RESEARCH

Open Access

Genetic Diversity and Breeding Signatures for Regional *Indica* Rice Improvement in Guangdong of Southern China

Yu Hang[1,2,3,4], Liu Yue[1,2,3,4], Sun Bingrui[1,2,3,4], Liu Qing[1,2,3,4], Mao Xingxue[1,2,3,4], Jiang Liqun[1,2,3,4], Lyu Shuwei[1,2,3,4], Zhang Jing[1,2,3,4], Chen Pingli[1,2,3,4], Pan Dajian[1,2,3,4], Chen Wenfeng[1,2,3,4], Fan Zhilan[1,2,3,4] and Li Chen[1,2,3,4*]

Abstract

As the pioneer of the Green Revolution in China, Guangdong province witnessed the improvement and spread of semi-dwarf *Xian/Indica* rice cultivars and possessed diverse rice germplasm of landrace and cultivars. A total of 517 accessions containing a core germplasm of 479 newly sequenced landraces and modern cultivars were used to reveal breeding signatures and key variations for regional genetic improvement of *indica* rice from Guangdong. Four subpopulations were identified in the collection, which including Ind IV as a novel subpopulation that not covered by previously released accessions. Modern cultivars of subpopulation Ind II were inferred to have less deleterious variations, especially in yield related genes. About 15 Mb genomic segments were identified as potential breeding signatures by cross-population likelihood method (XP-CLR) of modern cultivars and landraces. The selected regions spanning multiple yield related QTLs (quantitative trait locus) which identified by GWAS (genome-wide association studies) of the same population, and specific variations that fixed in modern cultivars of Ind II were characterized. This study highlights genetic differences between traditional landraces and modern cultivars, which revealed the potential molecular basis of regional genetic improvement for Guangdong *indica* rice from southern China.

Keywords Rice, Yield improvement, Resequencing, Breeding signature, GWAS

Background

Rice (*Oryza sativa*) feeds more than half of the world's population, and rice yield is vital for world food security. Rice genetic improvement in China has facilitated the increase of its production over the past several decades. Guangdong of southern China witnessed the breeding, introduction and spread of semi-dwarf *indica* rice accessions. Since then, rice yield has been increased by about three-fold with the breakthrough of high-yield cultivars. However, with the spread of modern cultivars, rice landraces that were grown by local farmers is gradually disappearing. For the goal of further production increasing, the usage of genetic diversity for the valuable germplasm needs to be enhanced in breeding programs. Tremendous efforts have been made by germplasm scientist for the collection and conservation of landraces and locally-improved traditional cultivars of southern China. These landraces and cultivars represent the rice genetic diversity of southern China before and after the rice "Green Revolution", which could be used to reveal genetic trajectory for regional *indica* rice breeding and phenotype enhancement. Revelation of the functional variations related to the success of rice breeding in southern

*Correspondence:
Li Chen
lichen@gdaas.cn
[1] Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China
[2] Key Laboratory of Genetics and Breeding of High Quality Rice in Southern China (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Guangzhou 510640, China
[3] Guangdong Key Laboratory of New Technology in Rice Breeding, Guangzhou 510640, China
[4] Guangdong Rice Engineering Laboratory, Guangzhou 510640, China

China will promote the utilization of genetic resources for future rice breeding. Moreover, characterization of the genome sequences, genetic diversity and functional variations of these germplasm collections is becoming very critical for the next potential breakthrough of rice production.

The advancement of sequencing technologies enabled the analysis of genetic diversity for large collection of germplasm, which promoted the revelation of domesticated loci, and accelerated the identification of functional genes. Genotypes of a large collection of 517 rice landraces were identified with onefold-coverage sequencing and accurate imputation method, population structure, genome-wide association analysis and haplotype analysis were conducted using about 3.6 million nonredundant SNPs (Huang et al. 2010). Thereafter, higher-depth sequencing with more than 15-fold coverage were conducted on 40 cultivated and 10 wild rice accessions to identify selection signatures during domestication using nucleotide polymorphisms (Xu et al. 2012), and larger collection of 446 wild diverse rice accessions and 1083 cultivated varieties were also genotyped by sequencing and used for the identification of 55 selective sweeps during domestication (Huang et al. 2012). Further, 10,074 F_2 lines from 17 representative hybrid rice combinations were genotyped, and heterosis related loci were identified (Huang et al. 2016). Release of sequencing data from "3000 Rice Genomes Project" (3KRGP) largely facilitated the identification of untapped variations and novel genes (Fuentes et al. 2019; Wang et al. 2018). Jointly data analysis of 3KRGP and Indian long and short grain germplasm identified the long low-diversity region harboring key gene regulating grain weight (Kumar et al. 2020). Recently, the sequencing of local germplasm and improved varieties illustrated the regional genetic diversity in detail. Sequencing of 239 *japonica* rice elites from China, Japan and Korea identified 1131 novel genes and artificial selection signals (Liu et al. 2021). Analysis of 672 Vietnamese rice genomes described their classification and identified 21 unique QTLs from 19 traits (Higgins et al. 2021). Genotyping and systemically phenotyping of 200 *japonica* rice varieties grown in central China over the past 30 years revealed the genetic factors regulating the balance of yield, quality and blast resistance (Xiao et al. 2021).

Artificial selection and breeding signatures during the succession of rice varieties makes deep insight into their genetic improvement. Low-coverage sequencing of 1479 landrace and modern cultivars from 73 countries revealed 200 regions were differentially selected between two major *indica* subpopulations, and yield was correlated with number of the signatures (Xie et al. 2015). Locally selective sweeps also showed pressure during artificial selection and farmer cultivation. Genotyping by sequencing of 108 core on-farm conserved rice landraces from Yunnan revealed 186 and 183 potential selective-sweep between different collection date (Cui et al. 2019). Different selection signature during breeding indicated by genetic differentiation between early and late cultivars of *indica* and *japonica* in Taiwan (Hour et al. 2020). As the pioneer of "Green Revolution" for *indica* rice in China, Guangdong province have rich diversity of germplasms. However, large-scale population genomics of rice landraces and improved varieties for the study of genetic diversity and identification of regional breeding signatures of are still lacking.

In this study, core germplasm of locally planted rice accessions of landrace and cultivars from Guangdong of southern China before and after rice "Green Revolution" were collected. Agronomic traits were systematically investigated by field experiments, and they were genotyped by 10-fold depth genome resequencing. Genetic diversity was analyzed and breeding signatures for modern cultivar subpopulation were identified and annotated by QTLs of eleven agronomic traits. Specific genetic variations and favorable alleles that fixed in subpopulation of modern cultivar were identified that can be used for molecular marker assisted breeding.

Results

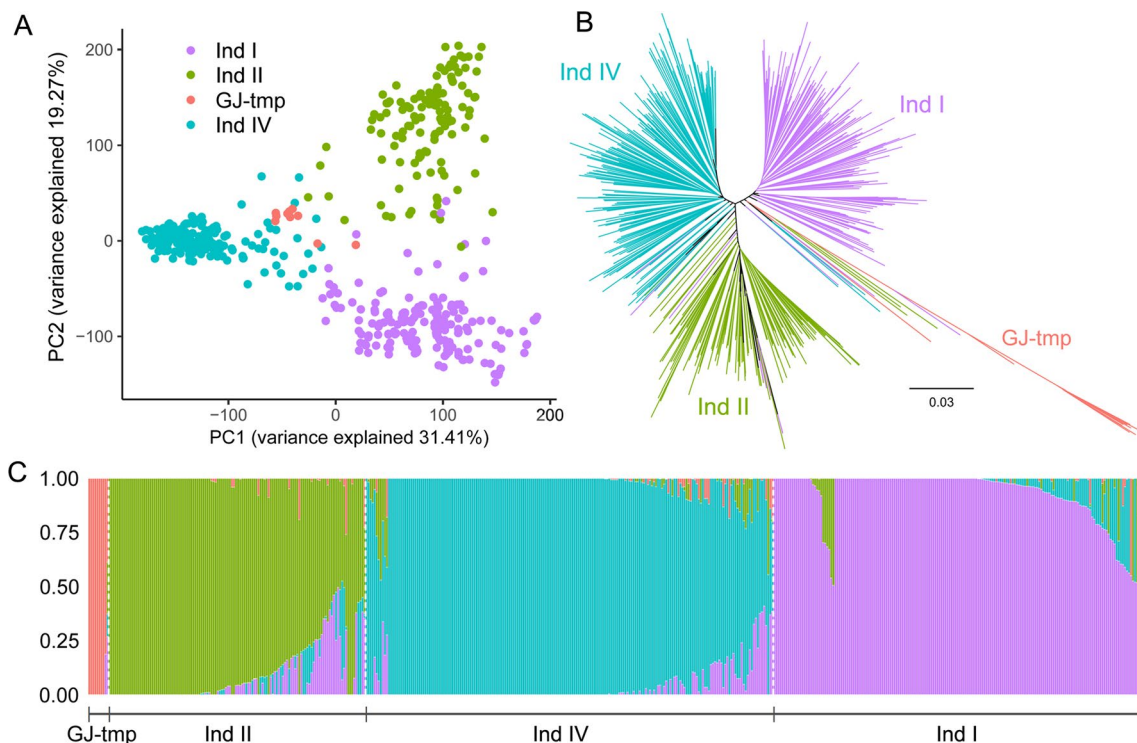### Genetic Diversity and Population Structure Analysis

A total of 517 accessions consisting mainly of *indica* rice germplasm, which including 358 landrace and 159 artificially improved cultivars from Guangdong, China were used to identify genomic variations (Additional file 1: Table S1). The 479 newly-sequenced accessions generated 24.23 million 100 bp pair-end sequencing reads for each accession. Quality assessment of these sequencing reads revealed the average Q30 base quality (99.9% base call accuracy) percent was 99.65% (Additional file 1: Table S2). The average mapping depth against the MSU7 reference genome was 12.02 with 91.68% coverage ratio (Additional file 1: Table S3). Averagely 2.04 million SNPs, 82.34 thousand insertions and 127.99 thousand deletions were identified for 517 *indica* rice accessions (Additional file 1: Table S4).

Population structure analysis was conducted by principal components analysis (PCA), phylogenetic and admixture analysis. In admixture analysis, group 1 contained 211 accessions (201 landraces and 10 cultivars), and group 2 contained 306 accessions (157 landraces and 149 cultivars) when subpopulation number (*k*) was 2. And three groups, namely group 1 (20 landraces and 116 cultivars), group 2 (189 landraces and 11 cultivars) and group 3 (149 landraces and 32 cultivars) can be identified when *k*=3 (Additional file 2: Fig. S1). Integrated with

PCA (Fig. 1A), phylogenetic tree (Fig. 1B) and admixture analysis (Fig. 1C), a total of 4 subpopulations were finally determined. With the reference of 182 accessions (with three subpopulation that named with Ind I, Ind II and Ind III) from RiceVarMap database, and the phylogenetic relationship with accessions from 3KRG, these 4 subpopulations were named as Ind I (landrace), Ind II (cultivar), Ind IV (landrace) and GJ-tmp in this study (Fig. 2A, Additional file 2: Fig. S2). GJ-tmp diverged from other subpopulations, and most of accessions from GJ-tmp subpopulation were glutinous rice landraces. Another subpopulation Ind I contained a total of 181 accessions, 151 accessions of which were landraces and 30 cultivars that were bred before 1980s. Subpopulation Ind II contained 126 accessions, which have 117 cultivars and 9 landraces. Subpopulation Ind IV have 189 landraces and 11 cultivars that were bred before 1980s (Additional file 1: Table S1). A recently released and refined *indica* reference genome (9311) was also used to call genetic variations and conduct population structure analysis, which obtained similar results for genetic clustering for all those accessions (Additional file 2: Fig. S3).

The length of linkage disequilibrium (LD) decay for subpopulation Ind I, Ind II and Ind IV other than the glutinous rice subpopulation GJ-tmp were estimated using the square of the correlation coefficient (r^2) between variations. LD decay distance for Ind IV, Ind I and Ind II were 61.0 kb, 110.1 kb and 219.8 kb, respectively. The extension of LD decay distance for Ind II indicated that the cultivar subpopulation Ind II underwent artificial selection pressure during the process of genetic improvement. Interestingly, the landrace subpopulation Ind I probably have selection effect by regional farmer breeders as its LD decay distance longer than landrace subpopulation Ind IV (Fig. 2B). Genetic diversity (*pi* and *Tajima's* D) and differentiation (*fst*) analysis were conducted for three main subpopulations of Guangdong *indica* rice. The *pi* values for Ind IV, Ind I and Ind II were 0.0031, 0.0029 and 0.0029, respectively. Ind IV have higher *pi* values, while Ind I and Ind II have similar values. The *Tajima's* D value for Ind IV was positive, while they were negative for Ind I and Ind II, which implying potential selection effect in subpopulation of Ind I and Ind II. Genetic divergence (*fst*) between Ind IV and Ind I is smaller than that of Ind IV and Ind II, which indicates Ind II was higher diverged from Ind IV than Ind I (Fig. 2C). Phylogenetic tree with 998 common wild rice (*Oryza rufipogon* Griff.) lines also indicates degree of differentiation from wild rice populations from high to low was Ind IV, Ind I and Ind II (Fig. 2D). Together with these results, we deduced

Fig. 1 Population structure analysis of Guangdong *indica* rice accessions. PCA plot (**A**), phylogenetic analyses (**B**) and population structure (**C**) showing genetic diversity and clustering of all accessions based on whole genome SNP variations. For PCA plot, PC1 (principal component 1) and PC2 (principal component 2) are showed on horizontal and vertical axes, and percentages of variance explained were noted in parentheses
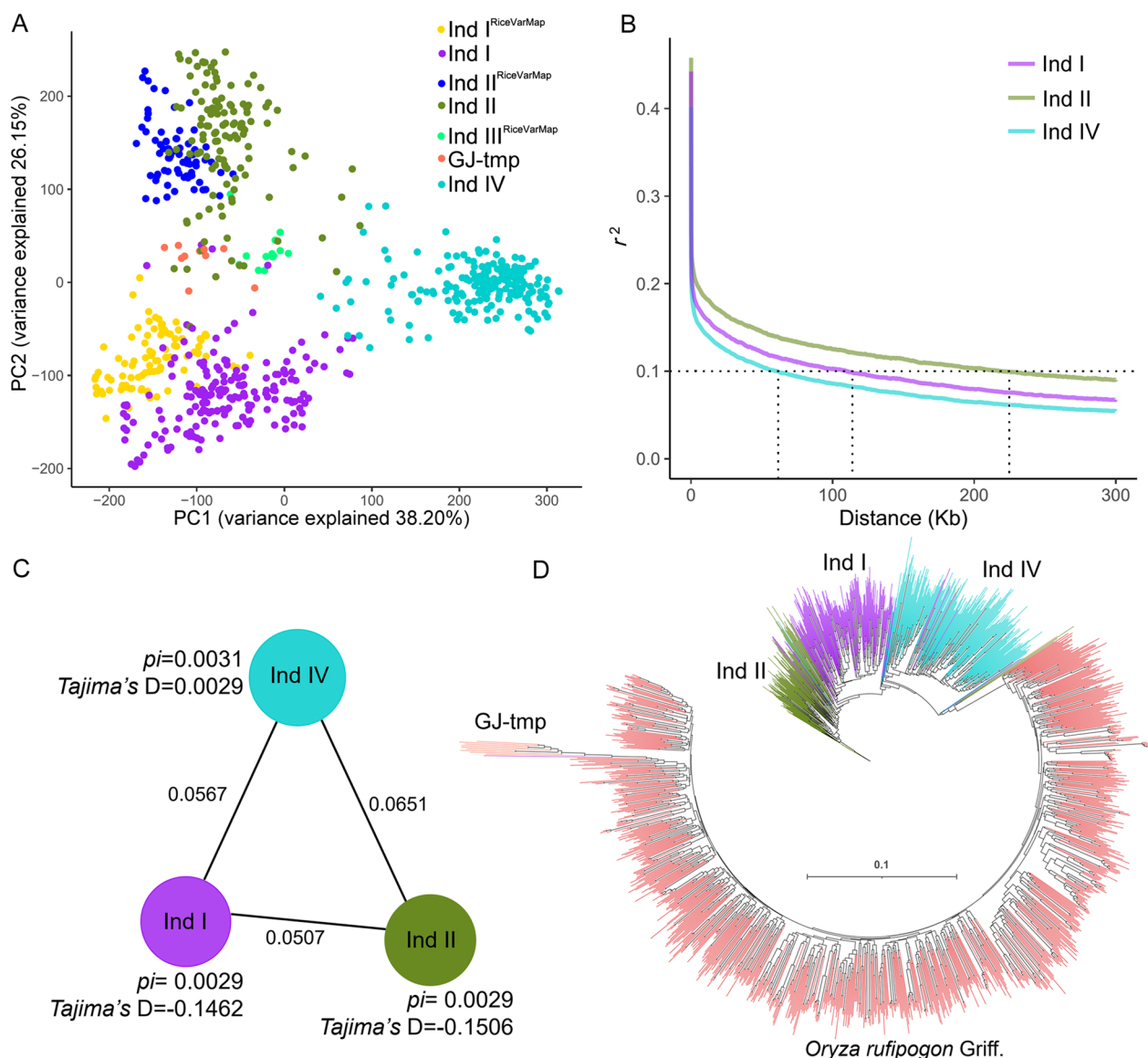
Fig. 2 Genetic diversity and phylogenetic relationship of subpopulations for Guangdong *indica* rice. **A** PCA plot depicting the comparison of genetic clusters of 479 Guangdong and 220 *indica* accessions from RiceVarMap database. **B** Linkage disequilibrium (LD) decay analysis of three main subpopulations. **C** Genetic diversity and differentiation of three subpopulations for Guangdong *indica* rice. The size of the circles represents the level of genetic diversity (*pi*) of the subpopulations, and, and length of lines represent *fst* values between subpopulations. **D** Phylogenetic tree for Guangdong *indica* rice and common wild rice (*Oryza rufipogon* Griff.)

that the genetic differences of these regionally cultivated rice lines were attribute to the cultivation period, as modern cultivars may have high speed and flexible distance in their seed dispersal.

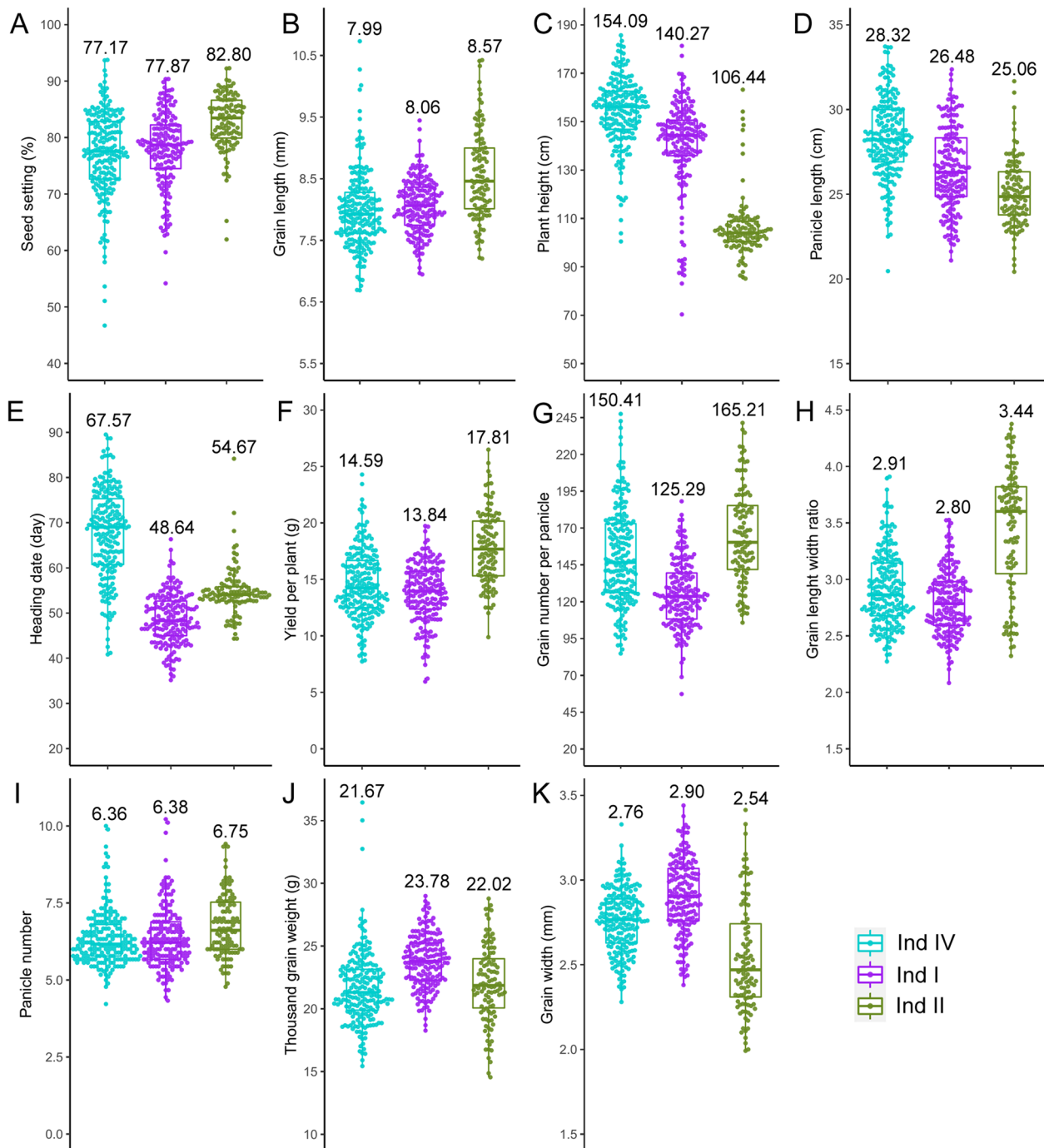**Phenotypic Comparison for Subpopulations**

The selection pressure by local breeders during rice improvement for the past half century largely changed the agronomic traits between traditional and modern

rice accessions. Genetic diversity and LD decay analysis indicates potential selection pressure in Ind I, and even stronger selection effect in modern cultivar subpopulation Ind II. The alteration of agronomic traits for these subpopulations recorded the trajectory of these selection effect. A total of eleven important agronomic including plant height (PH), heading date (HD), yield per plant (YPP), panicle number (PN), grain number per panicle (GNPP), seed setting (SS), thousand grain weight (TGW), panicle length (PL), grain length (GL),

grain width (GW) and grain length width ratio (GLWR) were investigated and analyzed.

During the improvement progress of Ind IV, Ind I and Ind II, values of eleven agronomic traits showed four different types of changing trends. Seed setting rate (Fig. 3a) and grain length (Fig. 3b) were increased during modern breeding process, as shown by the comparison of Ind IV, Ind I and Ind II. Plant height (Fig. 3c) and panicle length (Fig. 3d) descended during this process, which represents the main phenotype alteration for semi-dwarf rice

Fig. 3 Phenotype comparison of eleven main agronomic traits for three main subpopulations of Guangdong *indica* rice. Subpopulations were ordered by LD decay distance from short to long, and average trait values for each subpopulation were noted above boxplots

cultivars that released during rice "green revolution" of southern China. Trait of heading date (Fig. 3e), yield per plant (Fig. 3f), grain number per panicle (Fig. 3g) and grain length width ratio (Fig. 3h) showed fluctuation of decline in Ind I and elevation in Ind II. Panicle number (Fig. 3i), thousand grain weight (Fig. 3j) and grain width (Fig. 3k) were raised in Ind I but decreased in Ind II. The increasing of thousand grain weight and grain length and grain width reflecting the selection of high yield rice lines with large grain size, while the breeding and application of high-quality "Simiao rice" with small and slender grains, the Guangdong *indica* rice showed decrease of these traits and the increase of grain length width ratio.

Frequency of Deleterious or Beneficial Allele During Genetic Improvement

Number of deleterious variations that encode adverse amino acid were predicted in three main subpopulations, and the number of accessions from landrace subpopulation Ind I and Ind IV were compared cultivar subpopulation Ind II. Firstly, deleterious variations identified by SIFT software showed the total count of deleterious variations were stepwise decreasing in Ind I (median number was 3255.0) and Ind II (median number was 3287.5) compared with Ind IV (median number was 3472.0), which implying these variations were lost during modern cultivars improvement under artificial selection pressure (Fig. 4A and Additional file 1: Table S5). Secondly, a total of 319 quantitative trait nucleotides (QTNs) of the 212 vital gene in rice of RiceNavi database were used to annotate accessions of three subpopulations (Additional file 1: Table S6). For all genes, the average inferior allele count of accessions in Ind IV, Ind I and Ind II were 52.19, 52.96 and 50.32, respectively. Inferior allele counts were 17.38, 16.34 and 15.72 for yield related genes and 29.83, 30.91 and 30.88 for Ind IV, Ind I and Ind II, respectively. These results suggesting that modern cultivars of subpopulation Ind II accumulated favorable alleles, especially for yield related genes during improvement (Fig. 4B and C). However, modern cultivars lost some favorable allele of stress responsive genes (Fig. 4D).

QTLs and Breeding Signatures of Guangdong Indica Rice

Breeding signatures of modern cultivar subpopulation Ind II were identified using significant distorted patterns in allele frequency of XP-CLR method. Modern cultivar subpopulation Ind II was respectively compared with landrace subpopulation Ind IV and Ind I. For the comparison of Ind IV and Ind II, a total of 150 genomic segments spanning 15.10 Mb potentially selected genomic regions by modern cultivar breeding were identified (Additional file 1: Table S7). And for the comparison of Ind I and Ind

II, a total of 146 genomic segments with genomic length of 14.59 Mb were identified (Additional file 1: Table S8).

Genome wide association study (GWAS) were conducted for eleven yield and yield-related traits and the effects of candidate genes were identified. For instance, *Ghd7.1/DTH7* explained 9.50% of heading date variances, *sd1* explained 15.07% of plant height variances, *GS3* explained 4.57%, 9.84% and 6.46% phenotype variances for thousand grain weight, grain length and grain length width ratio, *GSE5* explained 11.40%, 22.85% and 11.49% phenotype variances for thousand grain weight, grain width and grain length width ratio, and GS5 explained 40.70% and 4.37% phenotype variances for grain width and grain length width ratio (Additional file 1: Table S9, Additional file 2: Fig. S4). Effect of allele combination were analyzed for plant height and grain size genes. Average plant height of accessions with combination of *sd1*[Hap1] and *Oshox4*[Hap2] was 116.49 cm, which significantly lower than 151.70 cm of *sd1*[Hap1] and *Oshox4*[Hap1] (Additional file 1: Table S10). A total of 25 major allele combinations of grain size genes *GS3*, *GSE5* and *GS5* were detected. Thousand grain weight ranged from 19.56 to 24.22 g, grain length ranged from 7.62 to 9.43 mm, grain width ranged from 2.30 to 2.96 mm, and grain length width ratio ranged from 2.66 to 3.99 for the accessions with the 25 allele combinations. For instance, the high quality "Simiao" rice Meixiangzhan2hao have the combination of *GS3*[Hap3], *GSE5*[Hap2] and *GS5*[Hap6] with thousand grain weight of 20.88 g and grain length width ratio of 3.99 (Additional file 1: Table S11). The phenotype effects of known genes that genotyped by RiceNavi were also evaluated, and several QTNs have potential effects on Guangdong indica rice phenotype. For instance, two variations of *Hd1* gene (9338004 and 9338220 on Chr6) shows effect to promoting heading date by about 9 days (Additional file 1: Table S12). Four genes that regulating eating quality were also genotyped. Ten accessions were genotyped to have fragrance allele of *Badh2* gene, and 9 of which are Ind II accessions. The only different site of two elite cultivars Huanghuazhan and Meixiangzhan-2hao was the presence and absence of the fragrance allele of *Badh2* gene (Additional file 1: Table S13).

XP-CLR analysis were conducted by comparing Ind II with Ind IV (Fig. 5A) and Ind I (Fig. 5B), and the QTLs were further used to annotate the regions of breeding signatures (Fig. 5C). A total of 24 and 23 intersections between agronomic QTLs and selected genomic regions for subpopulation Ind IV and Ind I, respectively. Known vital yield and yield-related genes under selection pressure were detected. For instance, *sd1*, *Oshox4* and *OsGA2ox5* were found under selection of plant height, *OsWDR5*, and *TAC3* were selected for the improvement of grain yield per plant, and favorable
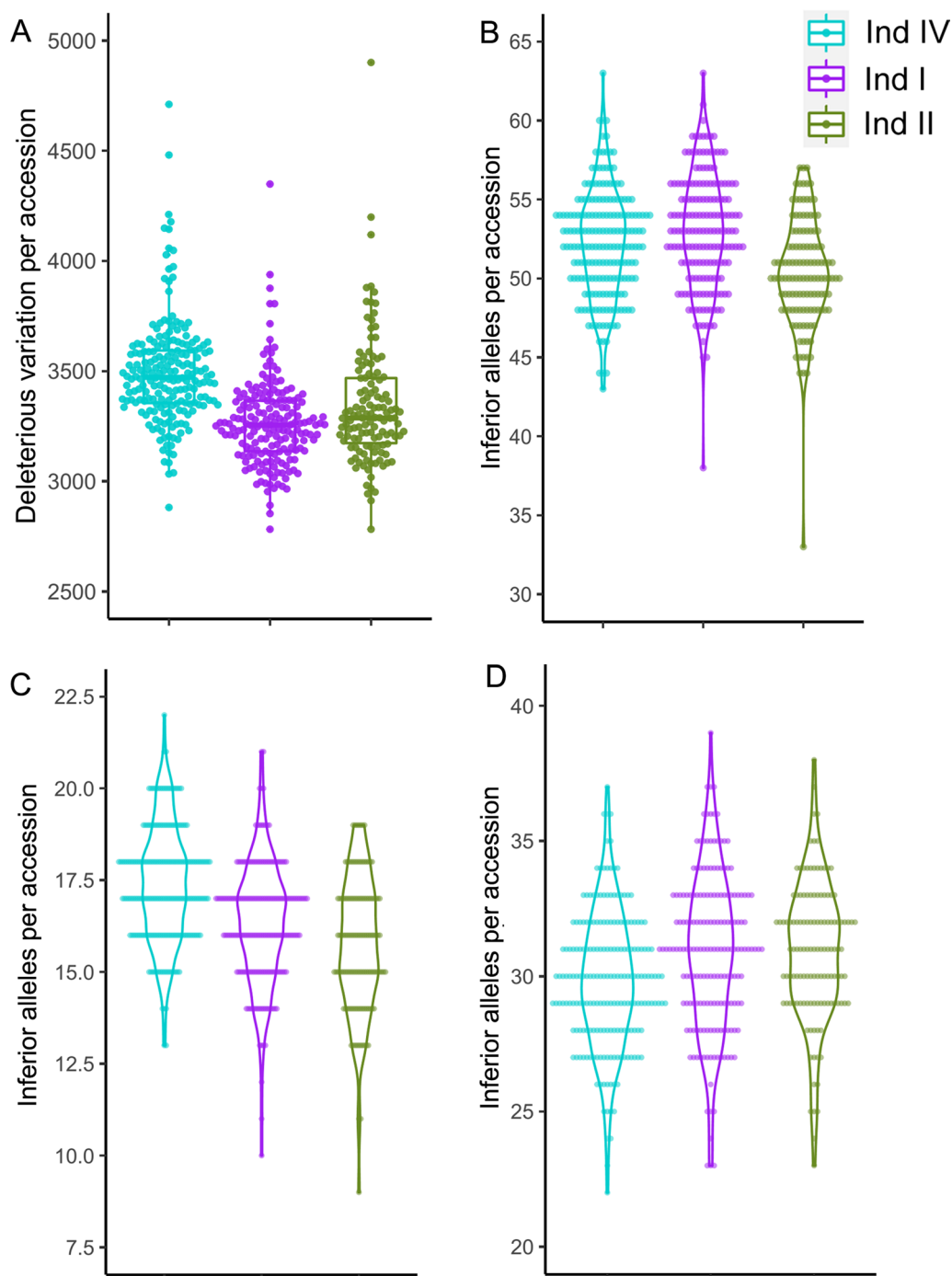
Fig. 4 Deleterious variations in accessions of three main subpopulations. **A** Number of deleterious variations that predicted by SIFT (sorting intolerant from tolerant) software. Number of inferior alleles of all genes (**B**), yield related genes (**C**) and stress related genes (**D**) that annotated using RiceNavi database. Subpopulations were ordered by LD decay distance from short to long

alleles of *GS3, Osmyb3* and *FLO13* were selected for grain length. Interestingly, different selected genes were detected for subpopulation Ind IV and Ind II. Plant height genes *sd1* and *Oshox4* were selected in Ind IV, and the selected genes in Ind I were *OsGA2ox5* and *Oshox4.* For grain yield per plant, *TAC3* and *OsWDR5* were under selection pressure in Ind I but not in Ind IV. *GS3* was selected for grain length in Ind IV but not in Ind I. *GSE5* and *OsDER1* were selected for grain width
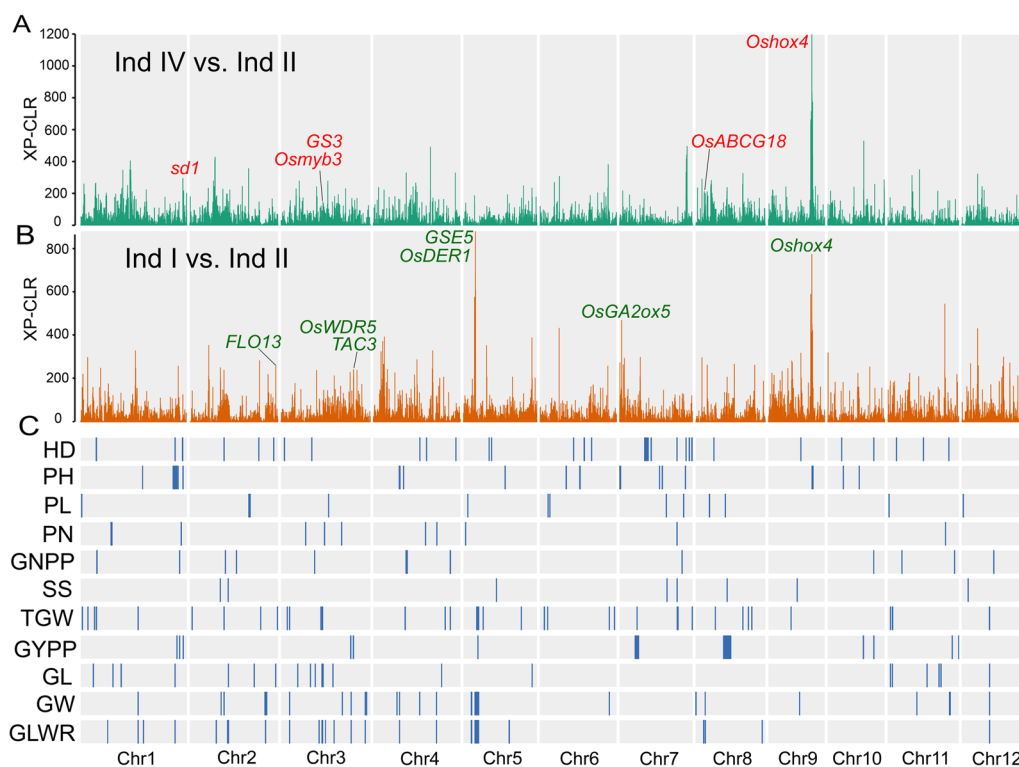
Fig. 5 Selection sweeps for modern cultivar Ind II and the annotation of breeding signatures by QTLs of eleven agronomic traits. Distribution of XP-CLR scores between modern cultivar Ind II and local landrace subpopulations of Ind IV (**A**) and Ind I (**B**). **C** Annotation of breeding signatures using QTLs identified by GWAS

and grain length width ratio in subpopulation Ind I, but *OsABCG18* was selected in Ind IV (Fig. 5).

**Allele Fixation During the Breeding Process of Guangdong Indica Rice**

Gene haplotype analysis were conducted to illustrate evolutionary relationship and identify selected alleles during modern breeding and improvement of subpopulation Ind II. A total of 6 potentially favorable alleles were fixed in Ind II for key yield related genes, those genes were *Oshox4* and *OsGA2ox5* for plant height, *TAC3* for tiller angle and yield, *GS3*, *Osmyb3* and *FLO3* for grain size and weight. Four main haplotypes were identified for plant height gene *Oshox4*, haplotype network analysis revealed that *Oshox4*-Hap3 (Ind II fixed haplotype) was derived from *Oshox4*-Hap2, following the variations of *Oshox4*-Hap1 (Ind IV and Ind I) and *Oshox4*-Hap4 (GJ-tmp). For two main haplotypes of plant height gene *OsGA2ox5*, most accessions of Ind II have *OsGA2ox5*-Hap2, while landraces of subpopulation Ind I and Ind IV have *OsGA2ox5*-Hap1. Tiller angle and yield related gene *TAC3* have four main haplotypes in these accessions, and *TAC3*-Hap3 are a fixed haplotype of Ind II modern cultivars. Grain length and weight gene *GS3* have five

main haplotypes, *GS3*-Hap3 was fixed during breeding and improvement process for Ind II. Hap3 of grain length gene *Osmyb3* was mainly selected for Ind II during modern rice breeding by one and two variations from *Osmyb3*-Hap4 and *Osmyb3*-Hap2. For the four haplotypes of grain weight gene *FLO13*, Hap3 is a predominant allele in Ind II, which was selected from *FLO13*-Hap4 with one missense variation (Fig. 6). In those modern cultivar fixed alleles, six Ind II specific variations were identified. *Oshox4*-Hap3 have one specific intron variation between exon 1 and exon 2, *TAC3*-Hap3 have three cultivar specific variations, *GS3*-Hap3 have one stop codon gained variation, and *FLO13*-Hap3 have one specific missense variation in Ind II (Fig. 7).

Discussion

The major breakthrough of "green revolution" leads quantum leaps of rice productivity (Cheng et al. 2020), intensive breeding efforts and artificial selection have facilitated the significant improvement of *indica* rice yield in Guangdong, where the "green revolution" started in China. Unlike previously reported selection analysis of rice accession from multiple geographic positions (Li et al. 2020a, b; Lv et al. 2020; Xie et al. 2015; Xu et al.
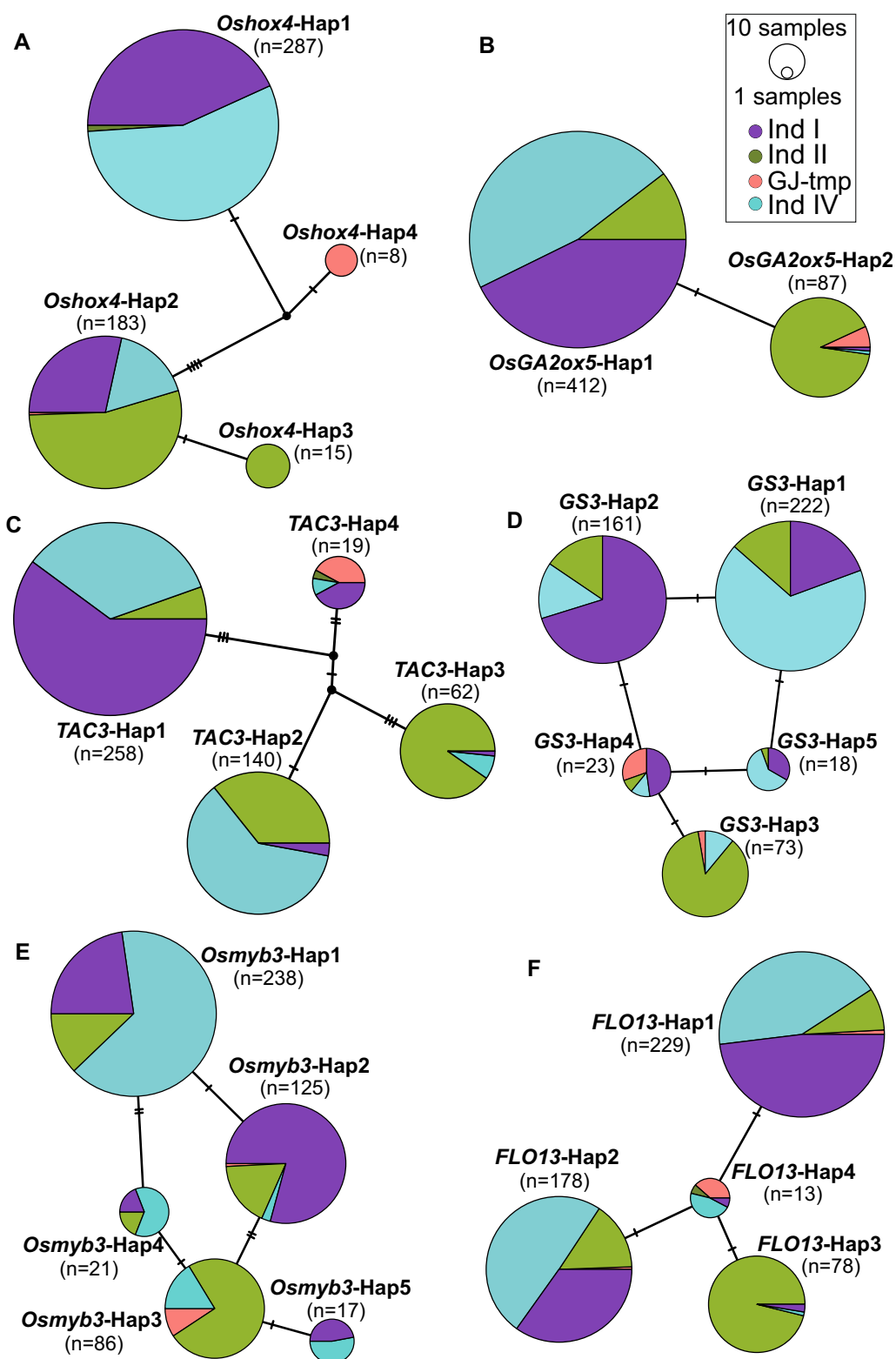
Fig. 6 Haplotype networks of six key yield related genes in selected regions. **A** *Oshox4* related to plant height, **B** *OsGA2ox5* related to plant height, **C** *TAC3* related to tiller angle, **D** *GS3* related to grain length and weight, **E** *Osmyb3* related to grain length, and **F** *FLO13* related to grain weight. Circle size is proportional to accession number of one haplotype. Number of ticks in network edges represents the variation counts between haplotypes
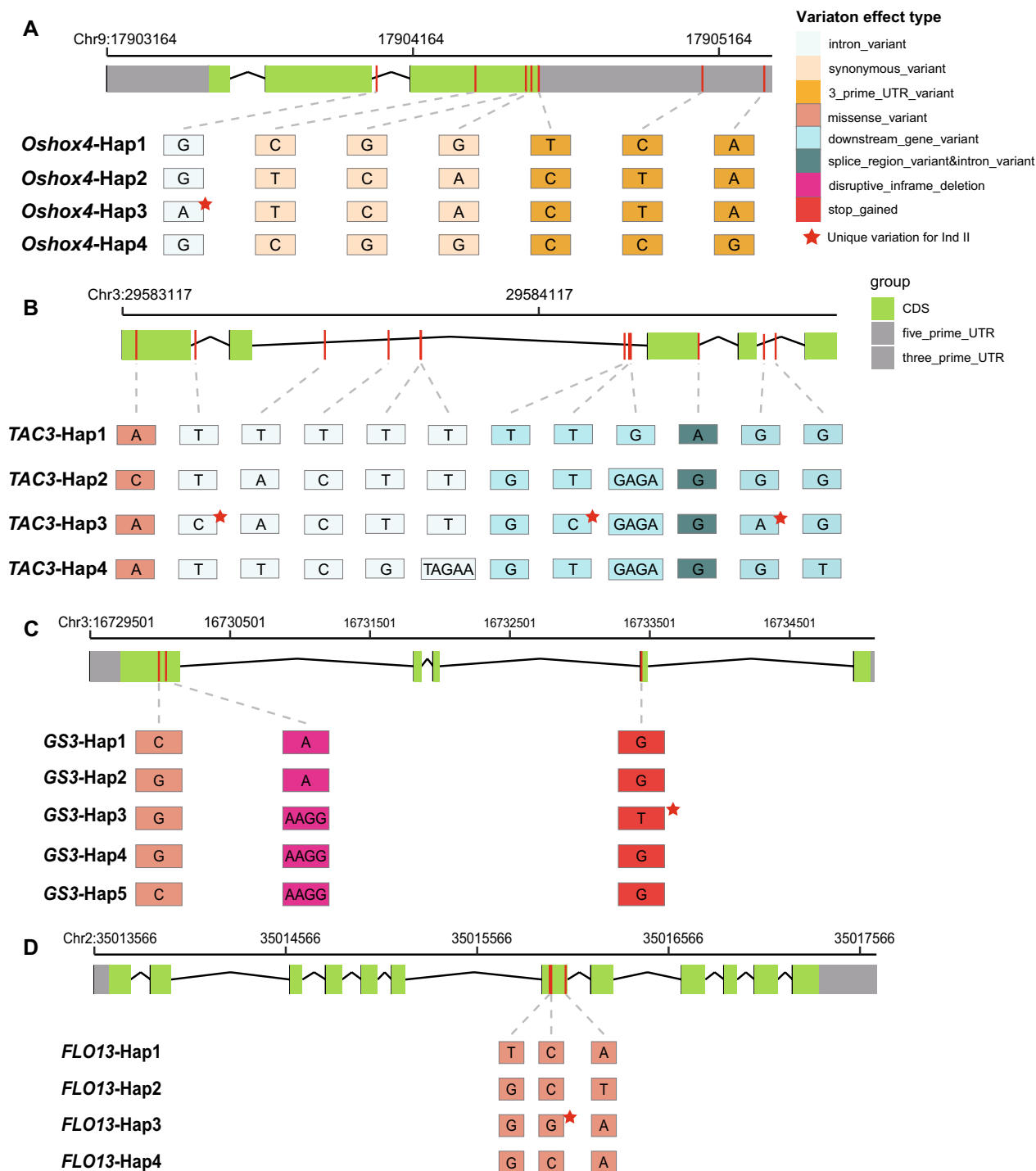
Fig. 7 Gene structure and haplotype specific variations of *Oshox4* (**A**), *TAC3* (**B**), *GS3* (**C**) and *FLO13* (**D**) for subpopulation Ind II. Background colors of bases represent annotation and effect level of the variation. Red stars mark the specific variations for the haplotype of Ind II

2016; Ye et al. 2022), we focusing on the locally adaptive selection of rice in Guangdong by comparing the genomic variations and agronomic traits of locally cultivated landraces by farmers before "green revolution" and modern improved cultivars. For example, modern breeding of rice quality started from the end of last century in Guangdong favors slender grain type for regional appetite in southern Aisa, which makes the increase of grain

length and width ratio, and decrease of grain weight in subpopulation of modern cultivars.

The artificial replacement of deleterious variations with favorable alleles during breeding and improvement are meaningful to meet social demands for crop and food. By integrating GWAS for agronomic traits and breeding signatures for selected regions, regionally selected key genes that influencing yield and yield related traits for Guangdong *indica* rice were identified. *Oshox4* was identified as a selected gene for plant height in our accessions, which plays negative function in gibberellin responses and influencing plant height and tiller number (Dai et al. 2008; Zhou et al. 2015). An Ind II specific haplotype (*Oshox4*-Hap3) were identified for cultivars. Another regulator for rice growth and architecture, *OsGA2ox5*, were also identified as selected gene of plant height for cultivar subpopulation Ind II when compared with Ind I (Lo et al. 2008), and *OsGA2ox5*-Hap2 was a selected haplotype by cultivars from Ind II. Tiller angle gene *TAC3* were selected for yield per plant of Ind II when compared with Ind I, and *TAC3*-Hap3 were selected by cultivars from Ind II (Dong et al. 2016). *GS3* and *Osmyb3* were identified to be selected for Ind II when compared with Ind IV, and the Hap3 of these two genes were selected for cultivars (Li et al. 2020a, b; Fan et al. 2006; Mao et al. 2010). Starch biosynthesis and grain weight gene *FLO13* was selected for Ind II compared with Ind I, and *FLO13*-Hap3 with a unique missense variant in cultivars was a potential selected haplotype (Hu et al. 2018). These selected favorable haplotypes are promising functional alleles regulating the yield improvement of Guangdong modern cultivars.

Conclusions
In summary, large-scale genomic and yield assessment of Guangdong landrace and modern cultivars promotes analysis of their diversity, classification and phylogenetic relationship. We revealed less deleterious variations number in modern cultivars than landrace. Selected genomic regions were also identified and annotated using GWAS of vital agronomic traits, which leads the identification of selected key genes during Guangdong rice breeding and improvement. These results shed light on regionally breeding trajectory and artificial selection, and provides valuable resources for rational design of molecular breeding.

Materials and Methods
Field Experimental Design and Phenotyping of Agronomic Traits
Agronomic traits of 479 accessions of Guangdong rice core germplasm was investigated on experimental field of Rice Research Institute, Guangdong Academy of Agricultural Sciences for two seasons of 2019 and 2021 under conventional field management. All accessions were planted in 1437 blocks under randomized complete block design with three replications. A total of eleven agronomic traits were investigated and calculated. Plant height (PH) was measured as length from the ground to the highest point of the plant, heading date (HD) was recorded when half of the plants in a block have reached the heading stage and days from sowing to heading is calculated, yield per plant (YPP) was the total weight of filled grains per plant, panicle number (PN) was the number of effective panicles per mature plant, grain number per panicle (GNPP) was the mean total number of grains per panicle on a single plant, seed setting (SS) was calculated as the percentage of filled grains to total grains per plant, grain length (GL), grain width (GW) and panicle length (PL) were measured when seeds are mature (Yu et al. 2020). Thousand grain weight (TGW) were calculated by the division of total weight to total filled grain number, and grain length width ratio (GLWR) was the division of GL to GW.

Genomic Resequencing and Variation Calling
Young leaves of 358 landrace and 121 improved cultivars of *indica* rice accessions from Guangdong province of southern China were collected to construct sequencing libraries according to the manufacturer's instructions, and qualified libraries were sequenced using Illumina HiSeq platform. A total of 38 Guangdong improved cultivars were collected from the NCBI SRA database with accession numbers of PRJNA321462, PRJNA522896 and PRJNA656900 (Additional file 1: Table S1). Quality of raw sequencing data were accessed using FastQC (v0.11.9) software (Andrews 2010), and low-quality data were trimmed using TrimGalore (version 0.6.6) to generate clean sequencing data. Clean data were mapped onto reference genome (MSU7) using BWA (0.7.17-r1188) software with default parameter (Li and Durbin 2009). MarkDuplicates in Picard (2.12.1) was used to eliminate PCR duplication and sorting BAM files, and genome-CoverageBed of bedtools (v2.27.1) was used to calculate genome coverage ratios. SNPs (single nucleotide polymorphisms) and InDels (insertions and deletions) were then called using HaplotypeCaller of Genome Analysis Toolkit (GATK, version 4.2.2.0) pipeline (McKenna et al. 2010), and annotated using SnpEff (4.3 s) with the GFF3 file of MSU7 reference genome (Cingolani et al. 2012).

Population Structure Analysis
Principal components analysis (PCA), phylogenetic and admixture analysis were employed to classify subpopulations (Yu et al. 2021). For population structure analysis, SNP variations were filtered using VCFtools (0.1.16)

software with parameter "–max-missing 0.95" and "–maf 0.05". PCA method with kmeans clustering algorithm in CropGBM software was used to reducing dimensions of genotypic data (Yan et al. 2021), and eigenvalues were calculated by plink software. Phylogenetic relationship was constructed using VCF2Dis software (https://github.com/BGI-shenzhen/VCF2Dis) and illustrated by using FigTree (v1.4.3) software (https://github.com/rambaut/figtree). ADMIXTURE (version 1.3.0) software (Alexander and Lange 2011) was used to analyze population structure with *k* values ranged from 2 to 12. The ancestry distributions of individuals were visualized using R script. Genetic diversity (*pi* and *Tajima's* D) and differentiation (*fst*) analysis were conducted using VCFtools (0.1.16) software with 100 kb sliding windows. Linkage disequilibrium (LD) decay for each subpopulation was estimated and plotted using PopLDdecay (Zhang et al. 2019).

National *indica* rice sequencing data from RiceVarMap database (Zhao et al. 2015), variations of 3024 3KRG accession from SNP-seek database (Locedie et al. 2017) and variations of 998 wild rice lines from our recently research (Zhang et al. 2022) that used to conduct population structure and phylogenetic analysis in this study were subjected to the same data processing pipeline. A recently released and refined *indica* 9311 reference genome was also used to check the results of population structure analysis (Wang et al. 2022).

Estimation of Variation Effects and Deleterious Mutation Prediction

Functional alteration of genomic variations was predicted using Sorting Intolerant From Tolerant 4G (SIFT 4G) software (Vaser et al. 2016). Variations with SIFT scores smaller than 0.05 was considered as putatively deleterious variations. Allele function for known genes with vital role in rice were annotated using RiceNavi database (Wei et al. 2021). Advantage and inferior allele were determined by manually check of allele functional alteration and its corresponding trait. Number of inferior alleles for accessions from each subpopulation were counted and plotted using boxplot or violin plot in R.

Identification of Breeding Signatures Using XP-CLR

Breeding signatures for artificial selection were identified by using the cross-population composite likelihood ratio test (XP-CLR) method (Chen et al. 2010) and its updated version of python module (https://github.com/hardingnj/xpclr). XP-CLR was conducted between subpopulations of landrace and cultivar with 10 kb sliding windows. Genomic segments with XP-CLR values above the 80th percentile was considered as putatively selected regions. Adjacent segments within 20 kb distance were then merged into longer blocks, and blocks shorter than 40 kb

were filtered out as such short blocks unlikely selected during the short history of modern rice improvement breeding. Long blocks with top 1% values of XP-CLR scores were finally considered as selected regions.

Genome-Wide Association Analysis

For genome-wide association analysis (GWAS), multi-sample VCF file of genomic variations was converted into plink file and variations were screened with parameters of "–geno 0.1 –mind 0.4 –maf 0.05". PCA analysis were conducted by plink software with five major components (Purcell et al. 2007). Kinship analysis and GWAS were conducted using GEMMA software using filtered genotypes and eleven agronomic traits (Zhou and Stephens 2012).

Gene Haplotype Reconstruction and Network Analysis

Software beagle (version 5.2) was used to impute missing genetic variations that generated by GATK (Browning et al. 2021). Genomic variations of selected genes were extracted based on the positions by using BCFTools (Li 2011). Haplotype network of these genes were constructed by our previously described method (Yu et al. 2021). Haplotype network was constructed and illustrated by Popart software (Leigh and Bryant 2015).

Abbreviations

SNP	Single-nucleotide polymorphisms
PCA	Principal component analysis
XP-CLR	The cross-population composite likelihood ratio test
QTLs	Quantitative trait locus
GWAS	Genome-wide association studies
LD	Linkage disequilibrium
QTNs	Quantitative trait nucleotides
KRG	3000 Rice genome project

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12284-023-00642-3.

Additional file 1: **Table S1**. Information for accessions used in this study. **Table S2**. Quality assessment of genome sequencing data. **Table S3**. Genome mapping quality and genome coverage of genome sequencing data. **Table S4**. Genomic variations for 517 indica rice accessions against the MSU reference genome. **Table S5**. Number of deleterious variations in subpopulation of landrace and cultivar. **Table S6**. Genotyping results 319 quantitative trait nucleotides (QTNs) of the 212 vital gene in rice from RiceNavi database. **Table S7**. Genomic segments that identified as breeding signatures between Ind IV and Ind II. **Table S8**. Genomic segments that identified as breeding signatures between Ind I and Ind II. **Table S9**. QTLs and known genes that identified by GWAS of eleven agronomic traits for Guangdong indica rice. **Table S10**. Allele combinations of plant height (PH) genes of Guangdong indica rice germplasm. **Table S11**. Allele combinations of thousand grain weight (TGW), grain length (GL), grain width (GW) and grain length width ratio (GLWR) genes of Guangdong indica rice germplasm. **Table S12**. Phenotypic effect assessment of known QTNs that genotyped by RiceNavi. **Table S13**. Genotyping of 4 eating quality genes.

Declarations

Ethics Approval and Consent to Participate
Not applicable.

Consent for Publication
Not applicable.

Competing Interests
The authors have declared that no competing interests exist.

References
Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinf 12:246

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Browning BL, Tian X, Zhou Y, Browning SR (2021) Fast two-stage phasing of large-scale sequence data. Am J Hum Genet 108:1880–1890

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20:393–402

Cheng F, Quan X, Zhengjin X, Wenfu C (2020) Effect of rice breeding process on improvement of yield and quality in China. Rice Sci 27:363–367

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. Fly 6:80–92

Cui D, Lu H, Tang C, Li J, Yu T, Ma X, Zhang E, Wang Y, Cao G, Xu F, Qiao Y, Dai L, Li R, Tian S, Koh HJ, Han L (2019) Genomic analyses reveal selection footprints in rice landraces grown under on-farm conservation conditions during a short-term period of domestication. Evolut Appl 13:290–302

Dai M, Hu Y, Ma Q, Zhao Y, Zhou D (2008) Functional analysis of rice *HOMEOBOX4* (*Oshox4*) gene reveals a negative function in gibberellin responses. Plant Mol Biol 66:289–301

Dong H, Zhao H, Xie W, Han Z, Li G, Yao W, Bai X, Hu Y, Guo Z, Lu K, Yang L, Xing Y (2016) 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

Fan C, Xing Y, Mao H, Lu T, Han B, Xu C, Li X, Zhang Q (2006) *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

Fuentes RR, Chebotarov D, Duitama J, Smith S, De la Hoz JF, Mohiyuddin M, Wing RA, McNally KL, Tatarinova T, Grigoriev A, Mauleon R, Alexandrov N (2019) Structural variants in 3000 rice genomes. Genome Res 29:870–880

Higgins J, Santos B, Khanh TD, Trung KH, Duong TD, Doai NTP, Khoa NT, Ha DTT, Diep NT, Dung KT, Phi CN, Thuy TT, Tuan NT, Tran HD, Trung NT, Giang HT, Nhung TK, Tran CD, Lang SV, Nghia LT, Van Giang N, Xuan TD, Hall A, Dyer S, Ham LH, Caccamo M, De Vega JJ (2021) Resequencing of 672 native rice accessions to explore genetic diversity and trait associations in Vietnam. Rice 14:1–16

Hour A, Hsieh W, Chang S, Wu Y, Chin H, Lin Y (2020) Genetic diversity of landraces and improved varieties of rice (*Oryza sativa* L.) in Taiwan. Rice 13:1–12

Hu T, Tian Y, Zhu J, Wang Y, Jing R, Lei J, Sun Y, Yu Y, Li J, Chen X, Zhu X, Hao Y, Liu L, Wang Y, Wan J (2018) *OsNDUFA9* encoding a mitochondrial complex I subunit is essential for embryo development and starch synthesis in rice. Plant Cell Rep 37:1667–1679

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang Q, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Huang X, Kurata N, Wei X, Wang Z, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501

Huang X, Yang S, Gong J, Zhao Q, Feng Q, Zhan Q, Zhao Y, Li W, Cheng B, Xia J, Chen N, Huang T, Zhang L, Fan D, Chen J, Zhou C, Lu Y, Weng Q, Han B (2016) Genomic architecture of heterosis for yield traits in rice. Nature 537:629–633

Kumar A, Daware A, Kumar A, Kumar V, Gopala KS, Mondal S, Patra BC, Singh AK, Tyagi AK, Parida SK, Thakur JK (2020) Genome-wide analysis of polymorphisms identified domestication-associated long low-diversity region carrying important rice grain size/weight quantitative trait loci. Plant J 103:1525–1547

Leigh JW, Bryant D (2015) POPART: full-feature software for haplotype network construction. Methods Ecol Evol 6:1110–1116

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

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

Li Q, Lu L, Liu H, Bai X, Zhou X, Wu B, Yuan M, Yang L, Xing Y (2020a) A minor QTL, *SG3*, encoding an R2R3-MYB protein, negatively controls grain length in rice. Theor Appl Genet 133:2387–2399

Li X, Chen Z, Zhang G, Lu H, Qin P, Qi M, Yu Y, Jiao B, Zhao X, Gao Q, Wang H, Wu Y, Ma J, Zhang L, Wang Y, Deng L, Yao S, Cheng Z, Yu D, Zhu L, Xue Y, Chu C, Li A, Li S, Liang C (2020b) 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

Liu C, Peng P, Li W, Ye C, Zhang S, Wang R, Li D, Guan S, Zhang L, Huang X, Guo Z, Guo J, Long Y, Li L, Pan G, Tian B, Xiao J (2021) Deciphering variation of

239 elite japonica rice genomes for whole genome sequences-enabled breeding. Genomics 113:3083–3091

Lo S, Yang S, Chen K, Hsing Y, Zeevaart JAD, Chen L, Yu S (2008) A novel class of gibberellin 2-oxidases control semidwarfism, tillering, and root development in rice. Plant Cell 20:2603–2618

Locedie M, Roven RF, Frances NB, Jeffery D, Juan MA, Dmytro C, Millicent S, Kevin P, Dario C, Alexandre P, Inna D, Victor S, Rod AW, Ruaraidh SH, Ramil M, Kenneth LM, Nickolai A (2017) Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res 45(D1):D1075–D1081

Lv Q, Li W, Sun Z, Ouyang N, Jing X, He Q, Wu J, Zheng J, Zheng J, Tang S, Zhu R, Tian Y, Duan M, Tan Y, Yu D, Sheng X, Sun X, Jia G, Gao H, Zeng Q, Li Y, Tang L, Xu Q, Zhao D, Huang Z, Lu H, Li N, Zhao J, Zhu L, Li D, Yuan L, Yuan D (2020) Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. Nature Commun 11:4778

Mao H, Sun S, Yao J, Wang C, Yu S, Xu C, Li X, Zhang Q (2010) Linking differential domain functions of the *GS3* protein to natural variation of grain size in rice. Proc Natl Acad Sci 107:19579–19584

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC (2016) SIFT missense predictions for genomes. Nat Protoc 11:1–9

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao U, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann J, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49

Wang S, Gao S, Nie J, Tan X, Xie J, Bi X, Sun Y, Luo S, Zhu Q, Geng J, Liu W, Lin Q, Cui P, Hu S, Wu S (2022) Improved 93–11 genome and time-course transcriptome expand resources for rice genomics. Front Plant Sci 12:769700

Wei X, Qiu J, Yong K, Fan J, Zhang Q, Hua H, Liu J, Wang Q, Olsen KM, Han B, Huang X (2021) A quantitative genomics map of rice provides genetic insights and guides breeding. Nat Genet 53:243–253

Xiao N, Pan C, Li Y, Wu Y, Cai Y, Lu Y, Wang R, Yu L, Shi W, Kang H, Zhu Z, Huang N, Zhang X, Chen Z, Liu J, Yang Z, Ning Y, Li A (2021) Genomic insight into balancing high yield, good quality, and blast resistance of japonica rice. Genome Biol 22:1–22

Xie W, Wang G, Yuan M, Yao W, Lyu K, Zhao H, Yang M, Li P, Zhang X, Yuan J, Wang Q, Liu F, Dong H, Zhang L, Li X, Meng X, Zhang W, Xiong L, He Y, Wang S, Yu S, Xu C, Luo J, Li X, Xiao J, Lian X, Zhang Q (2015) Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. Proc Natl Acad Sci 112:E5411–E5419

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111

Xu Q, Yuan X, Wang S, Feng Y, Yu H, Wang Y, Yang Y, Wei X, Li X (2016) The genetic diversity and structure of Indica rice in China as detected by single nucleotide polymorphism analysis. BMC Genet 17:1–8

Yan J, Xu Y, Cheng Q, Jiang S, Wang Q, Xiao Y, Ma C, Yan J, Wang X (2021) LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol 22:1–24

Ye J, Zhang M, Yuan X, Hu D, Zhang Y, Xu S, Li Z, Li R, Liu J, Sun Y, Wang S, Feng Y, Xu Q, Yang Y, Wei X (2022) Genomic insight into genetic changes and shaping of major inbred rice cultivars in China. New Phytol 236:2311

Yu H, Shahid MQ, Li Q, Li Y, Li C, Lu Z, Wu J, Zhang Z, Liu X (2020) Production assessment and genome comparison revealed high yield potential and novel specific alleles associated with fertility and yield in neo-tetraploid rice. Rice 13:32

Yu H, Li Q, Li Y, Yang H, Lu Z, Wu J, Zhang Z, Shahid MQ, Liu X (2021) Genomics analyses reveal unique classification, population structure and novel allele of neo-tetraploid rice. Rice 14:16

Zhang C, Dong S, Xu J, He W, Yang T (2019) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 35:1786–1788

Zhang J, Pan D, Fan Z, Yu H, Jiang L, Lv S, Sun B, Chen W, Mao X, Liu Q, Li C (2022) Genetic diversity of wild rice accessions (*Oryza rufipogon* Griff.) in Guangdong and Hainan provinces, China, and construction of a wild rice core collection. Front Plant Sci 13:999454

Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, Xing Y, Chen L, Xie W (2015) RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res 43(D1):D1018–D1022

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44:821–824

Zhou W, Malabanan PB, Abrigo E (2015) *OsHox4* regulates GA signaling by interacting with DELLA-like genes and GA oxidase genes in rice. Euphytica 201:97–107

Publisher's Note