nature genetics

ARTICLES

https://doi.org/10.1038/s41588-020-00769-9

A quantitative genomics map of rice provides genetic insights and guides breeding

Xin Wei[1,4], Jie Qiu[1,4], Kaicheng Yong[1], Jiongjiong Fan[1], Qi Zhang[1], Hua Hua[1], Jie Liu[1], Qin Wang[1], Kenneth M. Olsen[2], Bin Han[3] and Xuehui Huang[1]✉

Extensive allelic variation in agronomically important genes serves as the basis of rice breeding. Here, we present a comprehensive map of rice quantitative trait nucleotides (QTNs) and inferred QTN effects based on eight genome-wide association study cohorts. Population genetic analyses revealed that domestication, local adaptation and heterosis are all associated with QTN allele frequency changes. A genome navigation system, RiceNavi, was developed for QTN pyramiding and breeding route optimization, and implemented in the improvement of a widely cultivated *indica* variety. This work presents an efficient platform that bridges ever-increasing genomic knowledge and diverse improvement needs in rice.

To meet the challenges of rapid population growth and environmental change, plant scientists are facing enormous challenges to improve crops for higher yield, better quality and stronger stress resistance[1]. Genetic and genomic approaches have been used for the improvement of many crops[2,3], including one of the most important staple cereals, rice (*Oryza sativa* L.). Results from extensive functional genomic studies have been transferred into practical rice breeding. Some studies using marker-assisted selection based on genes that underlie quantitative traits (QTGs) for rice disease resistance, submergence tolerance and eating quality, were shown to be effective[4–8]. Recently, a molecular design approach was proposed for rice breeding[9], and validated by the development of new elite varieties through pyramiding multiple major QTGs underlying plant architecture, grain shape and starch synthesis[10].

While substantial progress has been achieved in rice breeding based on advances in genomics and genetics, further efforts are needed to close the gaps between genomic studies and practical breeding, so that breeding can be performed in a rapid, high-throughput and precise manner that takes full advantage of whole-genome information. To date, several hundred QTGs have been identified[11,12] and several thousand rice accessions have been sequenced[13–18]. However, only a few QTGs are clearly mapped in genome sequences, with clear allelic states and phenotypic effects being available for each accession. This situation is caused partly by the lack of integration of available QTG data, which are spread out over thousands of publications. Furthermore, many of these QTGs have not been converted to precise genomic information. In the field of human medical research, comprehensive information on numerous genetic disease-associated variants and oncogenesis-related driver mutations have been integrated and used successfully to interpret whole-genome sequencing results that were obtained for use in health risk assessments and clinical examinations[14]. In rice, in addition to many neutral trait-associated markers, further integration of the data of causal genetic variants underlying agronomic traits will enable precise breeding design[19]. This integrated information can also provide important basic insights into mechanisms of adaptation that occurred during rice domestication and improvement.

Hence, a universal, rapid and precise breeding system integrating the knowledge from genetic mapping and functional analyses of hundreds of QTGs is needed for the rice community. In the area of geoinformatics, the development of map navigation applications based on global positioning systems (GPS) has greatly facilitated the ability of car drivers to cope with complex road conditions. GPS map navigation involves locating the precise position of the driver, designing the optimum route to the destination and guiding the driver all the way to the destination. Analogously, we generated a rice gene encyclopedia containing all known trait-related causative variants, created a collection of rice varieties covering these variants and developed a genome navigation system for breeding. This bioinformatics system allows the user to determine the exact allelic status at hundreds of QTGs for any rice line, estimate the success rate of each given breeding route and select the best genotypes among the progeny in each breeding generation. We also tested the genome navigation system by successfully implementing a speedy and customized improvement for a well-known high-yield *indica* variety.

Results

A comprehensive catalog of causative variants. Through a comprehensive literature search, the abstracts of a total of 29,994 articles related to rice genes and quantitative trait loci (QTL) were downloaded and curated manually (Extended Data Fig. 1). Genes identified from artificial mutants (for example, ethylmethanesulfonate (EMS), T-DNA and gamma-ray-induced mutants), reverse genetic approaches (for example, RNA interference (RNAi), overexpression and clustered regularly interspaced short palindromic repeats (CRISPR)) and genetic mapping but without functional validation were all removed from the collection, as well as those specific to *Oryza glaberrima* (African cultivated rice) or wild rice (*Oryza rufipogon*). A total of 562 alleles in 225 QTGs were identified from 299 papers published from 1995 to 2020 (Supplementary Dataset 1). The QTGs included genes involved in grain yield, grain quality, stress tolerance and many other traits (Fig. 1a and Extended Data Fig. 2). Further analysis using a method based on an integration of multiple genomics approaches revealed 348 causative variants (that is, QTNs) corresponding to these QTGs, including 207 single nucleotide poly-

[1]Shanghai Key Laboratory of Plant Molecular Sciences, College of Life Sciences, Shanghai Normal University, Shanghai, China. [2]Department of Biology, Washington University in St Louis, St Louis, MO, USA. [3]National Center for Gene Research, CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China. [4]These authors contributed equally: Xin Wei, Jie Qiu. ✉e-mail: xhhuang@shnu.edu.cn

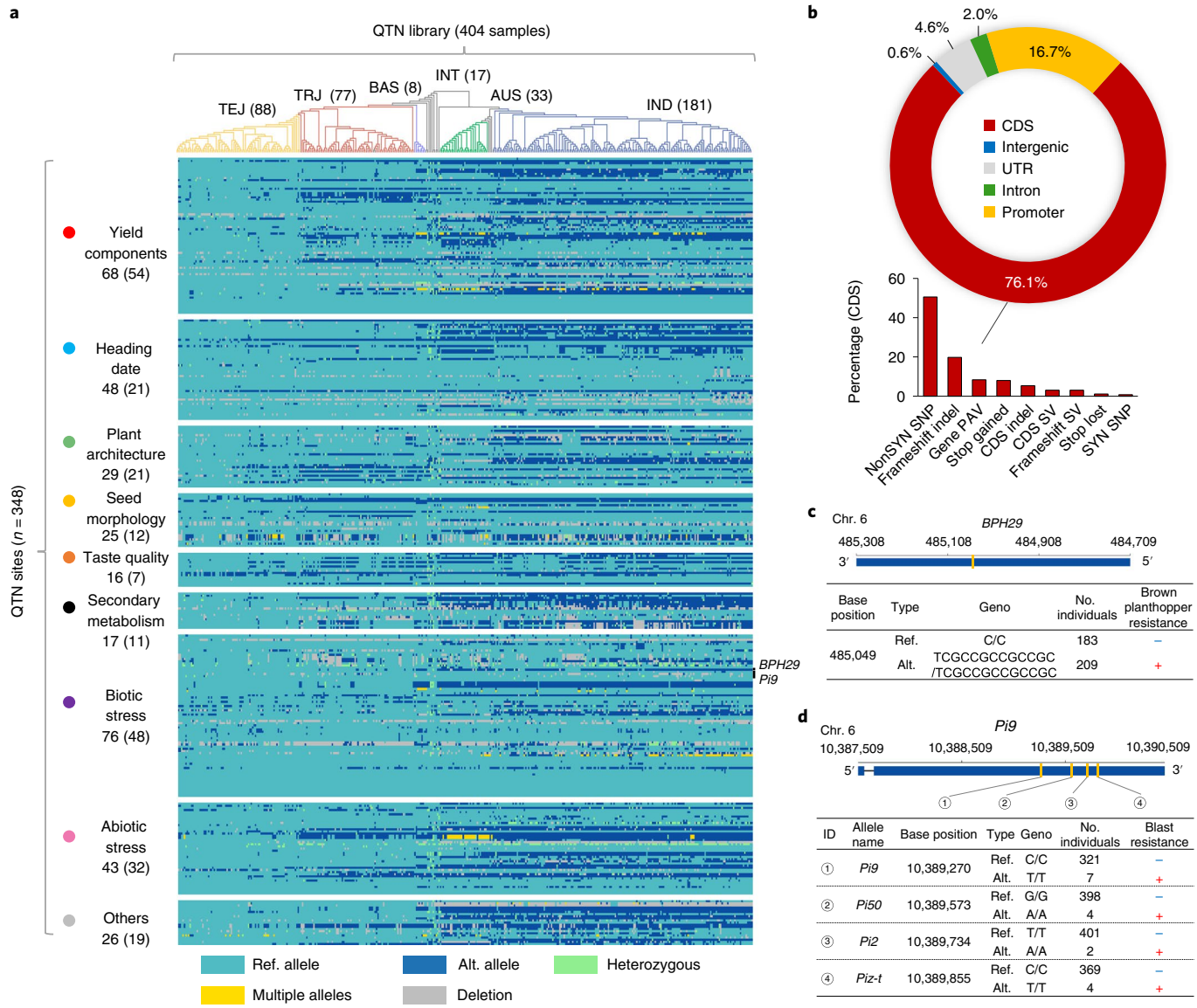


Fig. 1 | Genotype matrix of 225 QTGs for a collection of 404 rice accessions. a, A total of 348 causative variants for 225 QTGs (Supplementary Dataset 1) were genotyped for 404 collected rice samples (QTN library, Supplementary Dataset 2). The rice collection includes basmati, tropical *japonica*, temperate *japonica*, *indica*, aus and intermediate type (abbreviated as BAS, TRJ, TEJ, IND, AUS and INT, respectively) as shown in the phylogenetic tree. The genetic composition of BAS, TEJ, TRJ, IND and AUS are colored with purple, orange, red, blue, green and gray, respectively. The QTGs are classified into nine functional categories including Yield components, Heading date, Plant architecture, Seed morphology, Taste quality, Secondary metabolism, Biotic stress, Abiotic stress and Others. Numbers of QTNs and QTGs (in brackets) for each trait are shown on the left. Light blue, dark blue, light green, yellow and gray boxes represent the genotype for the Nipponbare reference (Ref.) allele, alternative (Alt.) allele, heterozygous, multiple alleles and deletion, respectively. **b**, The percentage summary for the genomic distributions of causative variants. CDS, coding sequence; indel, insertion-deletion; NonSYN, nonsynonymous; SYN, synonymous; PAV, presence and absence variation. The detailed percentage for effects of causative variants in the coding regions is listed below. **c**, The genomic location for the causative site of *BPH29* and its allelic distribution in the QTN library. Chr., chromosome. **d**, Causative sites of *Pi9* and their allele distributions in the QTN library.

morphisms (SNPs), 90 small insertions and deletions (indels) and 51 structural variants (SVs) (Fig. 1b). Here, SVs include large indels, presence and absence variation (PAV), and copy number variation in the rice genome. All of the causative variants were projected uniformly onto the rice reference genome (Nipponbare MSU 7.0). We examined the functional effects of these causative variants in the rice genome. Based on their locations and functional consequences, the 348 QTNs were classified into five types, of which the coding variant type was divided further into nine subtypes (Fig. 1b). Most causative variants (76.1%) are located within coding regions of the QTGs, including 48.5% with large effects on protein coding (for example, frameshift indel, stop codon gained). This distribution differs from maize, where the variation in nongenic regions makes a large contribution to quantitative traits[20]. This contrast may reflect differences in genetic architecture observed for many quantitative traits between rice and maize, where quantitative traits such as flowering time are controlled by a few large-effect QTLs in rice[21,22] but numerous small-effect QTLs in maize[23].

When several QTNs are located within the same gene, the QTG contains three or more allelic forms, which often results in

incremental differences in the target trait. A typical example of a multiple-allele gene is *Waxy*, which has multiple QTNs. There are, in total, eight *Waxy* alleles that affect amylose content, which ranges from <2% to 28% in rice grains[24]. Based on our comprehensive catalog of causative variants in the rice genome, we inspected the allelic forms of all QTGs. Of the 225 genes, 169 are bi-allelic (Fig. 1c) whereas 56 are represented by multiple alleles (Fig. 1d). Therefore, multiple-allele genes make up a substantial proportion of the available quantitative trait variation: 24.9% of the QTGs and 51.3% of the QTNs. As displayed in Extended Data Fig. 3a, many important genes related to stress resistance and heading date are represented by multiple alleles, and some of the alleles are of very low frequency in rice varieties (for example, the large deletion allele of *Ghd7* that shortens flowering time and plant height[25]).

A collection of diverse rice accessions containing the majority of the QTG alleles was constructed by collecting landraces and cultivars representing a wide geographic distribution and germplasm from a number of rice research groups, especially materials with rare alleles. From these collections, a QTN library consisting of a total of 404 rice accessions was established (Supplementary Dataset 2). The QTN library was purified by self-pollination and then sequenced with an average genomic coverage of 24.3× for each accession. Our phylogenetic analysis indicated that the library contains 181 *indica*, 88 temperate *japonica*, 77 tropical *japonica*, 33 aus and 8 basmati rice accessions. Because of the presence of SVs in the catalog and the difficulty of SV genotype calling, five variation calling methods were integrated to determine the QTN genotypes of the 404 accessions (Methods and Extended Data Fig. 1). In total, 95.5% of the alleles in the catalog were detected in the QTN library, demonstrating the large allelic diversity of the accession collection. In particular, the QTN library contains a number of rare but valuable alleles, including 50 accessions with rare alleles for disease resistance, 5 for yield increase, 13 for taste quality and 9 for abiotic stress resistance (Extended Data Fig. 3b). For example, among the 404 accessions, only 10 contain the increased-yield allele of *LAX1* (ref. [26]), and only 7 accessions contain the *Pi9* resistance allele[27]. These accessions are important donors to introduce rare but valuable alleles into gene pools of modern cultivars for rice breeding. In addition, based on the causative variant catalog, genome data of 3,010 diverse rice accessions[15] were used to examine QTN genotypes (Extended Data Fig. 4). Taking into consideration the QTNs of SNPs and indels, it was estimated that 90.1% alleles in the QTN catalog were included in the 3K collection[15]. The results indicate an enrichment of the reported QTNs despite the sample number ($n=404$) of the library being relatively small.

Estimation of QTN effects. We collected these large genomic and phenomic datasets and performed genome-wide association studies (GWASs) uniformly in eight cohorts, including the QTN library and seven GWAS cohorts that have been reported previously[15,17,21,28–31]. A total of 470 associated loci corresponding to 69 QTGs were identified (Fig. 2a and Supplementary Dataset 3). These QTGs underlie 50 agronomic traits that are associated mostly with yield components, heading date, plant architecture and taste quality. Although the phenotypes of the cohorts were analyzed in various environments, the major QTGs were identified repeatedly in multiple cohorts. For example, *GW5* and *GS3* are associated with seed morphology[32,33], *Waxy* with taste quality[34] and *Hd1* with heading date[35] (Fig. 2a). In contrast, QTGs with relatively minor effects and those with relatively low minor allele frequency tend to be identified in fewer cohorts. For example, *HESO1*, which is related to heading date, was identified only in the cohort of 176 Japanese rice accessions[17].

We then estimated allelic effects of the QTGs using the QTN maps and the phenotypic data in the eight GWAS cohorts. With correction for population structure, allelic effects were calculated for each of the 69 QTGs (Fig. 2b–f and Supplementary Dataset 3). Based on the quantitative estimation, a few QTGs showed very large effects, such as *sd1* and *dep1* in reducing plant height[36,37], *hd3a* in promoting heading date[38] and *ipa1* for increasing grain numbers[39]. In various environments, the estimated allelic effects of the same QTGs generally showed the same trends, but with different scales of effect (Extended Data Fig. 5). If we take plant height QTGs as an example, the trend of the allelic effects is taller or shorter but the difference in height determined by *OsSPY* (ref. [16]) varies from 3.70 cm to 18.56 cm depending on diverse cohort environments. For *Hd1*, the effect on heading date varies from −15.22 to 7.03 days in 11 locations that span a latitudinal gradient. This type of variation is probably due to the influences of population composition (for example, *indica* or *japonica* subspecies) and phenotyping environments (for example, long- or short-day conditions).

Meanwhile, phenotypic effects of the total 225 QTGs were examined according to the functional data from the 299 source papers (Supplementary Dataset 1) and recorded uniformly in a qualitative way. For example, the '12 bp insertion' allele in the QTN site of *BPH29* (ref. [40]) resulted in brown planthopper resistance (Fig. 1c) while the 'T' allele in the QTN site of *Pi9* led to stronger blast resistance (Fig. 1d). For 69 QTGs, the analyses of eight GWAS cohorts provided more precise effects. For example, the alternative alleles in *Hd3a* resulted in a decrease of 14.84 ± 1.60 days heading date in South China (~30° N), 7.69 ± 0.92 days in North China (~40° N) and 3.56 ± 0.49 days in Northeast China (~45° N). Through comparisons, the trend of the allelic effects of 94.1% QTGs estimated from GWAS is in the same direction as those from functional experiments, suggesting the effect estimation from the large GWAS datasets is reliable. Taken together, the allelic effect for each of the 69 QTGs was quantified whereas the effect of the remaining QTGs was described qualitatively, all of which were added into the QTG map annotations to facilitate breeding design.

Genetic drag in breeding. When plotting the QTGs onto rice chromosomes, we found their genomic distribution to be quite uneven. Based on the genetic map[41], many adjacent QTGs tended to co-occur within the same interval of 2 cM (Fig. 3a), which may lead to genetic drag. In other words, the introgression of superior alleles at some genes also introduced inferior alleles at linked loci during backcrossing due to high linkage but opposite phases. Hence, the potential genetic drag effects among the QTGs were investigated in a genome-wide manner based on the exact locations and the effect annotations in the QTN map. Any possible disadvantageous alleles located less than 2 Mb away from the advantageous allele were screened or examined in each line of the QTN library. On average, potential genetic drag was detected in about 25% of the genome in each line (Fig. 3b), with over 20 hotspots (Extended Data Fig. 6 and Supplementary Table 1). These results suggest that genetic drag is a common obstacle in rice breeding. Hence, special attention will be needed to break these linkages; this may be assisted by the use of molecular markers.

Furthermore, two major types of linkage drag were observed. One involves deleterious loci adjacent to QTGs with rare alleles but having opposite genetic phases. For instance, *TAC3* in most rice accessions is present as a superior allele[42], while the major alleles in *LOX-3* and *OsTB1* are typically inferior[43,44] (Fig. 3c). Thus, potential inferior alleles that are tightly linked to rare and valuable alleles should be taken into consideration when these rare alleles are introduced. The other common type of linkage drag observed is the occurrence of adjacent QTGs that are divergent in *indica* and *japonica*. For example, *Waxy*[24] and *BPH29* (ref. [40]) are represented by their superior alleles in *japonica* (93%) and *indica* (89%), respectively (Fig. 3c). For the latter type, the allelic distribution of all QTG was investigated among rice groups. According to the QTN map, in total 83 QTNs (36.9%, see Extended Data Fig. 3c) are highly differentiated between *indica* and *japonica* rice (for example, *COLD1*

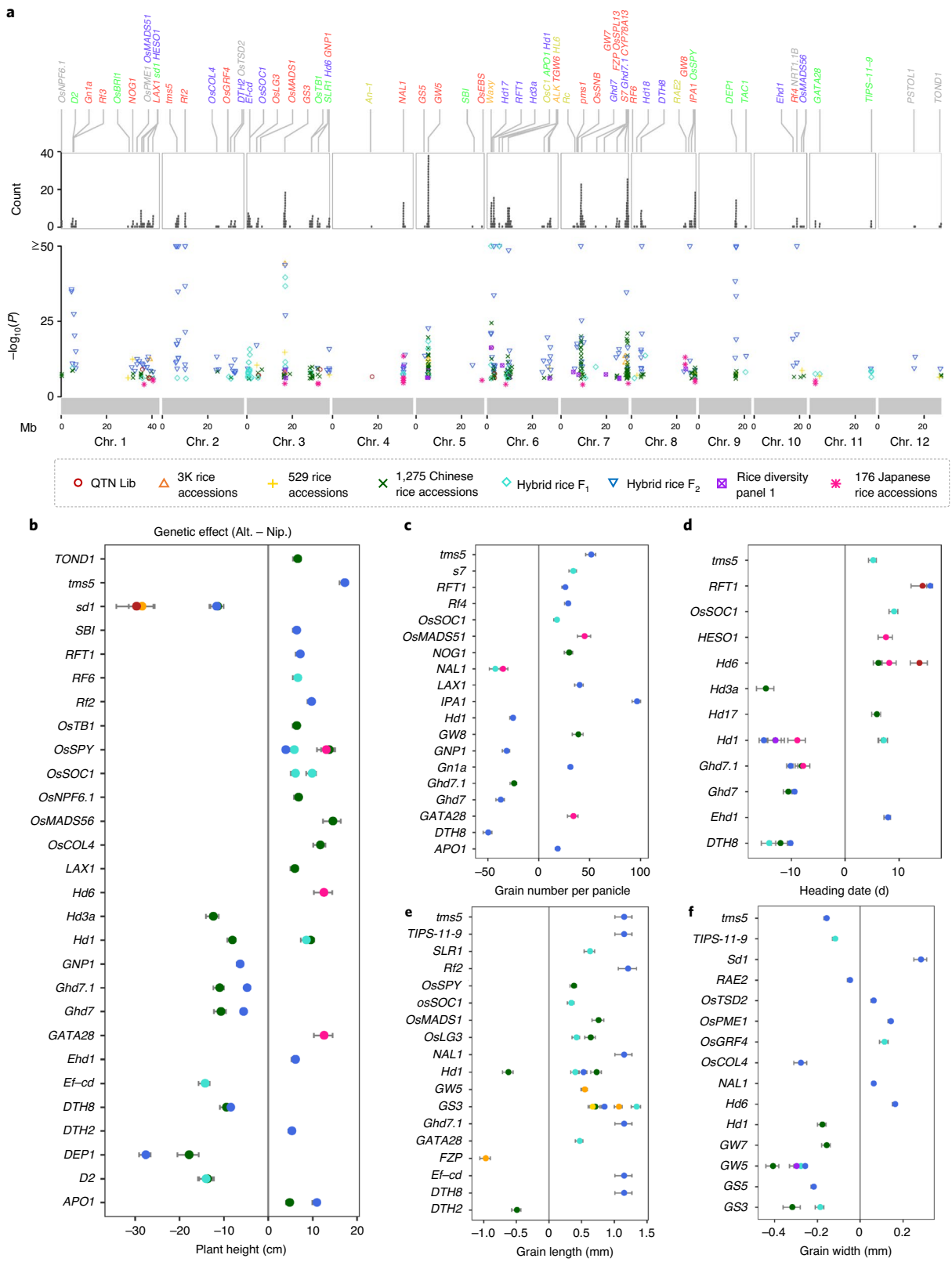


Fig. 2 | GWAS loci associated with important agronomic traits summarized from eight GWAS cohorts. a, Negative $\log_{10}$ *P* values (*y* axis) are plotted against positions (*x* axis) for significant peak SNPs summarized from eight GWAS cohorts. The peak significant SNPs from different cohorts are indicated by symbols with different shapes and colors as indicated in the legend box at the bottom of the figure. The QTGs that physically overlap with the GWAS peaks are labeled on top. The names are colored based on the functional categories as in Fig. 1a. **b**–**f**, The estimated phenotypic effects of homozygous alternative alleles relative to homozygous Nipponbare alleles from eight GWAS cohorts are shown jointly for each QTG. The bars indicate standard errors estimated by the GCTA software package. The phenotypes displayed include plant height (**b**), grain number per panicle (**c**), heading date (**d**), grain length (**e**) and grain width (**f**) under long-day conditions. For each cohort, the QTGs with most significant *P* value were collected and shown. The effects of *Hd1* under short-day conditions are shown. The full details of all QTG effects are listed in Supplementary Dataset 3.

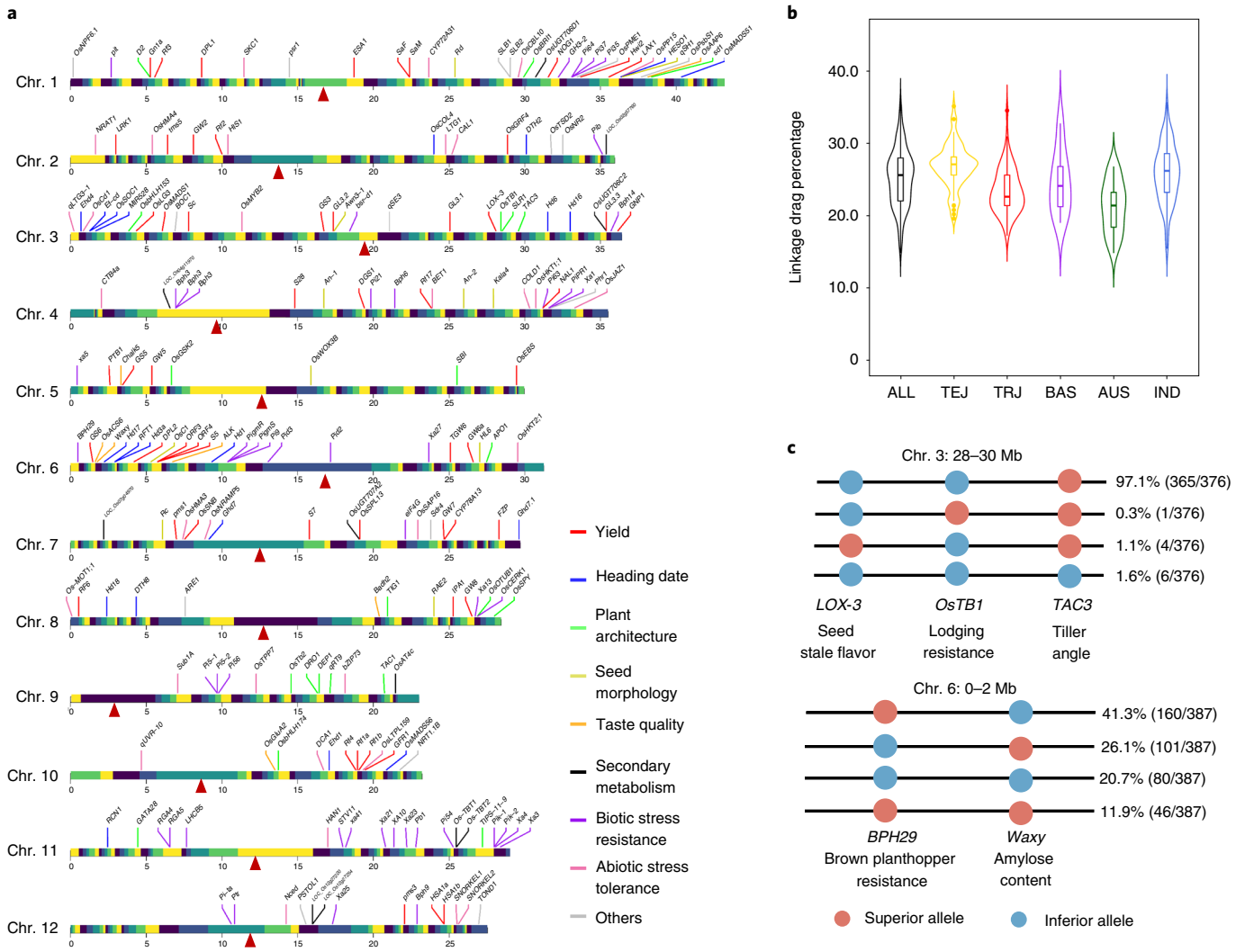


Fig. 3 | Genomic distribution and linkage drag for QTGs in rice genome. a, The genomic locations of 222 QTGs are illustrated along the 12 rice chromosomes (three mitochondrion QTGs not shown). The functional category of each QTG is indicated by the colored lines pointing to QTG labels. Each chromosome is colored in segments representing a genetic distance of 2 cM estimated from a previously constructed genetic map[41]. The positions for the centromeres of the 12 chromosomes are shown with a red triangle. **b**, Genome-wide percentage of linkage drag calculated for rice accessions in five groups. For the box-and-whisker plots, the middle line indicates the median value. The top whisker denotes the third quartile plus 1.5× the interquartile range (IQR), and the bottom whisker denotes the first quartile minus 1.5× the IQR. The points indicate outliers. **c**, Two examples of linkage drag identified from the QTN map. Top, a linkage drag genomic region containing *LOC-3*, *OsTB1* and *TAC3*. Bottom, another example of linkage drag, *BPH29* and *Waxy*. The percentage of each haplotype combination is listed on the right. The superior and inferior alleles are indicated by red and blue ovals, respectively.

with a major effect on chilling tolerance[45] and *NRT1.1B* with a major effect on nitrogen use efficiency[46]). Therefore, when subspecies intercrossing is performed for QTG pyramiding of *indica–japonica* differentiation traits, high-density genotyping is needed to trace recombination events in breeding populations to exclude highly linked and subspecies-specific inferior QTGs.

Genetic findings based on the high resolution of QTNs. We further examined the roles that QTNs play in genome evolution, heterosis and local adaptation. Conservation of the QTN sites was analyzed by Cnspipeline[47] and greenINSIGHT[48], respectively. Conservation scores and ρ scores of the QTNs were compared with variants within the same QTGs, and we observed significantly higher scores for QTNs, especially for nonsynonymous SNPs (nonSYN) and loss-of-function (LoF) variants in both methods (Fig. 4a), suggesting that NonSYN SNPs and LoF variations of QTNs tend to appear in the high conservation sites of the rice genome. SIFT (sorting intolerant from tolerant)[49] values were further calculated to evaluate the conservation of nonsynonymous SNPs, and we consistently found the same trend. For example, QTNs in *Waxy* and *Ehd1* had more conservative SIFT scores than other nonsynonymous SNPs (Fig. 4b). For the promoter QTNs, we found that the QTNs are likely to occur in regions that are close to the start site of the coding region (Extended Data Fig. 7a). In addition, 15 upstream region QTNs (including the promoter and 5′ untranslated region (UTR)) were detected in the open chromatin regions identified by ATAC- and FAIRE-seq (Extended Data Fig. 7b), indicating that QTNs are likely to reside in open chromatin regions. Therefore, the analysis suggests that plant conservation genomic maps[47,48] could facilitate determination of causative sites after candidate genes have been identified by QTL fine-mapping or GWAS. In addition, our constructed rice QTNs could also be a useful machine-learning training source to detect causative sites responsible for important agronomic traits for rice and other crops in the future.

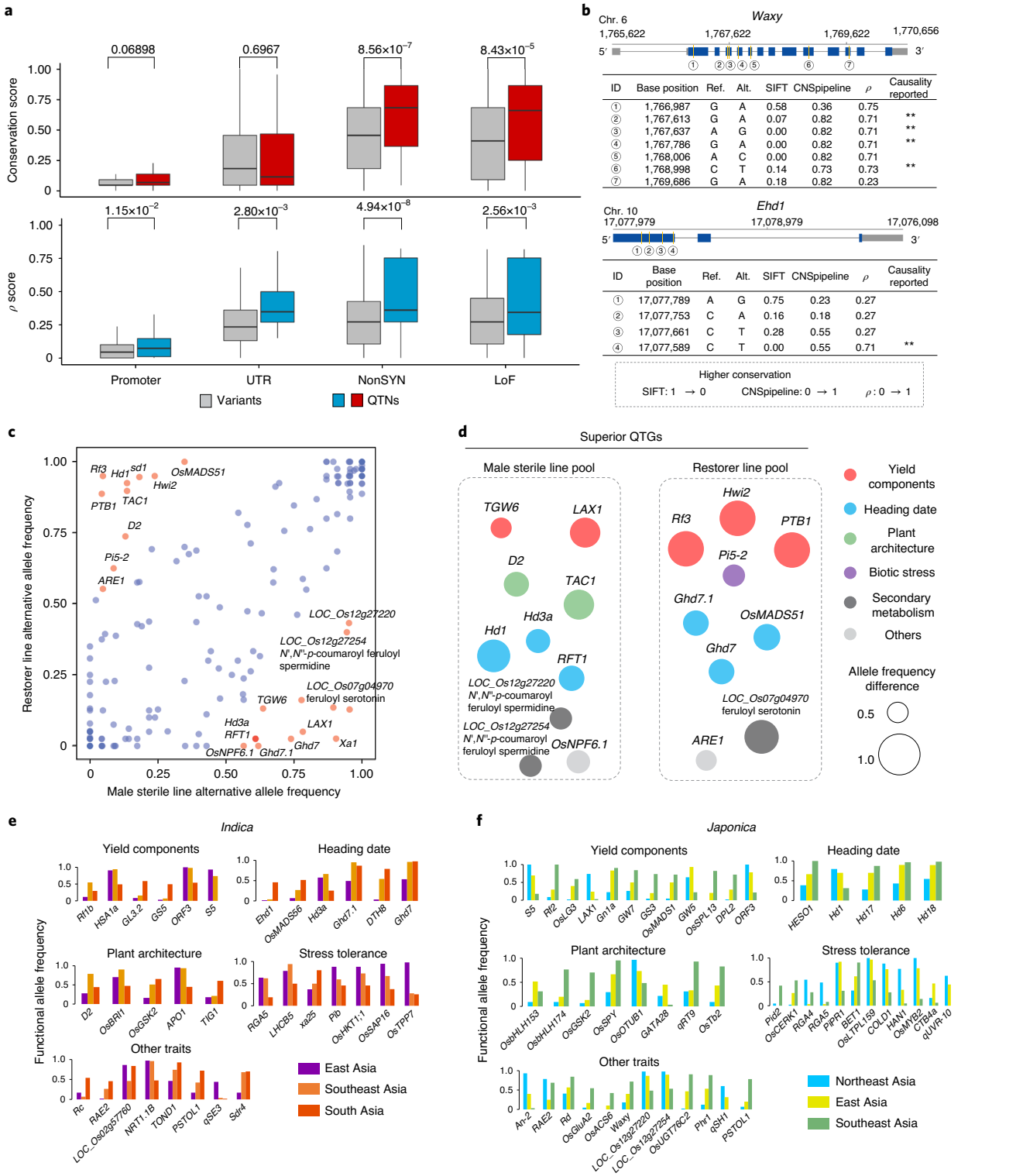


Fig. 4 | Genetic investigation of the QTNs. a, Comparison of conservation score and ρ score of QTNs and variants in the same QTGs. NonSYN and LoF refer to nonsynonymous SNPs and loss-of-function variations. For the box-and-whisker plots, the middle line indicates the median value. The top whisker denotes the third quartile plus 1.5× the IQR, and the bottom whisker denotes the first quartile minus 1.5× the IQR. The *P* values were calculated based on one-sided Wilcoxon test. **b**, SIFT score, conservation score and ρ score of nonsynonymous (NonSYN) SNPs in *Waxy* and *Ehd1*. High value of CNSpipeline and ρ scores or low value of SIFT indicate higher conservation of the variants. **c**, Alterative allele frequency of QTNs in male sterile lines and restorer lines of three-line hybrid rice. QTNs are indicated by solid circles, with orange or purple color indicating significantly (AFD > 0.4 and $P < 0.05$ by chi-squared test) or nonsignificantly differentiated QTNs, respectively. **d**, Superior allele frequency of QTGs in male sterile lines and restorer lines of three-hybrid rice. The alleles that contribute to high yield and strong disease resistance are defined as superior alleles. The QTNs are indicated by circles and the color represents group of traits. The size of the circle represents allele frequency. **e**, Allele frequency of QTNs in *indica* from East Asia, South Asia and Southeast Asia. **f**, Allele frequency of QTNs in *japonica* from East Asia, Southeast Asia and Northeast Asia. All QTGs shown have highly differentiated allele frequency (AF > 0.4). Detailed functional allele frequencies of each QTG are listed in Supplementary Datasets 4 and 5.

Allele frequencies of all QTNs in two heterotic groups (the restorer lines and male sterile lines of three-line hybrid rice) were surveyed to determine heterosis-related QTGs (Fig. 4c). Nine QTGs with superior alleles in restorer lines and ten QTGs with superior alleles in male sterile lines were identified, respectively (Fig. 4d). Among the ten QTGs in male sterile lines, three (*LAX1*, *TAC1* and *Hd3a*) have been identified previously through genetic mapping to be crucial for heterosis[21]. Besides these QTGs, *D2*, *TGW6*, *Hd1*, *RFT1*, *OsNPF6.1*, *LOC_Os12g27220* and *LOC_Os12g27254,* which contribute to the erect tiller, larger grains, later heading date, higher nitrogen use efficiency and stronger abiotic stress tolerance, respectively, were also identified in male sterile lines. Furthermore, QTNs that were related to higher seed setting rate, stronger blast resistance, later heading date and higher nitrogen use efficiency were found in restorer lines such as *PTB1*, *Pi5-2*, *Ghd7* and *ARE1*. All of these QTGs together result in higher yield of hybrid rice. The new findings will be important resources for rice heterosis research and provide strong support for further study in the near future.

The ability of rice to adapt to local growing environments was critical to the crop's expansion beyond its center of domestication (subtropical Asia) and into rice-farming regions worldwide. Based on the analysis of allele frequencies for QTNs in different areas of rice cultivation across Asia, we found that the allele frequency of 75 QTNs related to environmental adaptation varied by regions between *indica* and *japonica* (Supplementary Datasets 4 and 5). For *indica*, accessions in East Asia have more early heading date alleles, more blast disease resistance alleles and stronger resistance to low-temperature germination, but fewer alleles for high mineral nutrition use efficiency (Fig. 4e), which is in line with the long day-length, serious disease stresses, low temperature and heavy use of fertilizer in this region. In parallel, higher cold tolerance and early heading date allele frequencies were found in *japonica* from Northeast Asia (Fig. 4f), and this pattern is consistent with the low temperature and long day-length in the planting season in Northeast Asia. These results indicate that evolution of QTGs have played a major role as targets of natural and artificial selection during the course of rice domestication and subsequent spread of the crop, allowing adaptation to the various environmental conditions where rice is now cultivated.

Allele frequencies of QTNs in wild rice, landraces and modern cultivars were calculated and the neutrally evolved four-fold degenerate (4DTv) allele frequency change was used as background to control for genetic drift during domestication or improvement (details in Methods). In total, 99 QTGs were identified as targets of selection for domestication-related (including early improvement traits) or modern improvement-related genes. The number of domestication-related QTGs identified in *japonica* exceeded those in *indica* in number (36 and 28, respectively, Extended Data Fig. 8a and Supplementary Dataset 6), while the number of QTGs involved in modern breeding of *indica* exceeded that of *japonica* (40 and 23, respectively). Furthermore, overlaps of domestication/improvement-related QTGs between *indica* and *japonica* were observed (Extended Data Fig. 8b); this finding agrees with previous reports of shared targets of selection between the two rice subspecies[13,50–52]. Typical domestication genes in both *indica* and *japonica* domestication include *sh4*, *OsLG1*, *An-2* and *Sdr4*, while shared loci associated with later improvement include *Waxy*, *ALK*, *OsC1* and *GATA28*.

By categorizing QTGs according to the phenotypes they control, we found that heading date, plant architecture, seed morphology and taste quality-related QTGs accounted for the largest proportion of loci that were targets of selection during domestication or improvement (Extended Data Fig. 8c). For example, among seven major QTGs that underlie eating quality, six have been targets of selection. Similarly, two-thirds (14 of 21) of heading date-related QTGs had been selection targets and 78.6% (11 of 14) were selected to be early heading date alleles. These results revealed that QTGs responsible for early heading date and favorable eating quality have been continuous selection targets during the course of rice domestication and improvement, with the improved phenotypes arising through progressive, cumulative genetic change involving multiple loci. The polygenic basis for these improvement traits differs from many traits selected earlier in the domestication process, which are often controlled by a relatively few genes of major effect[53–55].

Notably, domestication and improvement did not always lead to accumulation of superior alleles of rice QTGs. Consistent with the genetic drag effects described above, we observed that inferior alleles constituted 46.4%, 30.6%, 22.5% and 16.7% of selectively favored alleles during *indica* domestication, *japonica* domestication, *indica* improvement and *japonica* improvement, respectively (Extended Data Fig. 8d). For example, brown-planthopper-susceptible alleles of *BPH29* were selected during the improvement of *Waxy* in *indica* improvement. This suggests that genetic drag has been pervasive in rice domestication and improvement.

Development and implementation of the RiceNavi system. To determine the key factors related to breeding route optimization, we performed in silico breeding by simulation with various scenarios (Supplementary Note). Benchmarking experiments indicate that use of the incremental mode, more backcrossing, larger populations and choosing QTG in regions with high recombination rates would help to boost breeding efficiency (Fig. 5). Similar to the GPS map and route planning algorithms in vehicle navigation systems, our constructed QTN map and breeding route optimization paved the way for us to develop a rice genome navigation system: RiceNavi (Fig. 6). Three main modules, including RiceNavi-QTNpick, -Sim and -SampleSelect, were created and loaded into the RiceNavi system (Supplementary Note).

RiceNavi was then applied to the genetic improvement of an elite *indica* variety Huanghuazhan (HHZ) for favorable rice fragrance, better adaptation to condensed planting and shorter growth period (Fig. 7a). Following RiceNavi guidance, we crossed the donor line (Basmati Surkh 89-15, QTG: *Badh2* (ref. [56]), *TAC1* (ref. [57]) and *OsSOC1* (ref. [58])) to HHZ and genotyped the plants in each backcross generation through whole-genome low-coverage multiplex sequencing (Fig. 7b). The introgression lines selected by RiceNavi with target genotypes (Extended Data Fig. 9 and Supplementary Fig. 1) were planted in southern China (Sanya, short-day conditions) and mid-latitude China (Shanghai, long-day conditions) for phenotyping. As expected, the introgression lines showed tight plant type, basmati-specific fragrance, earlier flowering time (short-day conditions), more tillers (long-day conditions) and higher yield in condensed planting conditions (Fig. 7 and Supplementary Table 2). More importantly, owing largely to the development of RiceNavi, we spent only 2.5 years to complete the entire process without introducing any linkage drag, which was much quicker and more precise than conventional breeding (typically >5 years[59]). Furthermore, we quantitatively evaluated the performances of applying the QTN-based breeding scheme, and the results showed the majority of predicted effects from QTNs are fully or partially realized (Supplementary Note, Extended Data Fig. 10 and Supplementary Dataset 7). Taken together, the RiceNavi system is able to guide breeding design in rice.

Discussion

The new genomic approach developed in the present study and implemented in the RiceNavi system enabled optimum designs of new varieties with precision and high efficiency. This platform was established based on the collective data from the abundant publications on rice genetics, population genomics and GWAS cohorts, and can be integrated efficiently with rice genomic data updates in the future. To improve any trait in a particular rice genetic

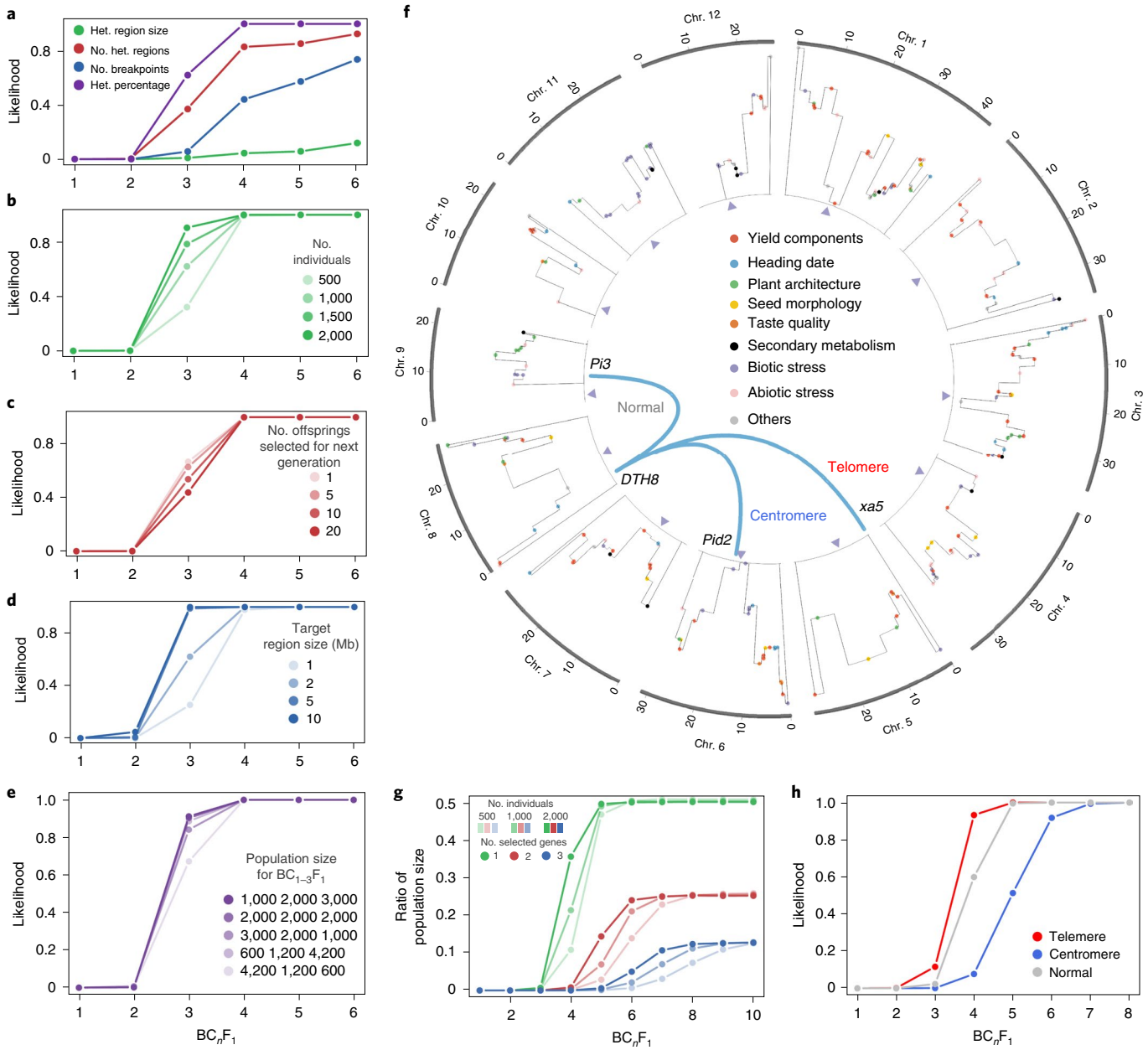


Fig. 5 | Benchmarking for different backcrossing breeding designs. The benchmarking was performed in silico by the RiceNavi software developed in this study. For each in silico experiment, 1,000 simulations were performed, with each simulation starting from F_1 to BC_nF_1 generation. **a**, The likelihood when using four different selection parameters for selection. For each generation, five individuals in BC_mF_1 were selected and crossed with the background parent to generate a $BC_{m+1}F_1$ population. In each generation, the likelihood is calculated based on the percentage of simulations that contain 'ideal' individuals. The 'ideal' individual is the one with only heterozygous (Het.) genotypes in target regions (<2 Mb) covering selected QTGs. **b**–**d**, Using low heterozygosity as the key factor, diverse parameters were then examined, which include population size (**b**), number of offspring selected for the next generation (**c**) and the size of genomic region that covers the selected gene(s) (**d**). **e**, Estimation for how the change of population size affected the likelihood. **f**, The relative likelihoods for each QTG locating in diverse genomic regions are visualized in the circos plot. The scattered points are colored according to their classified functions. **g**, Estimation of how the number of genes selected and population size affect the number of breeding generations. **h**, Estimation of the likelihood when one gene (*DTH8* in this case) is selected jointly with another gene locating in centromere (*Pid2*), telomere (*xa5*) and another region (*Pi3*).

line, it is important to know the corresponding genes and all the causative variant information related to the trait of interest. In the past 20 years, a large proportion of QTGs have been identified in rice, including those with major or modest effects, providing a basis for designed rice breeding. The reported QTGs explained a large proportion of phenotypic variation for yield-related traits[21]. The momentum of rice functional genomics studies remains strong, with >30 QTGs reported per year in the past 2 years. With continuous updates, the genome navigation system will be become an increasingly precise and powerful tool for rice breeding, including hybrid rice breeding. The key genes and loci contributing to yield heterosis, hybrid sterility and fertility restoration[21,29,60–65] have been identified and are already included in RiceNavi. Using the RiceNavi system, parental lines, including sterile, maintainer and restorer lines, can be improved by well-designed routes to keep a strong level of heterosis.

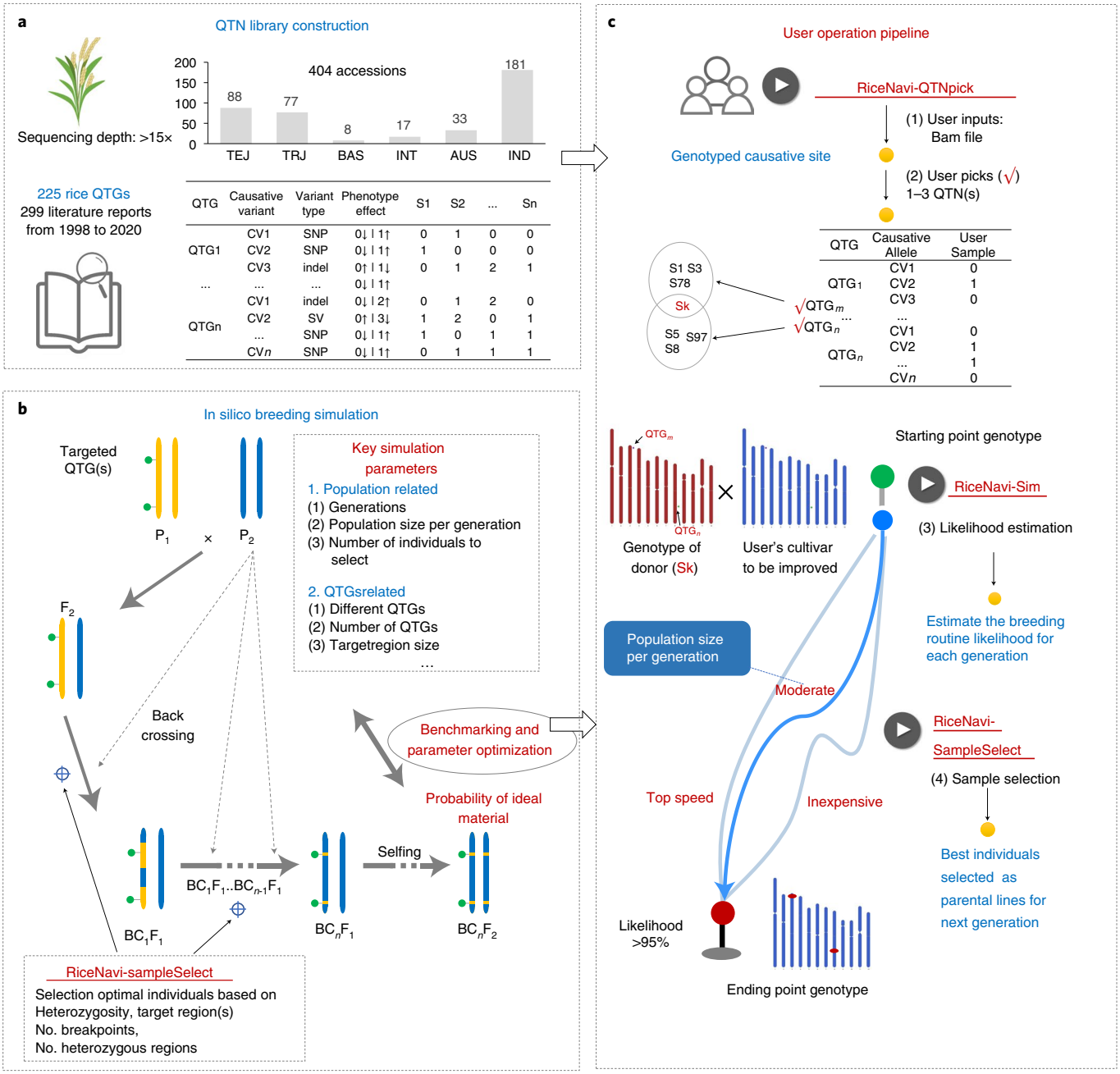


Fig. 6 | A schematic diagram for RiceNavi implementation and user operation. a, Construction of QTN library. A QTN library matrix table was constructed based on the 348 genotypes of causative sites for the 404 rice accessions. The phenotypic effects for causative alleles compiled from 299 publications from 1995 to 2020 are also listed in the matrix. **b**, In silico breeding simulation by RiceNavi. A backcrossing breeding process simulated by RiceNavi-Sim. During each backcrossing step, individuals with heterozygous loci covering the selected genes of the donor parent (P_1) are selected by RiceNavi-SampleSelect to backcross with the background parent (P_2). Diverse simulation parameters are set for benchmarking of their effect on the probability to obtain the ideal genotype. **c**, User operation pipeline of RiceNavi. The GATK4 gvcf file of the user's rice accession (or one accession in QTN library) is loaded into the RiceNavi program, and the genotype for each QTN site of the user's accession is compared with that of 404 rice accessions. Users can pick 1–3 QTNs and select one individual with these superior alleles. After choosing the donor individual from our QTN library, RiceNavi-Sim can help the user to estimate the likelihood of producing the ideal plant material given a selected population size per generation. During real breeding process, RiceNavi-SampleSelect can help the user to select ideal individuals as parental lines for next generation.

Because of the limited information on potential genetic interactions available at present, the outcome cannot be predicted precisely when pyramiding multiple QTNs underlying the same trait due to the complexities of epistasis/environmental interactions. Therefore, the proposed breeding strategy used in this study involves 'slight modification', that is, the creation of near-isogenic lines (NILs) with only a few QTNs introduced at a time (for example, typically, one QTN for grain yield and another QTN for disease resistance). Even so, in our quantitative evaluation, some NILs did not display expected performance, probably due to epistasis and environmental interactions or QTN effect errors. Hence, a better understanding of the genetic network in rice traits, including QTN–QTN interactions

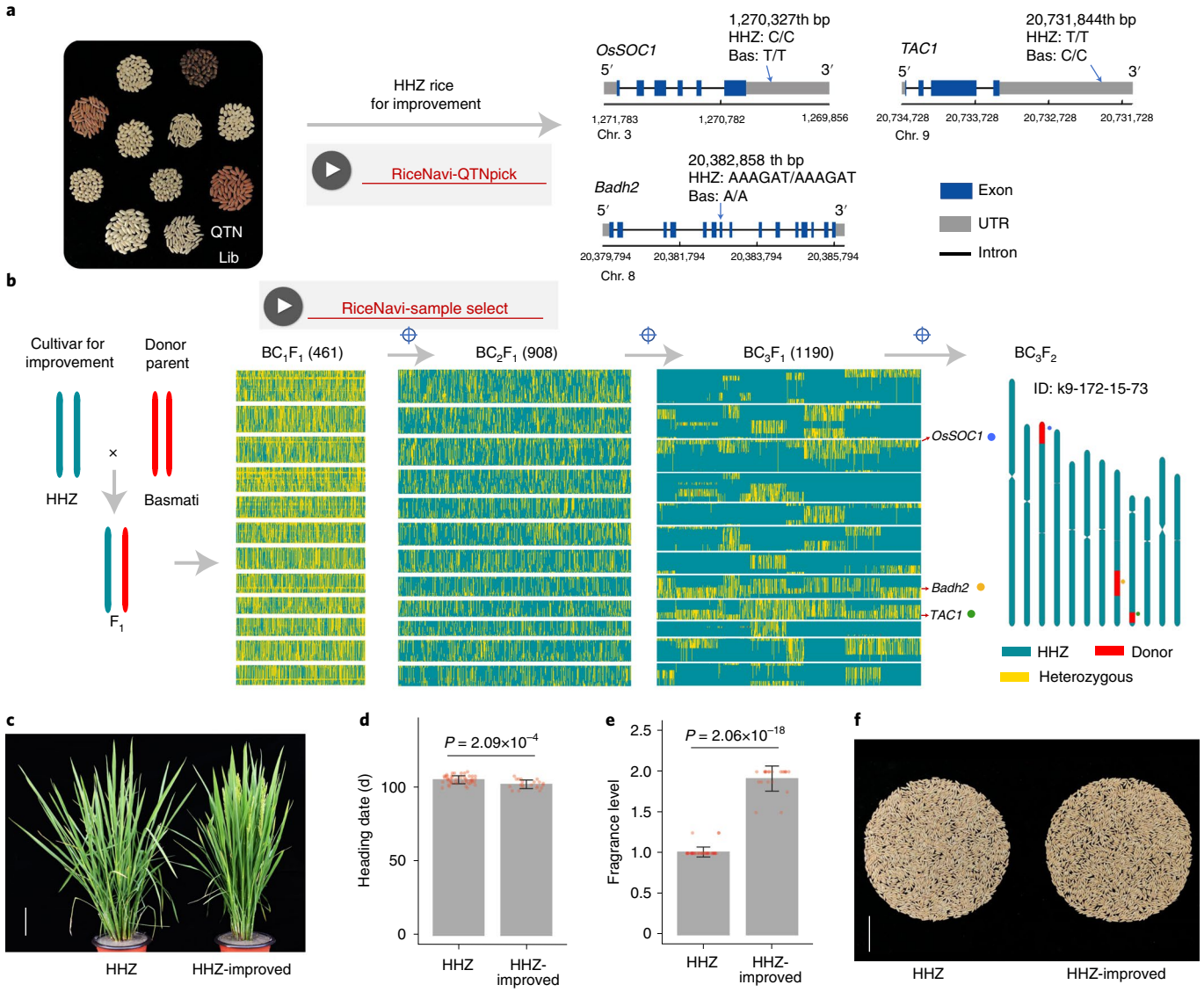


Fig. 7 | Improvement of the Huanghuazhan cultivar by implementation of RiceNavi. a, Detection of beneficial alleles in QTN library for improvement of Huanghuazhan (HHZ) by RiceNavi-QTNpick. As the first step, RiceNavi-QTNpick was used for comparisons of genotypes between HHZ and accessions in the QTN library (Lib). One rice accession (Basmati) was found with three QTGs (*OsSOC1*, *Badh2* and *TAC1*) harboring superior alleles for improvement of HHZ. **b**, A backcrossing breeding process with Basmati as the donor parent for improving three QTGs in HHZ. The selected donor parent (Basmati) was crossed with HHZ to construct the BC_nF_1 populations. The samples of each generation were low-pass sequenced for genotyping. The genotypes for the HHZ background, donor Basmati and heterozygous are color coded as light blue, red and yellow, respectively. Individuals with heterozygous genomic segments covering the selected three genes are suggested by RiceNavi-SampleSelect and chosen manually as backcrossing parents for next generation. In the BC_3F_2 population, one individual is found with the homologous donor alleles (red) only at the segments covering the selected three QTGs. **c**, The plant architecture of HHZ and HHZ-improved (scale bar, 10 cm), the latter showing a narrowed tiller angle. **d**,**e**, Comparison of heading date (**d**) and fragrance level (**e**) between HHZ (left) and HHZ-improved (right) in Sanya. Error bars indicate standard deviations. The *P* values are calculated based on two-tailed *t*-test. **f**, Comparison of grain size and grain yield per plant for HHZ and HHZ-improved in Sanya; scale bar, 10 cm. A comparison of all traits is provided in Supplementary Table 2.

and QTN–environment interactions, will improve the precision of phenotypic prediction. Once the genetic data of these interactions is available, the information will be added to the RiceNavi platform for updating. We expect that advances in functional genomics, including more investigation into epistatic interactions, will enhance further data integration and system updates, promoting a more powerful breeding guiding system for rice.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00769-9.

Received: 20 May 2020; Accepted: 11 December 2020;
Published online: 1 February 2021

References

1. Hickey, L. T. et al. Breeding crops to feed 10 billion. *Nat. Biotechnol.* **37**, 744–754 (2019).

2. Takeda, S. & Matsuoka, M. Genetic approaches to crop improvement: responding to environmental and population changes. *Nat. Rev. Genet.* **9**, 444–457 (2008).
3. Wallace, J. G., Rodgers-Melnick, E. & Buckler, E. S. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu. Rev. Genet.* **52**, 421–444 (2018).
4. Hasan, M. M. et al. Marker-assisted backcrossing: a useful method for rice improvement. *Biotechnol. Biotechnol. Equip.* **29**, 237–254 (2015).
5. Septiningsih, E. M. et al. Development of submergence-tolerant rice cultivars: the Sub1 locus and beyond. *Ann. Bot.* **103**, 151–160 (2009).
6. Singh, S. et al. Pyramiding three bacterial blight resistance genes (*xa5*, *xa13* and *Xa21*) using marker-assisted selection into indica rice cultivar PR106. *Theor. Appl. Genet.* **102**, 1011–1015 (2001).
7. Suh, J.-P. et al. Development of resistant gene-pyramided *japonica* rice for multiple biotic stresses using molecular marker-assisted selection. *Plant Breed. Biotech.* **3**, 333–345 (2015).
8. Chen, T. et al. Genetic improvement of *japonica* rice variety Wuyujing 3 for stripe disease resistance and eating quality by pyramiding *Stv-bi* and *Wx-mq*. *Rice Sci.* **23**, 69–77 (2016).
9. Qian, Q., Guo, L., Smith, S. M. & Li, J. Breeding high-yield superior quality hybrid super rice by rational design. *Natl Sci. Rev.* **3**, 283–294 (2016).
10. Zeng, D. L. et al. Rational design of high-yield and superior-quality rice. *Nat. Plants* **3**, 17031 (2017).
11. Ikeda, M., Miura, K., Aya, K., Kitano, H. & Matsuoka, M. Genes offering the potential for designing yield-related traits in rice. *Curr. Opin. Plant Biol.* **16**, 213–220 (2013).
12. Li, Y. et al. Rice functional genomics research: past decade and future. *Mol. Plant* **11**, 359–380 (2018).
13. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
14. Knoppers, B. M., Zawati, M. H. & Senecal, K. Return of genetic testing results in the era of whole-genome sequencing. *Nat. Rev. Genet.* **16**, 553–559 (2015).
15. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
16. Yano, K. et al. GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc. Natl Acad. Sci. USA* **116**, 21262–21267 (2019).
17. Yano, K. et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* **48**, 927–934 (2016).
18. Zhao, Q. et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284 (2018).
19. Ramstein, G. P., Jensen, S. E. & Buckler, E. S. Breaking the curse of dimensionality to identify causal variants in Breeding 4. *Theor. Appl. Genet.* **132**, 559–567 (2019).
20. Li, X. et al. Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res.* **22**, 2436–2444 (2012).
21. Huang, X. et al. Genomic architecture of heterosis for yield traits in rice. *Nature* **537**, 629–633 (2016).
22. Huang, X. et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **44**, 32–39 (2012).
23. Buckler, E. S. et al. The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
24. Zhang, C. et al. *Wx(lv)*, the ancestral allele of rice *Waxy* gene. *Mol. Plant* **12**, 1157–1166 (2019).
25. Xue, W. et al. Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat. Genet.* **40**, 761–767 (2008).
26. Gao, Z.-Y. 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* **110**, 14492–14497 (2013).
27. Qu, S. H. 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* **172**, 1901–1914 (2006).
28. Zhao, K. et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2**, 467 (2011).
29. Huang, X. et al. Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat. Commun.* **6**, 6258 (2015).
30. Xie, W. et al. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc. Natl Acad. Sci. USA* **112**, E5411–E5419 (2015).
31. Li, X. 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.* **63**, 1688–1702 (2020).
32. Shomura, A. et al. Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**, 1023–1028 (2008).
33. Fan, C. 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.* **112**, 1164–1171 (2006).
34. Wang, Z. Y. et al. The amylose content in rice endosperm is related to the post-transcriptional regulation of the *waxy* gene. *Plant J.* **7**, 613–622 (1995).
35. Yano, M. et al. *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering time gene *CONSTANS*. *Plant Cell* **12**, 2473–2484 (2000).
36. Sasaki, A. et al. A mutant gibberellin-synthesis gene in rice. *Nature* **416**, 701–702 (2002).
37. Huang, X. et al. Natural variation at the *DEP1* locus enhances grain yield in rice. *Nat. Genet.* **41**, 494–497 (2009).
38. Kojima, S. et al. *Hd3a*, a rice ortholog of the *Arabidopsis FT* gene, promotes transition to flowering downstream of *Hd1* under short-day conditions. *Plant Cell Physiol.* **43**, 1096–1105 (2002).
39. Zhang, L. et al. A natural tandem array alleviates epigenetic repression of *IPA1* and leads to superior yielding rice. *Nat. Commun.* **8**, 14789 (2017).
40. Wang, Y. et al. Map-based cloning and characterization of *BPH29*, a B3 domain-containing recessive gene conferring brown planthopper resistance in rice. *J. Exp. Bot.* **66**, 6035–6045 (2015).
41. Huang, X. et al. High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19**, 1068–1076 (2009).
42. Dong, H. et al. A novel tiller angle gene, *TAC3*, together with *TAC1* and *D2* largely determine the natural variation of tiller angle in rice cultivars. *PloS Genet.* **12**, e1006412 (2016).
43. Shirasawa, K., Takeuchi, Y., Ebitani, T. & Suzuki, Y. Identification of gene for rice (*Oryza sativa*) seed lipoxygenase-3 involved in the generation of stale flavor and development of SNP markers for lipoxygenase-3 deficiency. *Breed. Sci.* **58**, 169–176 (2008).
44. Yano, 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* **8**, 303–314 (2015).
45. Ma, Y. et al. *COLD1* confers chilling tolerance in rice. *Cell* **160**, 1209–1221 (2015).
46. Hu, B. et al. Variation in *NRT*1.1B contributes to nitrate-use divergence between rice subspecies. *Nat. Genet.* **47**, 834–838 (2015).
47. Liang, P. P., Saqib, H. S. A., Zhang, X. T., Zhang, L. S. & Tang, H. B. Single-Base resolution map of evolutionary constraints and annotation of conserved elements across major grass genomes. *Genome Biol. Evol.* **10**, 473–488 (2018).
48. Joly-Lopez, Z. et al. An inferred fitness consequence map of the rice genome. *Nat. Plants* **6**, 119–130 (2020).
49. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
50. Molina, J. et al. Molecular evidence for a single evolutionary origin of domesticated rice. *Proc. Natl Acad. Sci. USA* **108**, 8351–8356 (2011).
51. Choi, J. Y. et al. The rice paradox: multiple origins but single domestication in Asian rice. *Mol. Biol. Evol.* **34**, 969–979 (2017).
52. Choi, J. Y. & Purugganan, M. D. Multiple origin but single domestication led to *Oryza sativa*. *G3 (Bethesda)* **8**, 797–803 (2018).
53. Li, C. B., Zhou, A. L. & Sang, T. Rice domestication by reducing shattering. *Science* **311**, 1936–1939 (2006).
54. Jin, J. et al. Genetic control of rice plant architecture under domestication. *Nat. Genet.* **40**, 1365–1369 (2008).
55. Ishii, T. et al. OsLG1 regulates a closed panicle trait in domesticated rice. *Nat. Genet.* **45**, 462–465 (2013).
56. Chen, S. et al. *Badh2*, encoding betaine aldehyde dehydrogenase, inhibits the biosynthesis of 2-acetyl-1-pyrroline, a major component in rice fragrance. *Plant Cell* **20**, 1850–1861 (2008).
57. Yu, B. et al. *TAC1*, a major quantitative trait locus controlling tiller angle in rice. *Plant J.* **52**, 891–898 (2007).
58. Lin, H., Ashikari, M., Yamanouchi, U., Sasaki, T. & Yano, M. Identification and characterization of a quantitative trait locus, *Hd9*, controlling heading date in rice. *Breed. Sci.* **52**, 35–41 (2002).
59. Li, J. et al. A practical protocol to accelerate the breeding process of rice in semitropical and tropical regions. *Breed. Sci.* **65**, 233–240 (2015).
60. Chen, J. et al. Genome-wide association analyses reveal the genetic basis of combining ability in rice. *Plant Biotechnol. J.* **17**, 2211–2222 (2019).
61. Li, D. et al. Integrated analysis of phenome, genome, and transcriptome of hybrid rice uncovered multiple heterosis-related loci for yield increase. *Proc. Natl Acad. Sci. USA* **113**, E6026–E6035 (2016).
62. Liu, J., Li, M., Zhang, Q., Wei, X. & Huang, X. Exploring the molecular basis of heterosis for plant breeding. *J. Integr. Plant Biol.* **62**, 287–298 (2020).
63. Ouyang, Y. & Zhang, Q. The molecular and evolutionary basis of reproductive isolation in plants. *J. Genet. Genomics* **45**, 613–620 (2018).
64. Wang, C. S. et al. Dissecting a heterotic gene through Gradedpool-Seq mapping informs a rice-improvement strategy. *Nat. Commun.* **10**, 2982 (2019).
65. Xie, Y., Shen, R., Chen, L. & Liu, Y. G. Molecular mechanisms of hybrid sterility in rice. *Sci. China Life Sci.* **62**, 737–743 (2019).

Methods

QTG collections and QTN identification. Articles related to QTLs and genes in rice were surveyed by the advanced search function in Web of Science databases published during 1991–2020 with the query TI = (rice OR *Oryza*) AND TS = (QTL OR QTLs OR gene OR allele OR haplotype OR variation OR mapping OR GWAS). The papers used in RiceNavi (version 1) were updated until 29 February 2020. In total, 29,994 articles published in 2,779 journals were obtained. Subsequently, 20,305 articles were selected from the Science Citation Index–Expanded journals that were related to Agriculture and Plant Sciences. We further read each article carefully to identify papers in which QTG functions were validated by transgenic experiments. We excluded QTGs identified from *O. glaberrima* and wild rice unless biotic- and abiotic-stress-related QTGs had been introduced into *O. sativa* cultivars. Finally, a total of 225 QTGs with 603 QTNs (with redundancies) were collected from 299 articles.

QTN identification was based mainly on the descriptions of the collected articles and checked by rice gene annotations. For multiple-allele QTGs, the position of each variant was recorded and linked with the corresponding alleles. The precise locations of QTNs were determined and anchored on the rice MSU v.7.0 reference genome (http://rice.plantbiology.msu.edu/) by BLAST alignment (v.2.7.1). For QTGs that were highly diverged from the Nipponbare reference allele and QTGs with presence and absence variation, coding sequences of the QTGs were downloaded and used as QTNs. The phenotypic effects on agronomic traits of QTG alleles were collected from these articles and recorded following the description in the articles.

QTN library planting and phenotyping. To ensure high-level genetic diversity, we first collected 84 accessions from Chinese minicore collections[66], as well as 15 aus accessions, 27 tropical *japonica* accessions and 5 basmati accessions from South Asia. We further collected diverse accessions with wide geographic distributions and different phylogenetic types. In particular, for rare alleles, the classical disease resistance varieties and the specific varieties used in the published papers were obtained one-by-one from the original researchers. In total, we collected 404 rice accessions from 26 countries worldwide. All samples were planted at the Experimental Station of Shanghai Normal University, Shanghai, China (121°26′ N, 30°58′ E) during May to October in 2018 and 2020. The planting density was 20 × 25 cm^2. Normal agricultural practice and field management was used.

Heading date was recorded when the first inflorescences emerged above the flag leaf sheath. Plant height, panicle number, leaf length and panicle length were calculated from three plants for each accession. Because of the limited quantity of seeds collected, a near-infrared reflectance spectrophotometry (NIRS) scanning approach was used for protein and amylose content detection. Approximately 10 g mature and shelled seed was used for protein and amylose quantification using a FOSS Infratec 1241 Grain Analyser (Foss NIRSystems Inc.) according to the manufacturer's instructions.

Genome sequencing and population genetics analysis. Paired-end sequence data in this study were generated using the Illumina HiSeq4000 platform. Reads of each rice accession were mapped to a reference genome (MSU v.7.0) using Bowtie2 v.2.3.2 (ref. [67]) with default settings. Consecutive steps using Samtools v.1.9 (ref. [68]) and genomic analysis toolkit (GATK) v.3.7 (ref. [69]) were applied for detection of variants. Potential polymerase chain reaction (PCR) duplicates were removed using 'Samtools rmdup'. Alignments around small indels were remapped with 'IndelRealigner', and raw variants were called based on the realigned BAM file. The resulting BAM files of each sample were used for the multisample variant genotyping. 'UnifiedGenotyper' in GATK was applied to generate the raw variant calls with parameters '-stand_call_conf 30, -stand_emit_conf 10'. To reduce the false discovery rate of the variants, the SNP calls were filtered according to the following threshold: QUAL < 30, QD < 2, MQ < 30, MQ0/DP > 0.1. Potential variant annotation and effect were predicted by SnpEff v.3.6 (ref. [70]). For GWAS, genotype imputation was performed by BEAGLE v.4.0 (ref. [71]) using the genotype likelihoods, and with ten iterations specified.

For population genetics analysis, subgroup assignments of the 404 accessions in the QTN library were performed based on the phylogenetic tree, principal component analysis (PCA) and population structure analyses. The phylogenetic tree was constructed using FastTree[72]. PCA was performed by SNPRelate v.0.9.19 (ref. [73]). FastStructure v.1.0 (ref. [74]) was applied to infer the ancestry of each rice accession. Using these approaches, the total 404 accessions were grouped into 88 temperate *japonica*, 77 tropical *japonica*, 8 basmati, 33 aus, 181 *indica* and 17 intermediate type rice.

QTN library genotyping. As there were multiple causative variation types, a hybrid variation genotyping strategy was adopted. For SNP and small indels, GATK4 HaplotypeCaller was used for genotyping in the first round, while some ungenotyped sites were then called by UnifiedGenotyper from GATK3 (ref. [69]). For genotyping structure variation, we used three approaches: Manta v.1.6 (ref. [75]) was first used to detect all potential SVs for each accession, and then the genotypes of the causative sites were extracted. For the QTGs whose sequences were present in the Nipponbare reference genome but absent for some accessions in the QTN library, we then determined the genotype of their presence or absence state by the average mapping depth of the QTG. If the average mapping depth of one accession for the QTG was less than 1×, the accession was determined as the absent genotype, otherwise as the present or Nipponbare genotype. Additionally, in some cases, the QTGs were identified in a few rare rice materials and the sequences were largely different from the Nipponbare reference. The sequences of these QTGs were collected from the literature and combined as another reference sequence (non-Nip-QTGs) for read mapping. The reads that could not be mapped to the Nipponbare genome were extracted, and then mapped to the 'non-Nip-QTGs' reference to determine the genotypes. In addition to the QTNs reported in previous studies, 21 QTNs with loss-of-function mutations in 16 QTG were newly identified in this work. Furthermore, redundant QTNs from the same allele (highly linked with each other) were excluded. In total, 348 QTNs were collected from the literature and this research. The whole pipeline of genotyping causative variants for the QTN library accessions is summarized in Extended Data Fig. 1.

GWAS in eight rice cohorts and QTG effect estimation. The genotype and phenotype data used were generated from this study and also collected from seven other cohorts, namely 529 rice accessions[30], 1,275 Chinese rice accessions[31], Hybrid rice F_1 (ref. [29]), Hybrid rice F_2 (ref. [21]), Rice diversity panel 1 (ref. [28]), 176 Japanese rice accessions[17] and 3K rice accessions[15]. The eight rice cohorts included a total number of 17,376 accessions and 270 trait replicates. The raw sequencing data of three cohorts, including 529 rice accessions, 1,275 Chinese rice accessions and 176 Japanese rice accessions, were downloaded, and the mapping and genotyping pipeline was performed similarly as with our QTN library as described above. Genotypic datasets of the other four cohorts were collected from previous studies[15,21,29,76]. GWAS was performed separately for each cohort by GCTA v.7.93.2 (ref. [77]) with the mixed linear model. The threshold for significant SNPs was 1×10^{-8} for the 'Hybrid rice F_2' cohort and 1×10^{-4} for the '176 Japanese rice accessions' cohort, and a P value threshold of 1×10^{-6} was set for the other six cohorts. The minor allele frequency (MAF) was set as 0.05 for all cohorts, except the 'Hybrid rice F_2' cohort (MAF > 0.02). The genetic effect of each significant SNP was extracted from the summary statistics output. The QTGs that were involved in similar agronomic traits and located around the same regions (<1 Mb) with GWAS peaks selected as causative candidate genes. The QTGs that had been confirmed in the previous GWAS were collected directly from the cohorts. To confirm the newly identified GWAS–QTNs, the linkage disequilibrium pattern of the peak SNP and the QTNs in the cohort was investigated and the QTNs with correlation $P < 0.05$ were used in the genetic estimation. For QTNs of indels and SVs, the correlation coefficient was evaluated in the QTN library. The positive and negative genetic effects were defined by the value of the correlation coefficients. Broad sense heritability (h^2) was calculated by $h^2 = \delta_a^2/(\delta_a^2 + \delta_e^2/l)$, where a and l are number of accessions and locations, respectively; δ_a^2 and δ_e^2 are components of variance for accessions and error, respectively.

Genetic analysis of QTNs. A genetic map constructed from a cross of 9311 and Nipponbare was used for the analysis[41] (Supplementary Dataset 8). The nucleotide-based conservation scores across the whole rice genome were analyzed by CNSpipeline[47]. The rice fitcon scores (ρ) were downloaded from a fitness consequence map[48]. A total of 181 QTGs were used for the examination, in which the 265 QTN sites (SNPs or small indels) in the QTGs could be genotyped by GATK. The conservation scores and ρ of the QTNs were compared with the same type of variants (for example, UTR, nonsynonymous SNP) annotated by SNPeff v.3.6 (ref. [70]). In addition, we also used SIFT[49] to measure the conservation level for nonsynonymous SNP sites for the rice QTGs. To investigate the extent to which QTNs reside in upstream regions of the translational start site and overlap the candidate open chromatin regions, we downloaded rice ATAC-seq data from previous studies[78,79] and FAIRE-seq data from another research project[80] (Supplementary Table 3). The ATAC-seq and FAIRE-seq data were processed with Bowtie2 for reads mapping, and Samtools rmdup was used for removing PCR duplicates. In addition, reads that could map to the mitochondrion or chloroplast genomes were filtered. Finally, open chromatin regions were determined by MACS2 (ref. [81]).

To identify the candidate QTNs responsible for heterosis in the three-line hybrid rice system, genome resequencing data from 23 restorer lines and 40 male sterile lines were used in previous research[29]. QTN sites were genotyped for all individuals of the two populations, and the alternative allele frequency differentiation (AFD) between the two populations was calculated. QTNs with AFD > 0.4 ($P < 0.05$, chi-squared test) were taken as candidate QTNs related to heterosis. QTGs that had multiple QTNs were checked and nonconforming QTGs were removed.

In our QTN library accessions, we classified *indica* rice into East, South and Southeast Asian groups, which included 59, 61 and 52 accessions, respectively. For *japonica*, the accession number for Southeast, Northeast and East Asian groups is 32, 44 and 30, respectively. The alternative allele frequency of each QTN was calculated for each different geographical population for comparison. The alternative allele frequencies of QTGs that have multiple QTNs were combined with the alternative allele frequency of several QTNs. The alternative allele frequencies of QTGs were further transferred to functional allele frequency according to the direction of genetic effect for the QTNs. Functional alleles of all QTGs are shown in Supplementary Datasets 4 and 5.

To investigate QTNs involved in rice domestication and improvement, genomic data from a total of 1,645 Asian rice samples were used. The samples included 169 wild rice, 134 *indica* landraces, 137 *japonica* landraces, 423 *indica* varieties and 775 *japonica* varieties (Supplementary Dataset 9). The allele frequency of each QTN was calculated for each group, and AFD for the processes of domestication ($AFD_{landrace} - AFD_{wild}$) and improvement ($AFD_{variety} - AFD_{landrace}$) was measured for *indica* and *japonica*, respectively. To minimize the influence of genetic drift during domestication or improvement, we used 4DTv sites as background. The allele frequency changes of each 4DTv site were also calculated, and the 2.5th and 97.5th percentiles of 4DTv allele frequency changes were used as thresholds.

RiceNavi development and benchmarking experiments. The RiceNavi software includes three packages, namely RiceNavi-QTNpick, -Sim and -SampleSelect (Fig. 6). The RiceNavi-QTNpick package can take a gvcf file of the rice sample generated from GATK4 as input or it may pick one accession in the QTN library as the receptor line, and call the genotype of the sample at the causative sites. The genotype of the sample is compared with those of the QTN samples. This further provides allele information of the user's sample and the samples harboring the alterative allele for each QTN, including the function of the alleles and potential allelic effects. The user can pick the beneficial QTN(s). After choosing the QTN(s), the donor sample list will be provided.

The Rice-SampleSelect package can select suitable genotypes to facilitate the breeding process. The input file for this package is a genotyping matrix, where each column represents samples, while each row is the binned genotype (for example, 0.3 Mb per bin). The genotyping matrix is generated by our constructed genotyping pipeline SEG-map for skim genome sequencing[82]. The Rice-SampleSelect package can output the summarized genotype characteristics for each individual of that population, such as the number of recombination breakpoints, heterozygosity across the whole genome, the number of heterozygous genomic blocks, the size of the heterozygous regions covering the targeted genes, etc. Samples with heterozygous genotypes on target genes are further ranked according to the whole genome heterozygosity level for ease of selection.

The RiceNavi-Sim package is implemented taking advantage of the PedigreeSim software[83]. The PedigreeSim software can simulate the genotype of the offspring if the genotypes of the parents and the genetic map are given. With the constructed rice genetic map, the genotype matrix of different generations (F_1 to BC_nF_1) for a breeding population can be simulated by RiceNavi-Sim. During each generation, RiceNavi-Sim adopted Rice-SampleSelect to select the best candidates as parental lines for the next generation. The simulation time can be set by the user. After all simulations are performed, the likelihood can be estimated. In each generation, the likelihood was calculated based on the percentage of simulations that have the 'ideal' individuals with only heterozygous genotypes in the regions covering selected gene(s) with 2 Mb as the default size.

Genetic improvement of HHZ using RiceNavi. The QTN library was planted on 25 May 2017 at Shanghai, and six groups of HHZ were grown in the next 6 weeks to facilitate cross pollination. HHZ was crossed successfully with Basmati in September 2017 and 19 hybrid seeds were obtained. F_1 individuals were backcrossed with HHZ in March 2018 at Lingshui, China and 1,080 hybrid seeds were obtained. In total, 461 BC_1F_1 individuals were resequenced and 138 individuals with ≤ 50% heterozygosity were crossed successfully with HHZ. Furthermore, 908 BC_2F_1 individuals were resequenced and ten individuals containing desired alleles of *OsSOC1*, *Badh2* and *TAC1* from Basmati were selected. A total of 1,190 BC_3F_1 individuals were resequenced and three individuals with the lowest heterozygosity were selected. Eight thousand BC_3F_2 individuals planted on 20 November 2019 at Sanya were screened by PCR-based markers. Of 23 individuals resequenced, 21 were confirmed to have homozygous target QTGs while no other chromosome segments were found. To validate the causative variation in *OsSOC1*, *Badh2* and *TAC1* in the improved HHZ, DNA of both HHZ and improved HHZ lines was extracted from fresh leaves using a Hi-DNAsecure Plant Kit (Tiangen, China) and PCR amplifications were performed with 0.5 U Tks Gflex DNA Polymerase (TaKaRa, Japan) using a ProFlex PCR System (Applied Biosystems, USA). The cycling conditions were 94 °C for 1 min followed by 35 cycles of 98 °C (10 s), 60 °C (15 s) and 68 °C (30 s). PCR products were electrophoresed in 1.5% agarose gels, and DNA fragments were purified using an AxyPrep DNA Gel Extraction Kit (Axygen, Germany). Sequencing reactions were performed using an ABI 3730XL automated sequencer (Applied Biosystems). The primers used for PCR amplification and Sanger sequencing were the same and are listed in Supplementary Table 4.

The improved HHZ was phenotyped for ten traits in Sanya in the winter of 2019 and Shanghai in the summer of 2020. Tiller angles of HHZ and the improved HHZ were measured at the beginning of heading date, and this was repeated three times by investigating three tillers. Grain length and grain width were measured by manual measuring of ten seeds. To test the fragrance of improved HHZ, the harvested grains were dried in an oven at 37 °C for 2 days and were evaluated by four experienced rice breeders. A double-blind test was performed, in which each sample was smelled and replicated two times. The fragrance level was recorded as 1 (non-basmati fragrance) and 2 (basmati-specific fragrance), respectively.

Statistical tests used. Details of the statistics applied are provided in the figure legends. Specifically, the two-tailed Student's *t*-test was performed to compare the phenotypic differences (heading date and fragrance level) between HHZ and the improved line. One-sided Wilcoxon test was used to compare the conservation score and ρ score of QTNs and variants of the QTGs. Statistical analyses were performed using R software (v.3.6.0, https://www.r-project.org/). GWAS in this work was performed with the mixed linear model using the GCTA software v.7.93.2 (ref. [77]).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw DNA sequencing data of the QTN library are deposited with GenBank under the bioproject accession no. PRJNA623686. A web-based version of RiceNavi is available from the website http://www.xhhuanglab.cn/tool/RiceNavi.html (supporting most browsers including Chrome, Firefox and Safari, but not Internet Explorer). In this web-based application, all functions in RiceNavi (QTNmap, QTNpick, Simulation and SampleSelect) can be accessed with user-friendly graphical interfaces.

Code availability

The source code of RiceNavi is available from both our laboratory website (http://www.xhhuanglab.cn/tool/RiceNavi.html) and the GitHub repository (https://github.com/xhhuanglab/RiceNavi). The other codes for the QTN-related analyses are also provided in the GitHub repository (https://github.com/xhhuanglab/QTN_scripts).

References

66. Wei, X. et al. Domestication and geographic origin of *Oryza sativa* in China: insights from multilocus analysis of nucleotide variation of *O. sativa* and *O. rufipogon*. *Mol. Ecol.* **21**, 5073–5087 (2012).
67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
68. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
69. McKenna, A. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
70. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
71. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
72. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
73. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
74. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
75. Chen, X. Y. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
76. Wang, D. R. et al. An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* **9**, 3519 (2018).
77. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
78. Wilkins, O. et al. EGRINs (environmental gene regulatory influence networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *Plant Cell* **28**, 2365–2384 (2016).
79. Reynoso, M. A. et al. Evolutionary flexibility in flooding response circuitry in angiosperms. *Science* **365**, 1291–1295 (2019).
80. Zhao, L. et al. Integrative analysis of reference epigenomes in 20 rice varieties. *Nat. Commun.* **11**, 2658 (2020).
81. Zhang, Y. et al. Model-based analysis of ChIP–Seq (MACS). *Genome Biol.* **9**, R137 (2008).
82. Zhao, Q., Huang, X. H., Lin, Z. X. & Han, B. SEG-Map: a novel software for genotype calling and genetic map construction from next-generation sequencing. *Rice* **3**, 98–102 (2010).
83. Voorrips, R. E. & Maliepaard, C. A. The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinform.* **13**, 248 (2012).

Acknowledgements

We are grateful to the China National Rice Research Institute, Institute of Crop Sciences of Chinese Academy of Agricultural Sciences, Institute of Plant Protection of Chinese Academy of Agricultural Sciences, Chinese Academy of Sciences Center for Excellence of Molecular Plant Sciences and Huazhong Agricultural University for providing valuable rice varieties (see Supplementary Dataset 2 for details). We thank P. Xu and J. Murray for their advice and assistance in the paper writing. This work was funded by the National Natural Science Foundation of China (grant nos. 91935301 and 31825015), Innovation Program of Shanghai Municipal Education Commission (grant no. 2017-01-07-00-02-E00039) and Program of Shanghai Academic Research Leader (grant no. 18XD1402900) to X.H. and the US National Science Foundation (Plant Genome Research Program, IOS-1947609) to K.M.O.

Author contributions

X.H. designed these studies and contributed to the original concept of the project. X.W., K.Y., J.F., Q.Z. and H.H. contributed to the collection, planting and phenotyping of the QTN library. Q.W. and J.L. performed the genome sequencing of the QTN library and breeding populations. J.Q., X.W. and X.H. performed QTN analysis, developed the RiceNavi system and implemented RiceNavi in practical breeding. X.H., J.Q., X.W., K.M.O. and B.H. analyzed the data and wrote the paper.

Competing interests

A patent on the QTN-based breeding selection method has been filed by Shanghai Normal University with X.H., X.W. and J.Q. as inventors. The remaining authors declare no competing interests.

Additional information

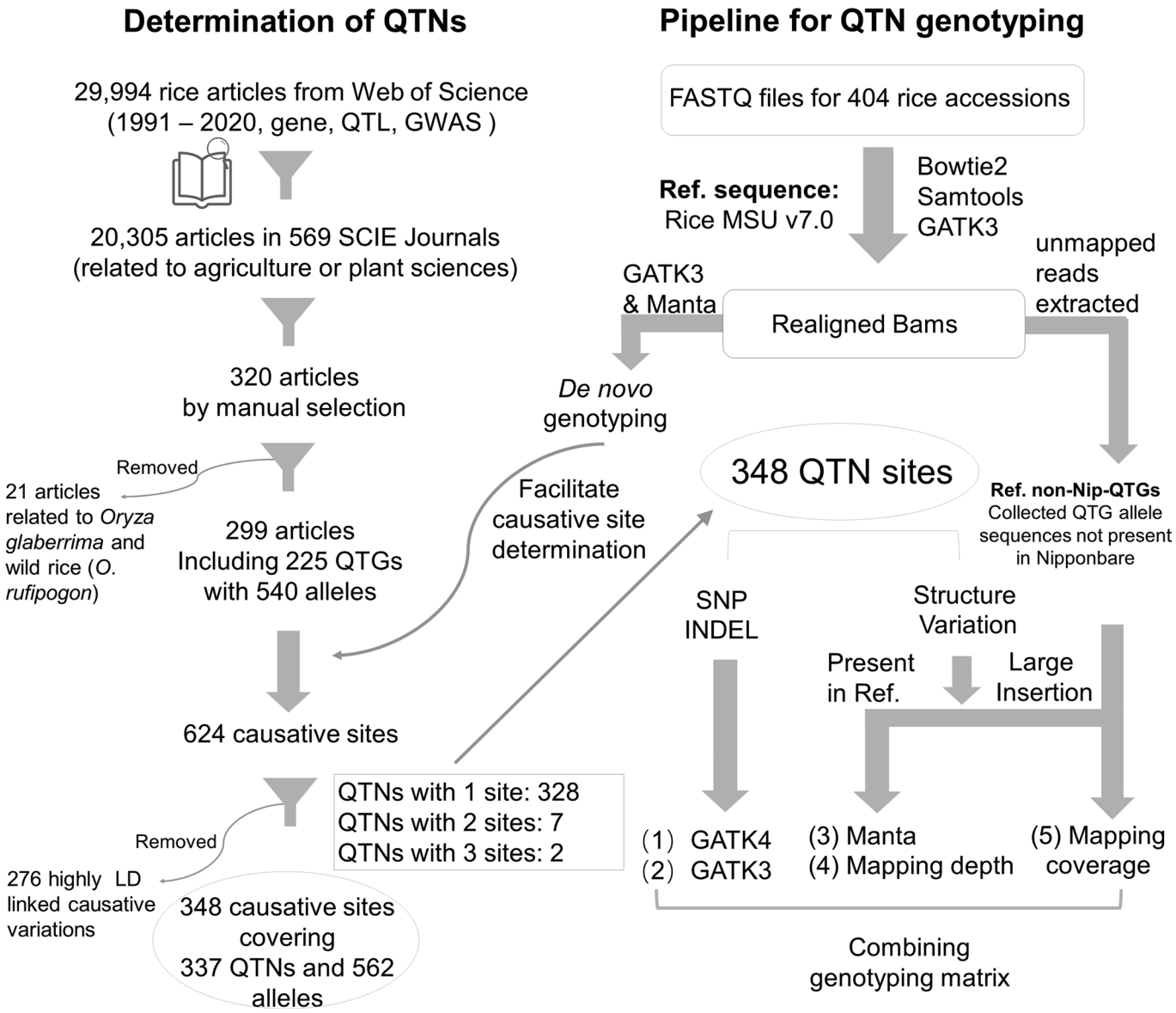
Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00769-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-020-00769-9.

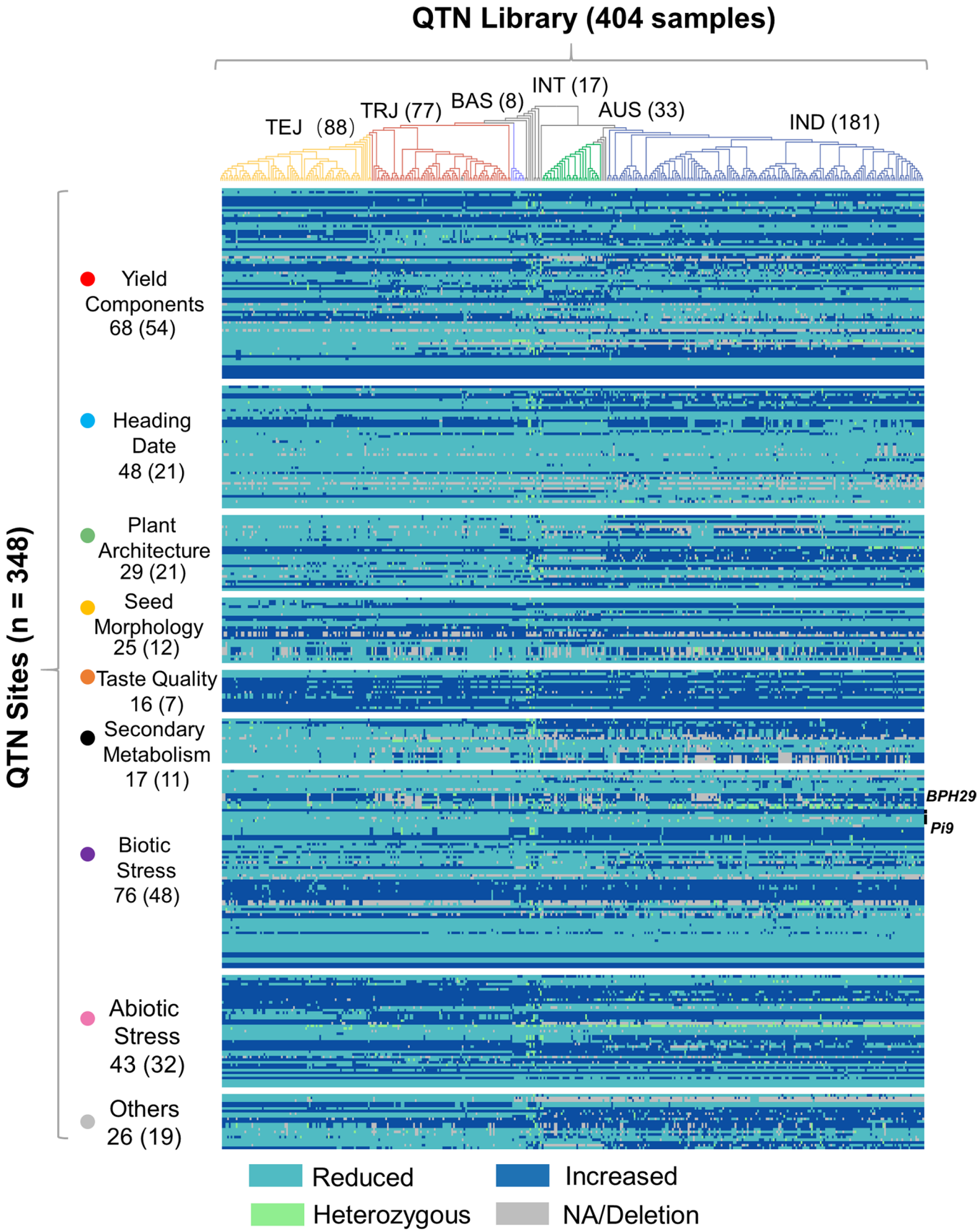
Correspondence and requests for materials should be addressed to X.H.

Peer review information *Nature Genetics* thanks Makoto Matsuoka and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

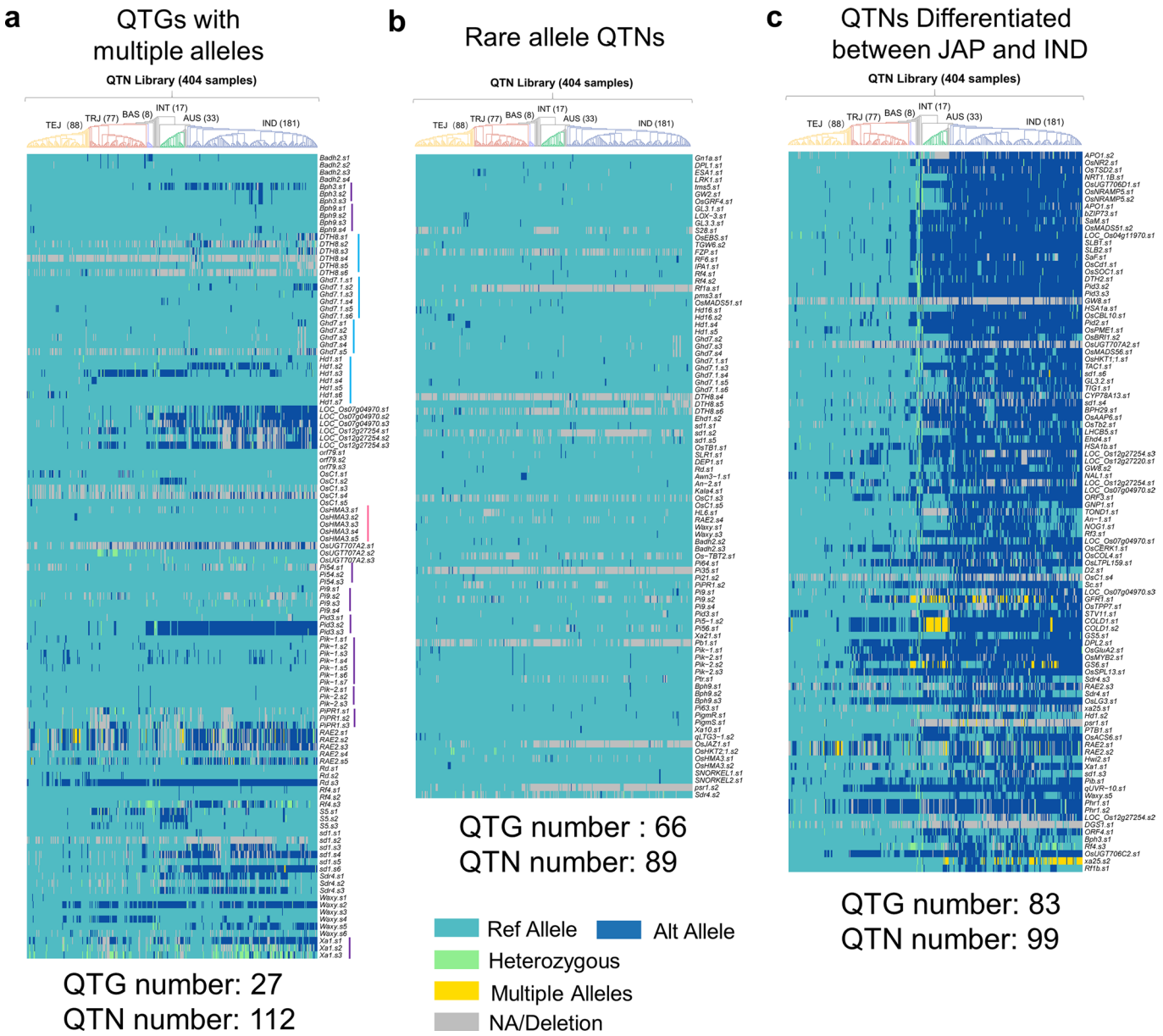
Reprints and permissions information is available at www.nature.com/reprints.



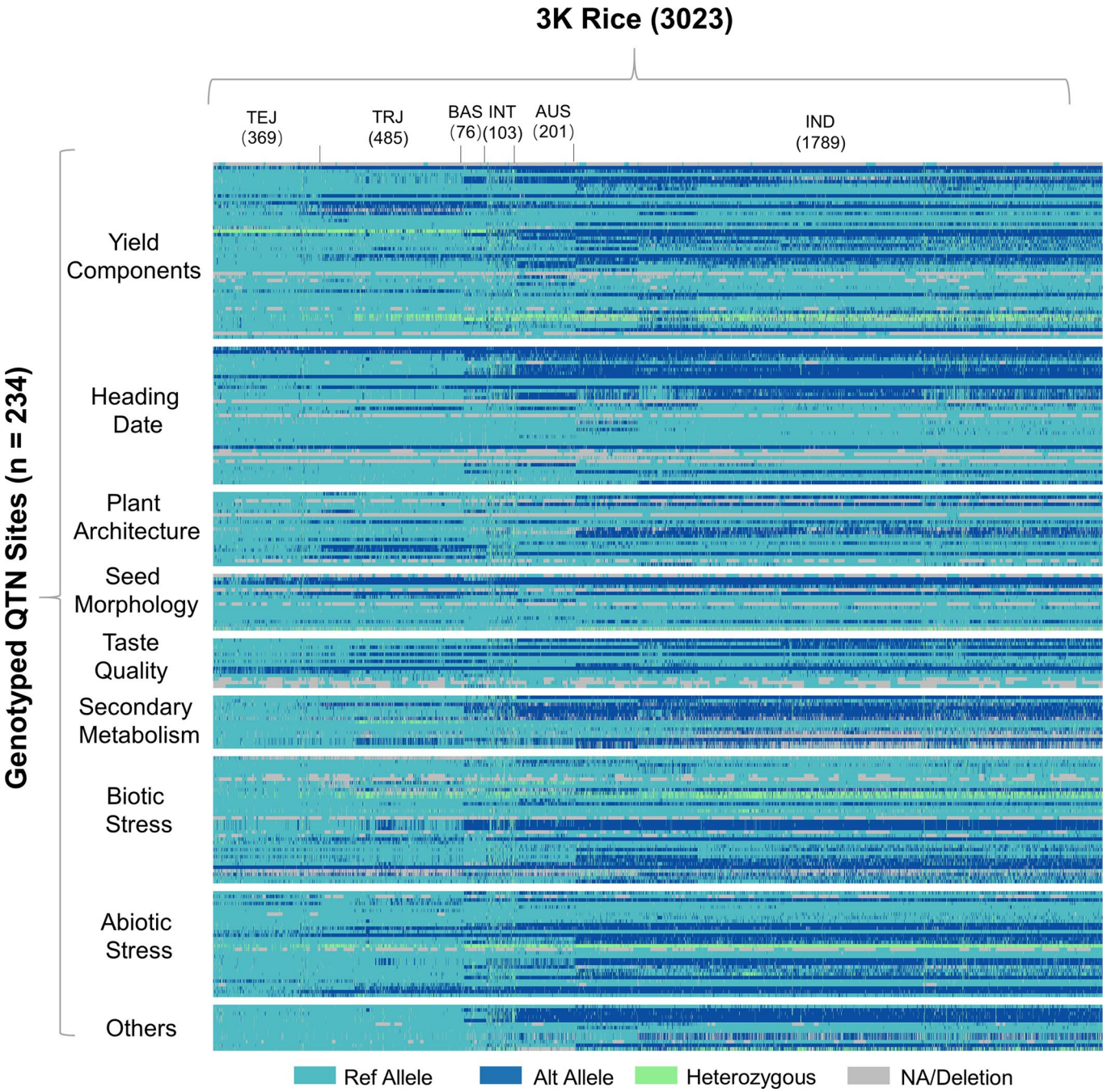
Extended Data Fig. 1 | The pipeline for 348 QTN site discovery and population genotyping. The procedure includes determination of QTNs according to research papers and QTN genotyping from whole-genome sequence data of rice accessions.



Extended Data Fig. 2 | Genotype matrix of 225 QTGs for QTN library colored by effect direction. The figure is another display mode of Fig. 1a (colored by effect direction, rather than alternative/reference in Fig. 1a). Here, dark green, dark blue and light green, yellow and gray boxes represent the genotype for the reduced allele, increased allele, heterozygous, NA and deletion, respectively.



Extended Data Fig. 3 | The matrix of QTN of different types for a collection of 404 rice accessions. a, QTGs with multiple (≥3) QTNs. QTNs related to heading date, biotic stress and abiotic stress are highlighted with blue, purple and pink bars, respectively. **b**, Rare allele QTNs. QTNs with low percentage of samples (≤2%) with alternative or heterozygous alleles are illustrated. **c**, QTNs differentiated between *japonica* (including tropical and temperate *japonica*) and *indica*. QTNs with allele frequency differentiated (AF > 0.4) between *japonica* and *indica* are shown. Light blue, dark blue and light green, yellow and gray boxes represent the genotype for the reference (MSUv7.0), alternative, heterozygous, multiple alleles and deletion, respectively.



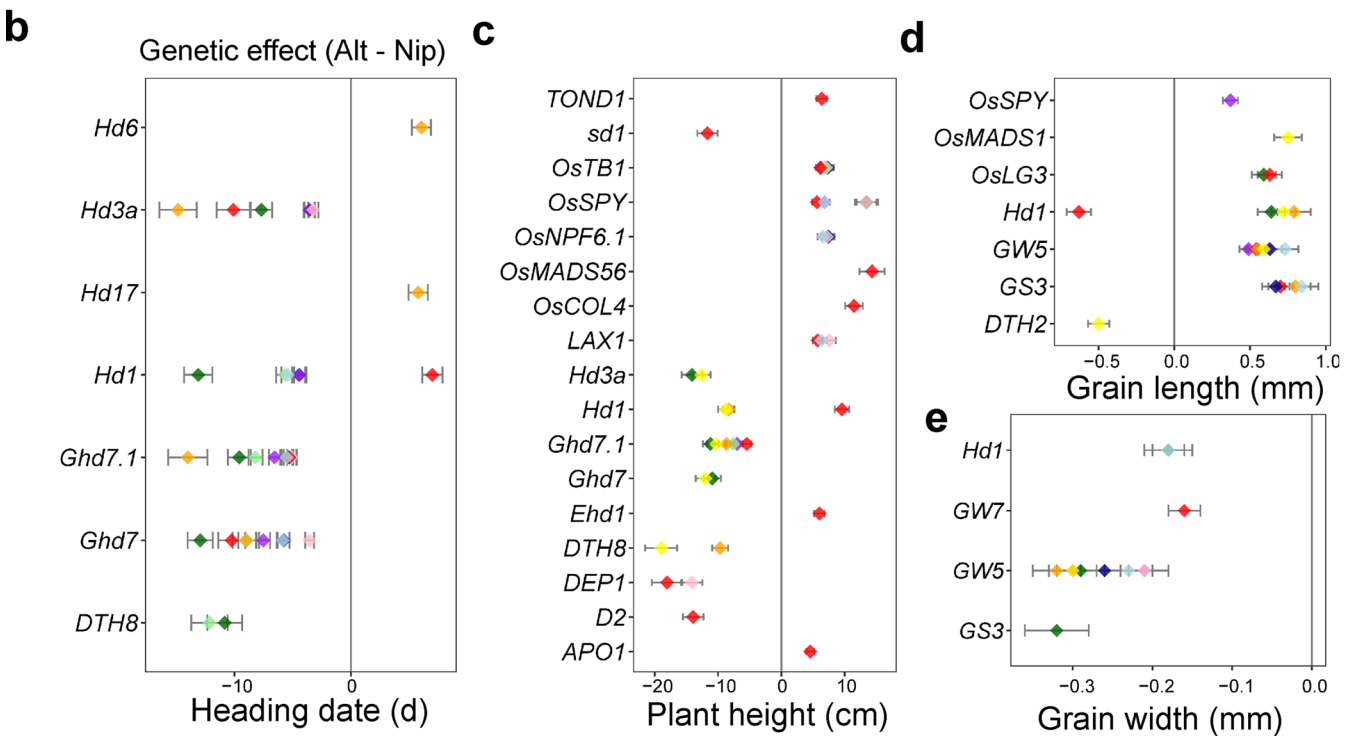
Extended Data Fig. 4 | The matrix of QTN for 3023 rice accessions. Light blue, dark blue and light green, yellow and gray boxes represent the genotype for the reference (MSUv7.0), alternative, heterozygous, and deletion, respectively.

a

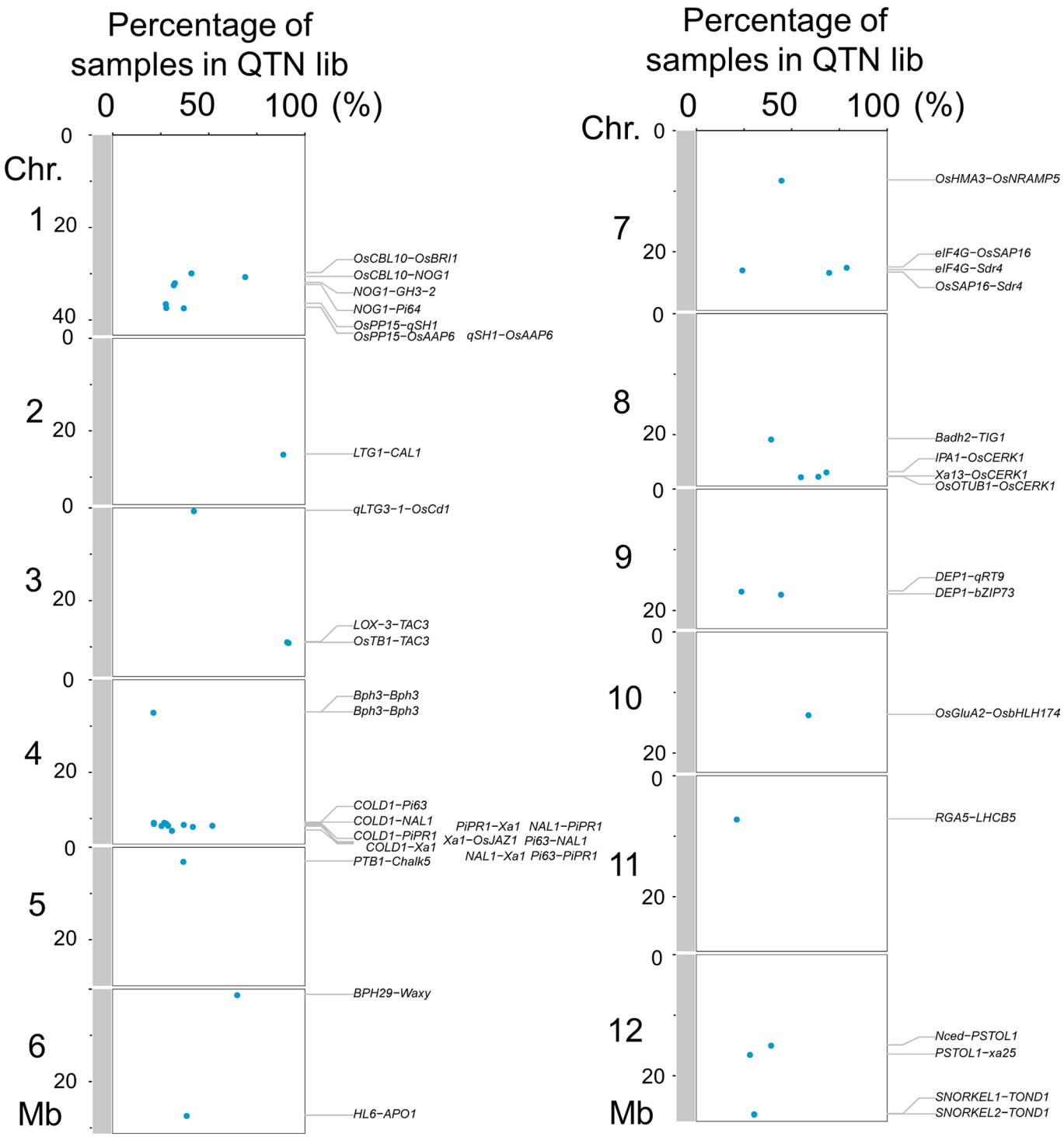
Location	Longitude	Latitude
Heihe	127.50	50.25
Jiamusi	130.40	46.82
Haerbin	126.96	45.55
Wuchang	127.17	44.93
Jilin	124.49	42.50
Beijing	116.23	40.22
Yangzhou	119.43	32.39
Wenjiang	103.86	30.68
Linshui	110.04	18.51

North

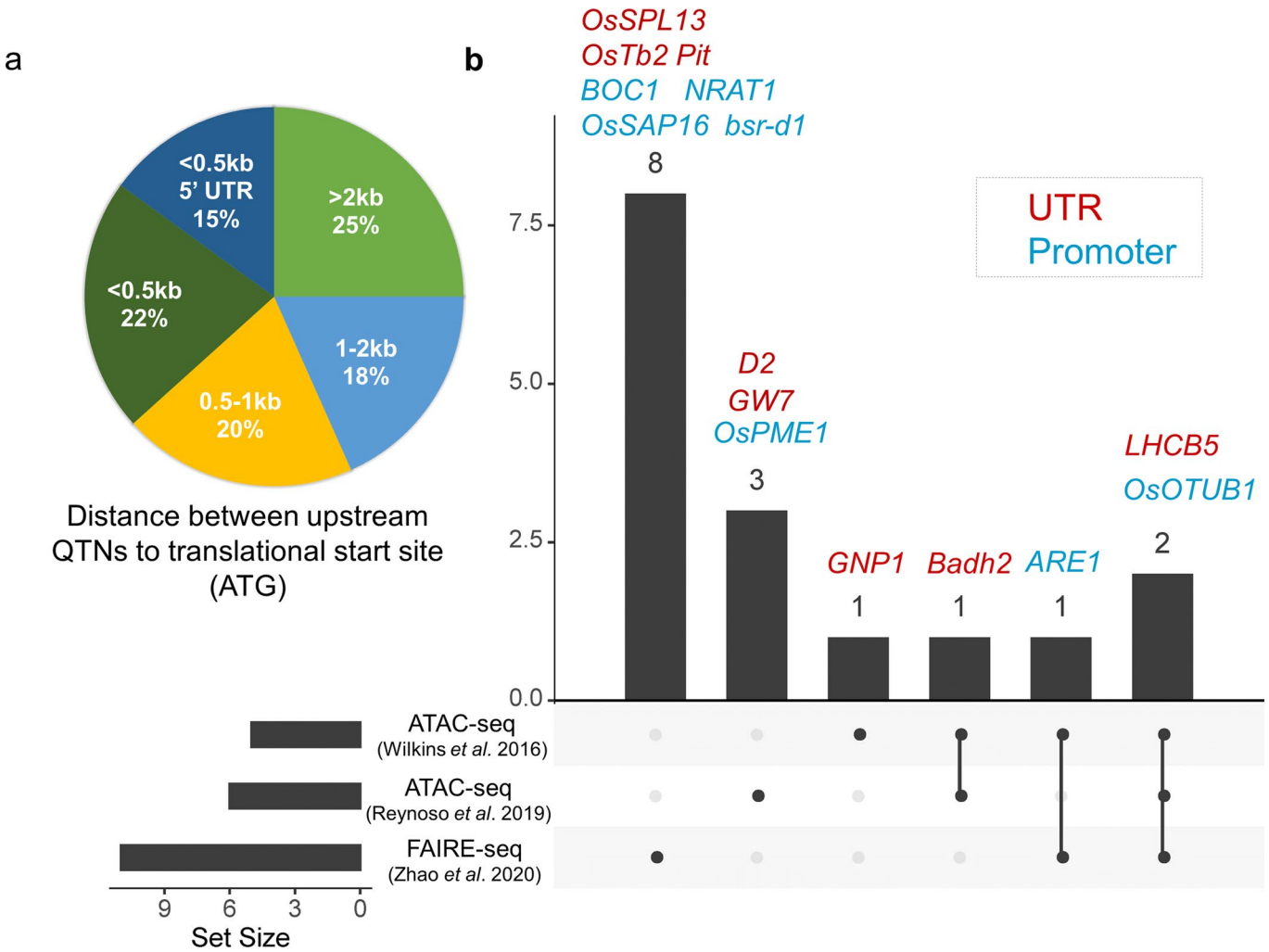
South



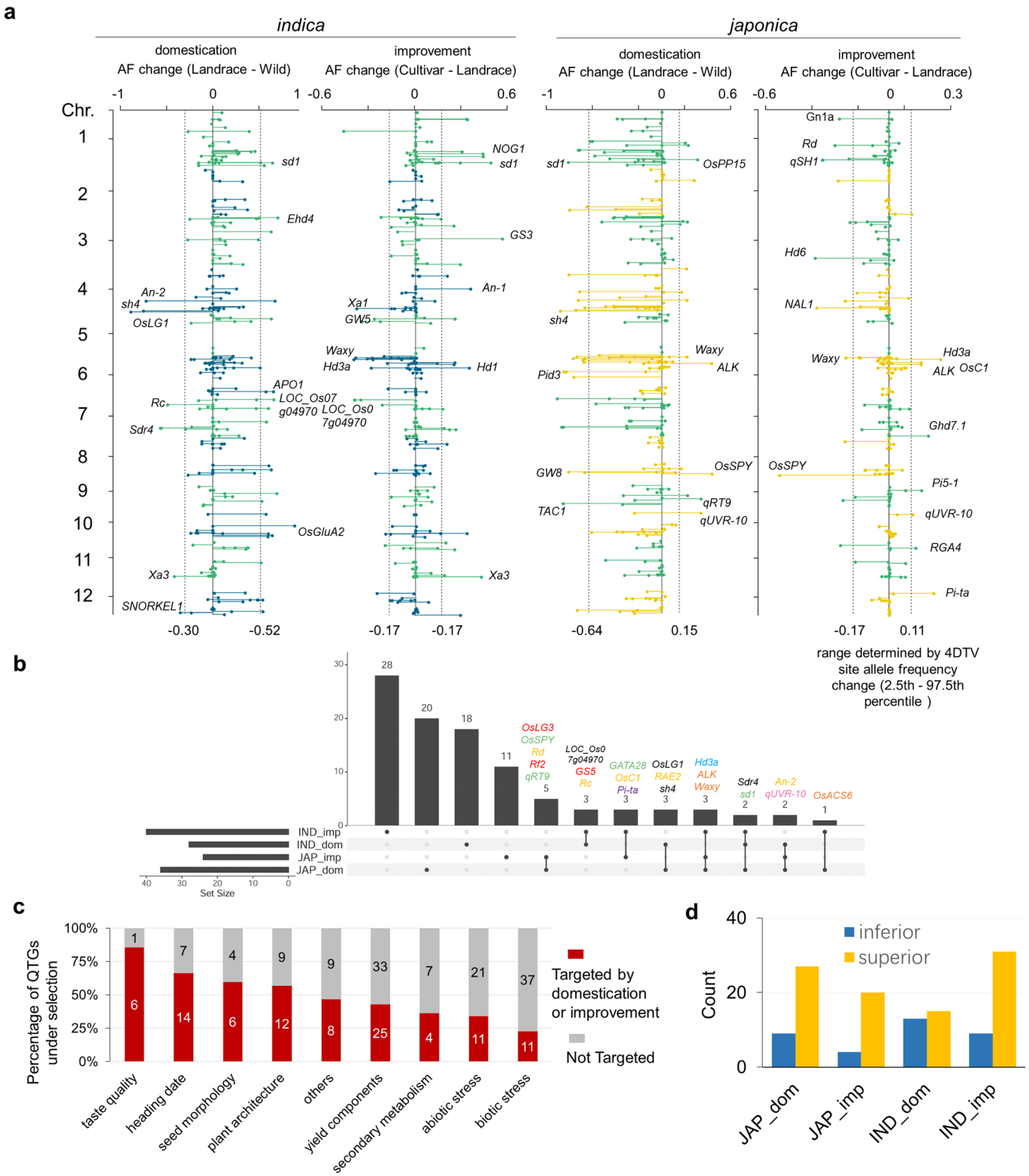
Extended Data Fig. 5 | Estimated phenotypic effects for QTGs controlling four agronomic traits. a, Geographical locations for 9 different environments in China. Longitude (° E) and latitude (° N) of the locations are shown. **b–d**, The estimated phenotypic effects of homozygous alternative alleles relative to homozygous Nipponbare are jointly shown for each QTG. The phenotypes displayed include heading date (**b**), plant height (**c**), grain length (**d**) and grain width (**e**). Colors represent different environments. The bars indicate standard errors estimated by GCTA package. The QTG effects from CNmix population in Beijing and from NE population in Lingshui are not showed. For QTNs in Lingshui, the QTNs with the peak *p*-value are selected.



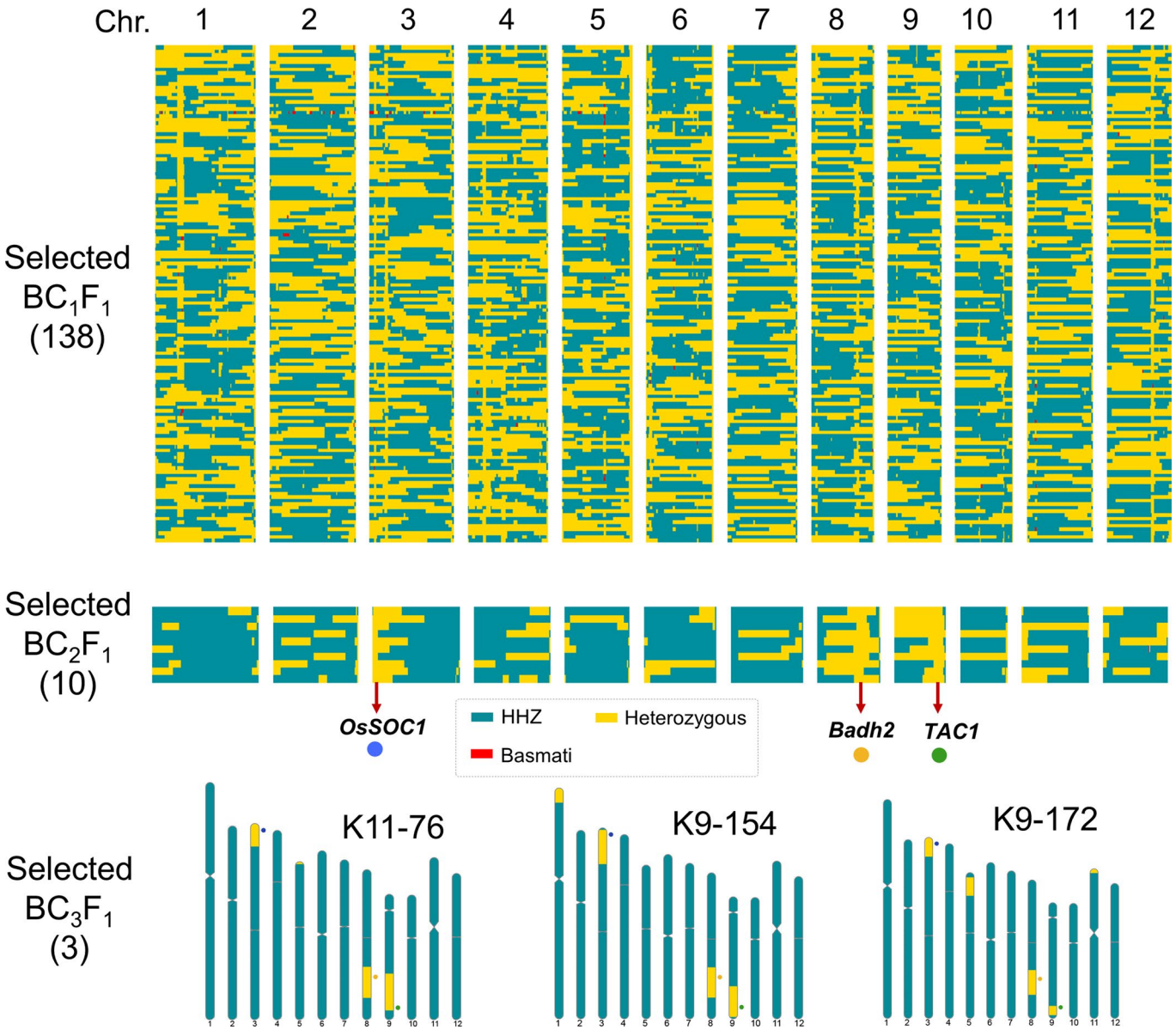
Extended Data Fig. 6 | Genomic distribution of linkage drag in the rice genome for QTN library. The candidate linkage drag (superior and inferior alleles located physically less than 2 Mb in distance) are labeled across the rice genome. The blue dots indicate the percentage of drag for the 404 QTN library accessions.



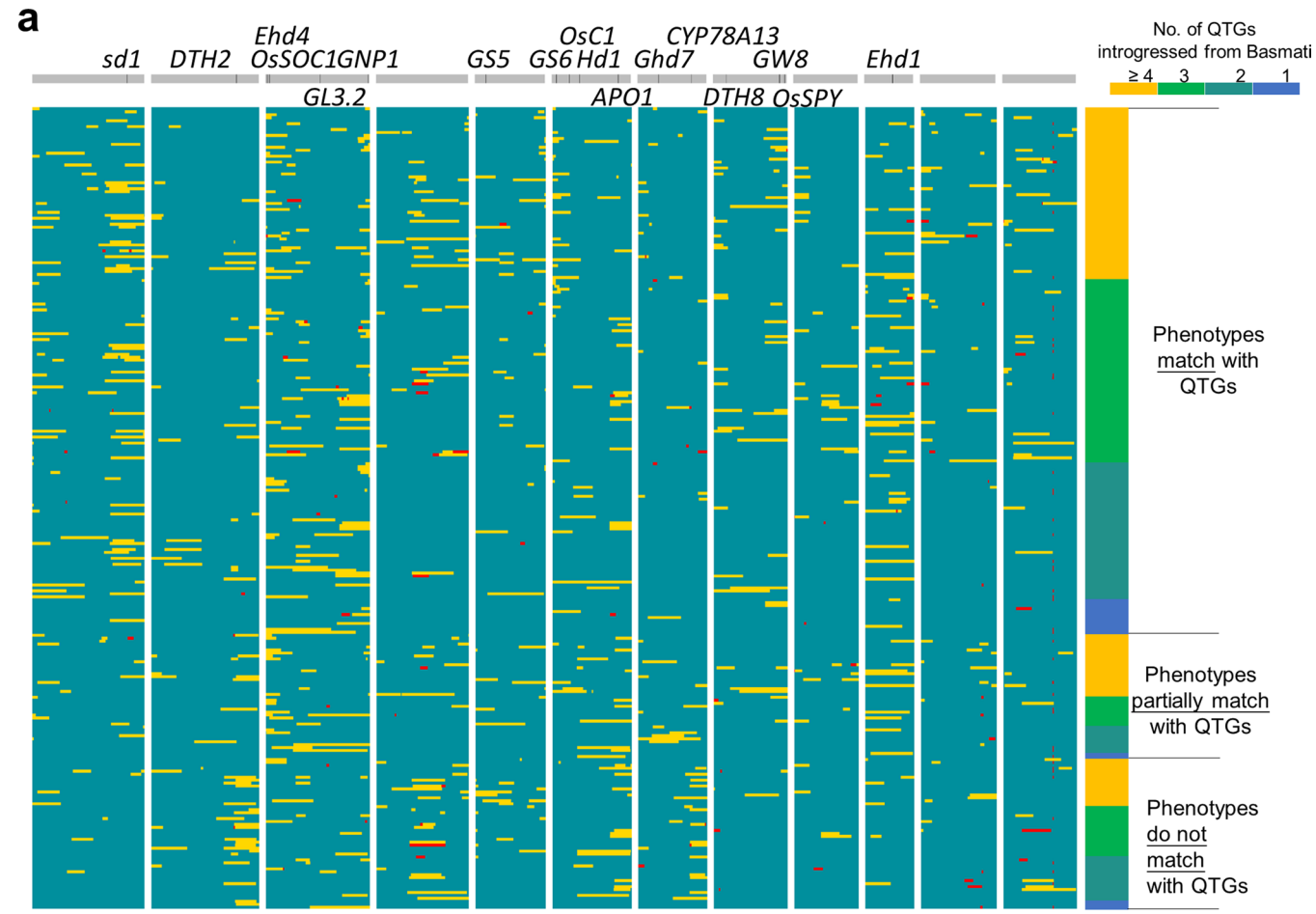
Extended Data Fig. 7 | Genomic characteristics for the QTNs in the UTR and promoter regions. a, Percentage of upstream QTNs of different distances to translational start site (ATG). **b**, Upstream QTN sites which resides in the open chromatin regions identified by ATAC- and FAIRE-seq.

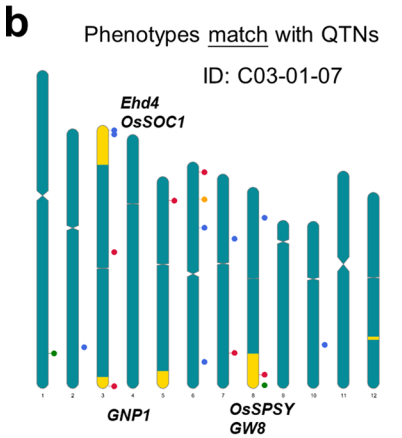


Extended Data Fig. 8 | The QTNs involved in the domestication and improvement. a, QTNs allele frequency change during the domestication & early variety improvement and modern variety improvement. QTNs with greatest allele changes are shown. Threshold is determined by the 4DTv sites and is indicated by dotted line. **b**, Groups of the domestication and improvement-related QTNs. QTNs shared by two kinds of domestication or improvement are shown. The color of the QTN names represents traits and is in line with Fig. 1a. **c**, Percentage of domesticated and improved QTGs in different agronomic traits. **d**, Number of QTGs with superior and inferior alleles.

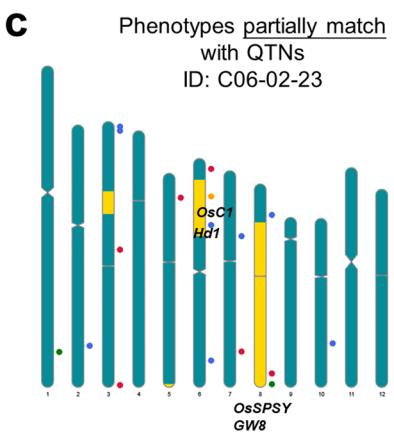


Extended Data Fig. 9 | The genotypes for the selected individuals of each generation during improvement of HHZ. The superior alleles of three QTGs (*OsSOC1*, *Badh2* and *TAC1*) are targeted during the breeding process for improvement of HHZ. The locations of the three QTGs are indicated by the red arrows. From BC_1F_1 to BC_3F_1, the numbers of selected individuals are 138, 10 and 3, respectively. The genotypes for the HHZ background, donor Basmati, and heterozygous are color coded as dark green, red, and yellow respectively.

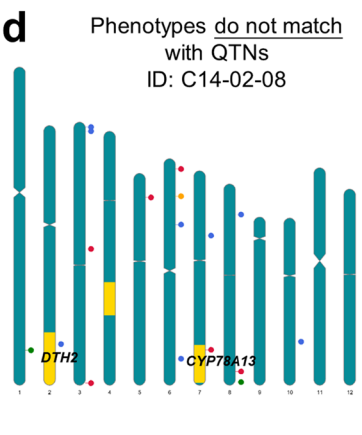




	Introgressed QTG(s) from Basmati	Expected	Real Traits
Plant Height	*OsSPY* & *GNP1*	↑	↑
Heading date	*Ehd4* & *OsSOC1*	↓	↓
Seed Length	-	-	-
Seed Width	*GW8*	↓	↓
Apiculus color	-	-	-



	Introgressed QTG(s) from Basmati	Expected	Real Traits
Plant Height	*OsSPY*	↑	↓
Heading date	*Hd1*	↑	↑
Seed Length	-	↓	↑
Seed Width	*GW8*	↓	↓
Apiculus color	*OsC1*	turn purple	turn purple



	Introgressed QTG(s) from Basmati	Expected	Real Traits
Plant Height	-	-	-
Heading date	*DTH2*	↓	↑
Seed Length	*CYP78A13*	↓	↑
Seed Width	*CYP78A13*	↓	↑
Apiculus color	-	-	-

Extended Data Fig. 10 | An examination for the extent to which introgressed segments from donor parents could match expected phenotypes. a, The genotypes of the 217 BC_3F_1 CSSLs that constructed by HHZ and Basmati. The genotypes for the HHZ background, donor Basmati, and heterozygous are color coded as dark green, red, and yellow respectively. Positon of the introduced QTNs is shown on the top. Number of QTNs that introduced into HHZ is shown on the right. **b**–**d**, Genotypes of three individuals of the CSSLs. QTNs are indicated by solid circles. The color represents the group of agronomic traits and is line with Fig. 1a. The change direction of the phenotype value is indicated by arrows. Red and blow arrows indicate increase and decrease of the traits, respectively.

nature research

Corresponding author(s): Xuehui Huang

Last updated by author(s): Nov 24, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data.
Data analysis	"Online Method" section: 1. Reads of each rice accession were mapped to reference genome (MSUv7.0) using BOWTIE2 v2.3.2 with default settings. Consecutive steps using Samtools v1.9 and GATK v3.7 were applied for detection of variants. 2. The phylogenetic tree was constructed with using Fasttree. PCA was performed by SNPRelate v0.9.19. FastStructure v1.0 was applied to infer the ancestry of each rice accession. 3. GWAS was performed separately for each cohort by GCTA v7.93.2 with the mixed linear model. 4. The conservation scores and fitcon scores (p) of the QTNs were compared to the same type of variants (e,g., UTR, nonsynonymous SNP) annotated by SNPeff v3.6. In addition, we also used SIFT to measure the conservation level for nonsynonymous SNP sites for the rice QTGs. 5. The precise locations of QTNs were determined and anchored on the rice MSU v7.0 reference genome by BLAST (v2.7.1) alignment. 6. Genetic improvement of HHZ was performed by using RiceNavi (v1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

"Data availability statement" section:
1. The raw DNA sequencing data of the QTN library are deposited in the GenBank under the bioproject accession number PRJNA623686.
2. A web-based version of RiceNavi is available in the website http://www.xhhuanglab.cn/tool/RiceNavi.html (supporting most browsers including Chrome, Firefox, and Safari, but not IE). In this web-based application, all functions in RiceNavi (QTNmap, QTNpick, Simulation and SampleSelect) can be accessed with user-friendly graphical interfaces.

The software package we developed is included in "Code availability statement" section:
3. The source code of of RiceNavi is available in both our lab web (http://www.xhhuanglab.cn/tool/RiceNavi.html) and the GitHub repository (https://github.com/xhhuanglab/RiceNavi.git).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	"GWAS in eight rice cohorts and QTG effect estimation" of "Online Methods" section: The genotype and phenotype data used was generated from this study and also collected from seven other cohorts. The eight rice cohorts included a total number of 17,376 accessions and 270 trait replicates.
Data exclusions	No samples were excluded in the analysis.
Replication	For the QTN library, phenotyping was conducted on three replicates for each line.
Randomization	Randomization is not involved in this study, because we used all the rice QTN information in public literatures, and we collected the genotypic and phenotypic data of all the rice accessions in the eight GWAS cohorts .
Blinding	Online Methods section: A double-blind test was performed, in which each sample was smelled and replicated two times. The fragrance level was recorded as 1 (non-basmati fragrance) and 2 (basmati-specific fragrance), respectively.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging