotype combination of *GL7*-multi-copy, *gs3* and *GW5*, resulting in long/slender grains comparable to DHX2 (Fig. 6D and E) that has a genotype combination of *gs3*, *GW5*, *osmads1* [48] and *GW6* (Fig. 6C and Supplementary Fig. S8). For yield improvement, ZKF5 harbors *qSB2*, *qSB8*, *qSB10* and *SCM2* (Fig. 6B and C, and Supplementary Fig. S8) and reached an average yield of 700.48 and 678.50 kilogram per mu (1 mu = 1/15 hectare) in 2 years of rice regional trials, which is about 8.36% and 10.22% higher than those of Jiyugeng (JYG), the officially designated control cultivar for regional trials, re-

spectively (Fig. 6F and Supplementary Table S11). Moreover, ZKF5 carries the blast resistance alleles of *Pib*, *qBR10*, *Pi-ta* and *Ptr* (Fig. 6B and C, and Supplementary Fig. S8), making it much more resistant to rice blast than JYG (Fig. 6G and Supplementary Fig. S9). Notably, compared with DHX2, ZKF5 showed excellent lodging resistance under natural field conditions (Fig. 6H), supporting the successful selection of appropriate plant height and culm strength.

## DISCUSSION

In this study, to accurately assess the trend of rice breeding in the new period, we took systematic analyses of the chronological characteristics of the past breeding practices in NEC from a genomic perspective. We genotyped 111 known QTLs and novel GWAS-QTLs in our rice collection (Supplementary Tables S5 and S7). Among them, some QTLs
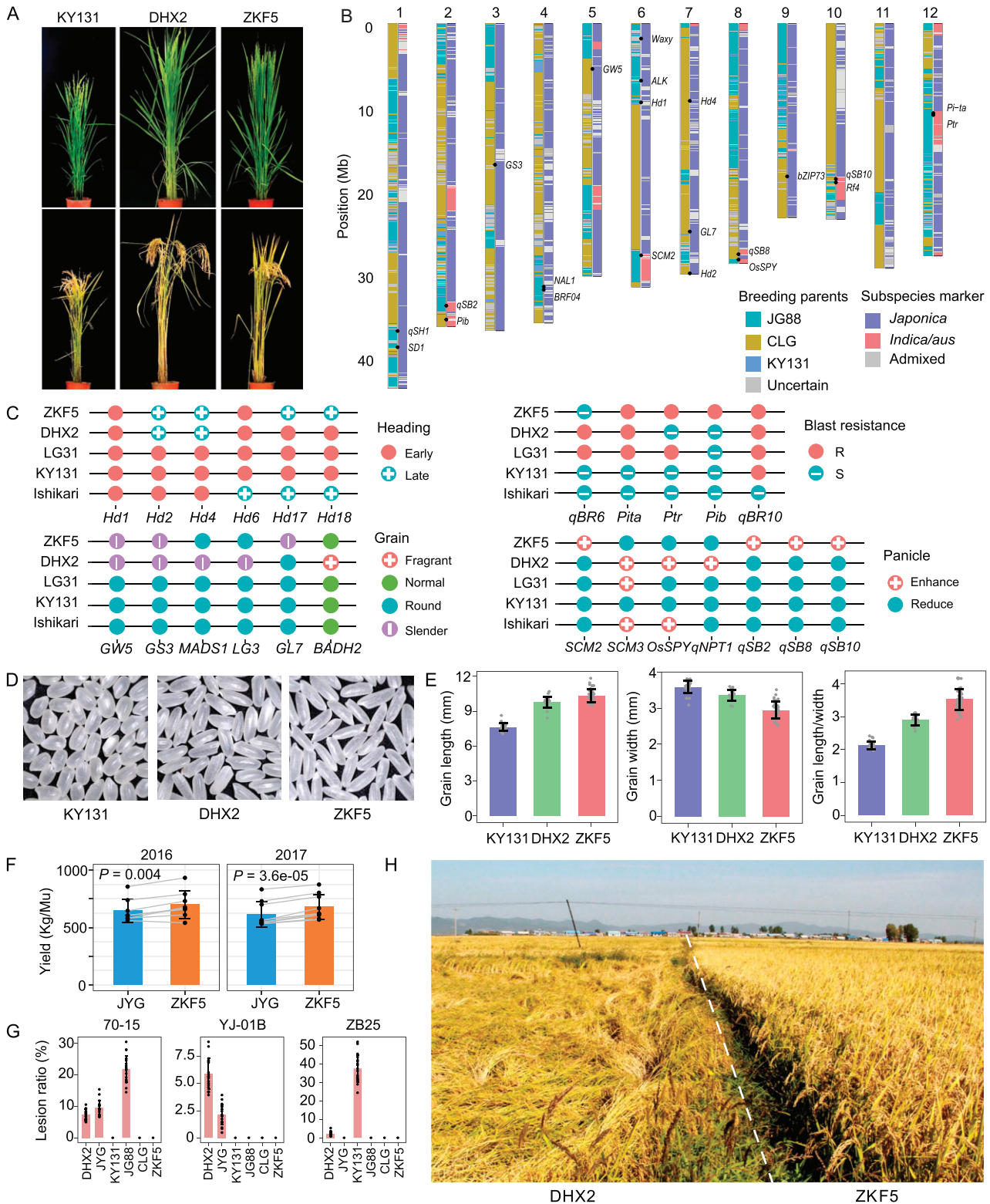
**Figure 6.** Genomic decoding of breeding history of an elite *japonica* cultivar, Zhongkefa5. (A) Plant phenotype of three elite cultivars: KY131, Daohuax-iang2 (DHX2) and Zhongkefa5 (ZKF5) at heading and mature stage. (B) Genome contribution of three breeding parents of ZKF5 and distribution of *indica* introgressions; favorable QTLs and breeding targets are marked on the chromosomes. (C) Allele combination of important QTLs in elite cultivars and early cultivar Ishikari-Shiroge (Ishikari) for comparison. For blast resistance QTLs, R for resistant allele and S for susceptible allele. (D) and (E) Grain phenotype of KY131, DHX2 and ZKF5. (F) Yield of ZKF5 compared to the control cultivar Jiyugeng (JYG) in nine experimental sites of Northern China in 2016 and 2017 (Supplementary Table S11); *P*-values are for Student's *t*-test. (G) Leaf lesion area ratio of ZKF5 and five other cultivars inoculated with different *Magnaporthe oryzae* isolates. (H) Field phenotype showing the lodging resistance performance of ZKF5 compared to DHX2.

contribute to the basic adaptation traits in NEC, such as *Hd1* for proper heading date [20], while others were alterable for designing specific cultivar for a specific need [17]. Some QTLs that have been widely used in NEC and with promising effect deserve further concern, such as *Hd2*, *Hd4*, *SCM3*, *SCM2*, *DEP1* and *qBR6*. Some QTLs that may introduce unfavorable traits should be avoided in breeding, such as the blast susceptible allele of *qBR10*. Also, some QTLs for disease resistance that may have made a historical contribution to resistance breeding in NEC, such as *Pi-ta* and *Ptr* [41], now face the threat of lost resistance due to long-term application, which strongly reminds rice breeders that introducing new resistance loci is needed. We noticed that many reported QTLs were not detected by using GWAS, which may be due to the influence of population structure, complex interaction between loci and environment effects or limitations of the algorithm, and new methods for GWAS may be able to get more results. Estimating the additive and dominance effects of QTL, as well as the controlling polygenic backgrounds, should be a promising approach to improve statistical power and avoid confounding effects. Recently, a comprehensive GWAS method for dissecting quantitative traits, 3VmrMLM, was developed and released in software package IIIVmrRMLM [49]. Adopting this method for our data confirmed some of the QTLs we found by using EM-MAX and also many other novel QTLs (Supplementary Table S12) that may be worth further verification. This suggests that reanalysing old data with new methods could yield more valuable information.

Although NEC cultivars are dominated by *japonica* rice background, introgressions from *indica* sub-species have also contributed to the breeding in NEC, as documented by previous studies [17,38]. Based on GWAS analysis, we confirmed that some chromosome segments that were introduced into *japonica* cultivars from inter-sub-species crossings harbor major QTLs contributing to resistance and yield traits, further highlighting the importance of broadening genetic diversity during NEC rice breeding. Although more introgression regions of *indica* sub-species are beyond the range of QTLs detected by our GWAS, it is more likely that these regions may be related to some important traits that we did not collect. Consistently with this speculation, by analysing the chronological distribution trend of these fragments, we found that some introgression chromosome regions appeared with increasing frequency in NEC rice cultivars over time (Supplementary Fig. S5), suggesting that these regions may be preserved due to an unknown role in NEC rice improvement. These regions deserve special attention in functional analysis and rice breeding in the future.

ZKF5 now serves as the main cultivar in HLJ and JL provinces with its high yield, good quality and stress-tolerant features. Our analysis has decoded the genomic history of its successful breeding and, moreover, based on our analysis, ZKF5 may be further improved by utilizing beneficial genes such as *Hd2*, *Hd4*, *Hd17* and *Hd18* for earlier heading and *SCM3*, *OsSPY*, *qNPT1* and nitrogen use efficiency-related alleles for higher yield, pinpointing the further breeding path. As a longer-term breeding goal in NEC, it is necessary to analyse the excellent gene alleles harbored in important loci such as *qCSSR5*, *qSB2*, *qSB8* and *qSB10* and in *indica* rice introgression fragments selected with high frequency, so as to achieve more precise and more comprehensive trait improvements, which are expected to further validate the guiding value of our analyses on BbD in practice. In addition, the construction of CSSLs to clarify the potential utilization value of other introgression fragments from different rice regions or sub-species, which have not yet been integrated into NEC cultivars, will also help to provide more diverse genetic resources for the future improvement of rice varieties in NEC.

It is worth noting that low genetic diversity is one of the biggest bottlenecks in rice breeding in NEC, which is bound to be aggravated in the long run if it is limited to select backbone parents in a specific group for improvement. In addition, BbD of rice varieties including ZKF5 at this stage is limited to the design and pyramiding of a handful of gene loci with large effects. The design and pyramiding of large-scale important agronomic trait loci, especially those with minor effects, will depend on the application of breeding strategies including genomic selection [4,50,51]. That is, BbD of backbone parents in NEC using genomic selection by referencing the core trait loci with chronological features as we provided (Figs 3–5) in the future may be a rational way to contribute to the innovation and diversification of breeding resources in this region.

Overall, our study analysed the chronological characteristics of rice breeding practice in NEC from the perspective of functional genomics and dissected basic information that was absent in the past but is very important for future BbD. Similar strategies can be applied to other rice or crop breeding regions, which may accelerate BbD of different crops worldwide.

## MATERIAL AND METHODS
### Plant materials and phenotype evaluation

The rice population was collected from multiple rice institutes in Heilongjiang, Jilin, Liaoning, Korea and Japan. Most of the agronomic trait phenotyping

was performed in Harbin (45°N, China) in 2014. Cold tolerance measurements were performed in the lab. Field blast measurements were performed in Shangzhi (44°N, China) in 2010 and 2011 according to the grading criteria formulated by the International Rice Research Institute [52]. CSSLs derived from a cross between Koshihikari and IR64 were kindly provided by Nagata *et al.* (https://www.rgrc.dna.affrc.go.jp/index.html.en) [53]. Details are provided in the Supplementary data.

## DNA sequencing and data processing

DNA libraries (400–500 bp) were prepared and sequenced by using a Hiseq2000 genome analyser (Illumina, San Diego, CA). Sequencing data were aligned to the Nipponbare reference genome (version 7.0) [54] using Burrows-Wheeler Alignment (BWA) Tool [55]. Data analysis were performed using samtools [56] and GATK [57] (version 2.7.2). Details are provided in the Supplementary data and Supplementary Fig. S10.

## Population structure inference

We performed population structure inference using Admixture [58]. Principle component analysis was performed using smartpca [59]. A neighbor-joining tree was constructed by using dna.dist and nj function in R package APE [60]. Details are provided in the Supplementary data.

## Genome-wide subpopulation ancestry inference and inter-subpopulation introgression inference

The procedure of subpopulation ancestry inference was described previously, based on 3K-SNP and 3K-HAP data sets [17]. We further used $f_{dM}$, a statistic related to Patterson's *D* implemented in Dsuite software [61], to test whether these putative introgression events overlap with Admixture events between subpopulations. Details are provided in the Supplementary data.

## Measuring nucleotide diversity and differentiation

Fixation index ($F_{ST}$) and nucleotide diversity ($\pi$) were calculated using VCFtools (version 0.1.15) [62]. Details are provided in the Supplementary data.

## Genotyping of known QTLs

Known QTLs (Supplementary Table S5) were collected and genotyped as described [63,64]. Gene copy number variation was detected as described [65]. Allele effects were defined according to their publications as follows: WT for functional wild-type; High for higher protein function or expression level; Low for lower protein function or expression level; LoF for loss-of-function due to broken protein or loss of gene sequence. Details are provided in the Supplementary data.

## Genome-wide association study

GWAS were performed using EMMAX with the mixed linear model (MLM) [66]. We also performed GWAS with a new method, IIIVmrMLM [49], with default parameters. Details are provided in the Supplementary data.

## Breeding procedure of ZKF5

CLG as the male parent was first crossed with KY131, and the resulting $F_1$ plants as the female parent was further crossed with JG88 to generate a segregating population, which was bred from $F_2$ generation to $F_9$ generation, so as to obtain stable lines. In the meantime, each generation was screened through agronomic preference evaluation in the field and molecular marker-assisted selection (Supplementary Table S10 and Supplementary Fig. S8) in the lab to confirm the genotype combination of yield, disease resistance and grain shape.

## DATA AVAILABILITY

Raw FASTQ reads for all rice accessions whose genomes were sequenced for this study have been deposited in the SRA under Bioproject accession number PRJNA844290.

## SUPPLEMENTARY DATA

Supplementary data are available at *NSR* online.

## AUTHOR CONTRIBUTIONS

C.C., J.L., J.T., C.L., Y.X. and Z.Cheng conceived the project; C.C., J.L. and J.T. designed and supervised the project; Z.Chen, J.T. and H.W. analysed the data; Q.B., M.W., H.Liu, X.L., H.Li and J.F. collected plant material and phenotype data; G.L., H.Y. and J.L. performed the genomic and phenotypic analyses of ZKF5; Y.L., Z.T. and S.K. performed phenotypic testing of CSSLs; Z.Chen, J.T., J.L. and C.C. wrote the manuscript with contributions from all authors.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Peleman JD and van der Voort JR. Breeding by design. *Trends Plant Sci* 2003; **8**: 330–4.
2. Ikeda M, Miura K and Aya K *et al.* Genes offering the potential for designing yield-related traits in rice. *Curr Opin Plant Biol* 2013; **16**: 213–20.
3. Zeng D, Tian Z and Rao Y *et al.* Rational design of high-yield and superior-quality rice. *Nat Plants* 2017; **3**: 17031.
4. Xiao N, Pan C and Li Y *et al.* Genomic insight into balancing high yield, good quality, and blast resistance of japonica rice. *Genome Biol* 2021; **22**: 283.
5. Qian Q, Guo L and Smith SM *et al.* Breeding high-yield superior quality hybrid super rice by rational design. *Natl Sci Rev* 2016; **3**: 283–94.
6. Fu YB. Understanding crop genetic diversity under modern plant breeding. *Theor Appl Genet* 2015; **128**: 2131–42.
7. Wang F, Longkumer T and Catausan SC *et al.* Genome-wide association and gene validation studies for early root vigour to improve direct seeding of rice. *Plant Cell Environ* 2018; **41**: 2731–43.
8. Prochnow A, Risius H and Hoffmann T *et al.* Does climate change affect period, available field time and required capacities for grain harvesting in Brandenburg, Germany? *Agric For Meteorol* 2015; **203**: 43–53.
9. Hui S, Li H and Mawia AM *et al.* Production of aromatic three-line hybrid rice using novel alleles of BADH2. *Plant Biotechnol J* 2022; **20**: 59–74.
10. Xin F, Xiao X and Dong J *et al.* Large increases of paddy rice area, gross primary production, and grain production in North-east China during 2000–2017. *Sci Total Environ* 2020; **711**: 135183.
11. Liu X and Mu Y. A brief introduction to the regional distribution of rice crops in Fengtian Province in the late Qing Dynasty and the early Republic of China (in Chinese). *J Dalian Univ* 2017; **38**: 60–6.
12. He Y. Discussion on the history of rice production in Heilongjiang province (in Chinese). *Ancient Modern Agric* 2019; **53**: 54–66.
13. Liu H, Liu Y and Chen W. The reviews and enlightenments on the breeding and popularization of rice variety Ishikari-shiroge derived ones. *J Heilongjiang Bayi Agric Univ* 2011; **23**: 8–12.
14. Liu H, Wang J and Zhao H *et al.* Analysis of founder parents for cultivar varieties in Heilongjiang Province (in Chinese). *J Northeast Agric Univ* 2011; **42**: 18–21.
15. Zhou Y, Li N and Dong G *et al.* Impact assessment of recent climate change on rice yields in the Heilongjiang Reclamation Area of North-east China. *J Sci Food Agric* 2013; **93**: 2698–706.
16. Wang X, Yang X and Lv S *et al.* The possible effects of global warming on cropping systems in China XI: the possible effects of climate warming on geographical shift in safe planting area of rice in cold areas and the risk analysis of chilling damage (in Chinese). *Sci Agric Sin* 2016; **49**: 1859–71.
17. Chen Z, Li X and Lu H *et al.* Genomic atlases of introgression and differentiation reveal breeding footprints in Chinese cultivated rice. *J Genet Genomics* 2020; **47**: 637–49.
18. Wang W, Mauleon R and Hu Z *et al.* Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 2018; **557**: 43–9.
19. Pritchard JK, Stephens M and Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–59.
20. Li X, Liu H and Wang M *et al.* Combinations of Hd2 and Hd4 genes determine rice adaptability to Heilongjiang Province, northern limit of China. *J Integr Plant Biol* 2015; **57**: 698–707.
21. Yan W, Liu H and Zhou X *et al.* Natural variation in Ghd7.1 plays an important role in grain yield and adaptation in rice. *Cell Res* 2013; **23**: 969–71.
22. Gao H, Jin M and Zheng X-M *et al.* Days to heading 7, a major quantitative locus determining photoperiod sensitivity and regional adaptation in rice. *Proc Natl Acad Sci USA* 2014; **111**: 16337–42.
23. Xue W, Xing Y and Weng X *et al.* Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nat Genet* 2008; **40**: 761–7.
24. Yano K, Morinaka Y and Wang F *et al.* GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc Natl Acad Sci USA* 2019; **116**: 21262–7.
25. Yano K, Ookawa T and Aya K *et al.* Isolation of a novel lodging resistance QTL gene involved in strigolactone signaling and its pyramiding with a QTL gene involved in another mechanism. *Mol Plant* 2015; **8**: 303–14.
26. Wang Y, Shang L and Yu H *et al.* A strigolactone biosynthesis gene contributed to the green revolution in rice. *Mol Plant* 2020; **13**: 923–32.
27. Duan P, Xu J and Zeng D *et al.* Natural variation in the promoter of GSE5 contributes to grain size diversity in rice. *Mol Plant* 2017; **10**: 685–94.
28. Fan C, Xing Y and Mao H *et al.* GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor Appl Genet* 2006; **112**: 1164–71.
29. Shi CL, Dong NQ and Guo T *et al.* A quantitative trait locus GW6 controls rice grain size and yield through the gibberellin pathway. *Plant J* 2020; **103**: 1174–88.
30. Qu S, Liu G and Zhou B *et al.* The broad-spectrum blast resistance gene Pi9 encodes a nucleotide-binding site-leucine-rich repeat protein and is a member of a multigene family in rice. *Genetics* 2006; **172**: 1901–14.

31. Wang ZX, Yano M and Yamanouchi U *et al.* The Pib gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. *Plant J* 1999; **19**: 55–64.

32. Ookawa T, Hobo T and Yano M *et al.* New approach for rice improvement using a pleiotropic QTL gene for lodging resistance and yield. *Nat Commun* 2010; **1**: 132.

33. Huang X, Qian Q and Liu Z *et al.* Natural variation at the DEP1 locus enhances grain yield in rice. *Nat Genet* 2009; **41**: 494–7.

34. Gao ZY, Zhao SC and He WM *et al.* Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc Natl Acad Sci USA* 2013; **110**: 14492–7.

35. Asano K, Takashi T and Miura K *et al.* Genetic and molecular analysis of utility of sd1 alleles in rice breeding. *Breed Sci* 2007; **57**: 53–8.

36. Deng Y, Zhai K and Xie Z *et al.* Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. *Science* 2017; **355**: 962–5.

37. Jiang N, Li Z and Wu J *et al.* Molecular mapping of the Pi2/9 allelic gene Pi2-2 conferring broad-spectrum resistance to magnaporthe oryzae in the rice cultivar Jefferson. *Rice* 2012; **5**: 29.

38. Sun J, Liu D and Wang J-Y *et al.* The contribution of intersubspecific hybridization to the breeding of super-high-yielding japonica rice in northeast China. *Theor Appl Genet* 2012; **125**: 1149–57.

39. Bryan GT, Wu KS and Farrall L *et al.* A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *pi-ta*. *Plant Cell* 2000; **12**: 2033–46.

40. Zhao H, Wang X and Jia Y *et al.* The rice blast resistance gene Ptr encodes an atypical protein required for broad-spectrum disease resistance. *Nat Commun* 2018; **9**: 2039.

41. Zhang S, Sun X and Chen T *et al.* Distribution of Pi-ta, Pi-5, Pi-km and Pi-b genes in japonica rice varieties (lines) and their relationship with neck blast resistance (in Chinese). *Jiangsu J Agri Sci* 2018; **34**: 961–71.

42. Shen R, Wang L and Liu X *et al.* Genomic structural variation-mediated allelic suppression causes hybrid male sterility in rice. *Nat Commun* 2017; **8**: 1310.

43. Ando T, Yamamoto T and Shimizu T *et al.* Genetic dissection and pyramiding of quantitative traits for panicle architecture by using chromosomal segment substitution lines in rice. *Theor Appl Genet* 2008; **116**: 881–90.

44. Sha H, Liu H and Zhao G *et al.* Elite sd1 alleles in japonica rice and their breeding applications in northeast China. *Crop J* 2022; **10**: 224–33.

45. Wang RS, Jiang GQ and Feng XM *et al.* Updating the genome of the elite rice variety Kongyu131 to expand its ecological adaptation region. *Front Plant Sci* 2019; **10**: 288.

46. Ye J, Niu XJ and Yang YL *et al.* Divergent Hd1, Ghd7, and DTH7 alleles control heading date and yield potential of japonica rice in northeast China. *Front Plant Sci* 2018; **9**: 35.

47. Xie LY, Lin ED and Zhao HL *et al.* Changes in the activities of starch metabolism enzymes in rice grains in response to elevated $CO_2$ concentration. *Int J Biometeorol* 2016; **60**: 727–36.

48. Yu J, Miao J and Zhang Z *et al.* Alternative splicing of OsLG3b controls grain length and yield in japonica rice. *Plant Biotechnol J* 2018; **16**: 1667–78.

49. Li M, Zhang YW and Xiang Y *et al.* IIIVmrMLM: the R and C plus plus tools associated with 3VmrMLM, a comprehensive GWAS method for dissecting quantitative traits. *Mol Plant* 2022; **15**: 1251–3.

50. Crossa J, Perez-Rodriguez P and Cuevas J *et al.* Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci* 2017; **22**: 961–75.

51. Xu Y, Ma KX and Zhao Y *et al.* Genomic selection: a breakthrough technology in rice breeding. *Crop J* 2021; **9**: 669–77.

52. IRRI. *Standard Evaluation System For Rice (SES).* Manila: International Rice Research Institute, 2002.

53. Nagata K, Ando T and Nonoue Y *et al.* Advanced backcross QTL analysis reveals complicated genetic control of rice grain shape in a *japonica* x *indica* cross. *Breed Sci* 2015; **65**: 308–18.

54. Kawahara Y, de la Bastide M and Hamilton JP *et al.* Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice* 2013; **6**: 4.

55. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–60.

56. Li H, Handsaker B and Wysoker A *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–9.

57. DePristo MA, Banks E and Poplin R *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011; **43**: 491–8.

58. Alexander DH, Novembre J and Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1655–64.

59. Reich D, Price AL and Patterson N. Principal component analysis of genetic data. *Nat Genet* 2008; **40**: 491–2.

60. Paradis E and Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019; **35**: 526–8.

61. Malinsky M, Matschiner M and Svardal H. Dsuite—fast D-statistics and related admixture evidence from VCF files. *Mol Ecol Resour* 2021; **21**: 584–95.

62. Danecek P, Auton A and Abecasis G *et al.* The variant call format and VCFtools. *Bioinformatics* 2011; **27**: 2156–8.

63. Li X, Chen Z and Zhang G *et al.* Analysis of genetic architecture and favorable allele usage of agronomic traits in a large collection of Chinese rice accessions. *Sci China Life Sci* 2020; **63**: 1688–702.

64. Wei X, Qiu J and Yong K *et al.* A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat Genet* 2021; **53**: 243–53.

65. Qin P, Lu H and Du H *et al.* Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 2021; **184**: 3542–58.

66. Kang HM, Sul JH and Service SK *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 2010; **42**: 348–54.