Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2023) **21**, pp. 202–218 doi: 10.1111/pbi.13938

Time-ordering *japonica/geng* genomes analysis indicates the importance of large structural variants in rice breeding

Yu Wang[1,2,3,†], Fengcheng Li[1,†], Fan Zhang[4,†], Lian Wu[1,†], Na Xu[1], Qi Sun[1], Hao Chen[1], Zhiwen Yu[1], Jiahao Lu[1], Kai Jiang[1], Xiaoche Wang[1], Siyu Wen[2,3], Yao Zhou[2,3], Hui Zhao[2,3], Qian Jiang[2,3], Jiahong Wang[5], Ruizong Jia[2,3], Jian Sun[1], Liang Tang[1], Hai Xu[1], Wei Hu[2,3], Zhengjin Xu[1], Wenfu Chen[1], Anping Guo[2,3,*] and Quan Xu[1,*]

[1]*Rice Research Institute of Shenyang Agricultural University, Shenyang, China*
[2]*Sanya Research Institute of Chinese Academy of Tropical Agricultural Sciences, Sanya, China*
[3]*Hainan Key Laboratory for Biosafety Monitoring and Molecular Breeding in Off-Season Reproduction Regions, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, China*
[4]*Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China*
[5]*Biomarker Technologies Corporation, Beijing, China*

Received 9 July 2022;
revised 23 August 2022;
accepted 29 September 2022.
**Correspondence* (Tel +86 024-88487011; fax +86 024-88417415; email kobexu34@syau.edu.cn (Q.X.) and Tel +86 0898-66962906; fax +86 0898-66962904; email gap211@126.com (A.G.))
†These authors contributed equally.

Keywords: *Oryza sativa*, *japonica/geng*, *de novo* assembly, structural variations, breeding process, gene editing.

Summary

Temperate *japonica/geng* (GJ) rice yield has significantly improved due to intensive breeding efforts, dramatically enhancing global food security. However, little is known about the underlying genomic structural variations (SVs) responsible for this improvement. We compared 58 long-read assemblies comprising cultivated and wild rice species in the present study, revealing 156 319 SVs. The phylogenomic analysis based on the SV dataset detected the putatively selected region of GJ sub-populations. A significant portion of the detected SVs overlapped with genic regions were found to influence the expression of involved genes inside GJ assemblies. Integrating the SVs and causal genetic variants underlying agronomic traits into the analysis enables the precise identification of breeding signatures resulting from complex breeding histories aimed at stress tolerance, yield potential and quality improvement. Further, the results demonstrated genomic and genetic evidence that the SV in the promoter of *LTG1* is accounting for chilling sensitivity, and the increased copy numbers of *GNP1* were associated with positive effects on grain number. In summary, the current study provides genomic resources for retracing the properties of SVs-shaped agronomic traits during previous breeding procedures, which will assist future genetic, genomic and breeding research on rice.

Introduction

China was one of the first countries to domesticate and cultivate rice, and it is now the world's largest producer and consumer of rice (Muthayya *et al*., 2014). The last century has witnessed quantum leaps in the rice productivity of China from 1.9 t/hm^2 in 1949 to 7.0 t/hm^2 in 2018 (http://faostat.fao.org/). This increased rice productivity could be primarily attributed to the introduction of semi-dwarf and hybrid rice varieties, and partially to the rapid expansions of high-yielding temperate GJ varieties. In the national new rice varietal trials in northeast China, the average yield of new temperate GJ varieties was from 8.3 t/hm^2 in 2004 to 9.0 t/hm^2 in 2018 (Fei *et al*., 2020). Despite tremendous gains in GJ rice breeding, the genomic modifications caused by yield enhancements obtained during previous breeding procedures are mainly unknown.

With the application of next-generation sequencing technology (NGS), diverse rice accessions have been sequenced in recent years (Huang *et al*., 2010, 2012). Population genomic research on the evolution and domestication of rice has progressed significantly (Wang *et al*., 2018; Zhang *et al*., 2021). The breeding signature, which was referred to as genomic changes associated with breeding efforts, was also identified using NGS (Chen *et al*., 2020; Cui *et al*., 2022; Xie *et al*., 2015). Although copy number variants (CNVs) and presence/absence variants (PAVs) are known to have played important roles in the genetic regulation of agronomical traits, short-read sequences by NGS have limited power in identifying these SVs (Cook *et al*., 2012; Deng *et al*., 2017; Hirsch *et al*., 2014; Hufford *et al*., 2012; Lu *et al*., 2015; Lye and Purugganan, 2019; Shomura *et al*., 2008; Xu *et al*., 2006). Although recently reported pan-genome analysis of 33 and 251 genetically diverse rice accessions have revealed hidden SVs among *Oryza sativa indica* (XI), GJ and *Oryza glaberrima* (Qin *et al*., 2021; Shang *et al*., 2022), it remains unclear how SVs behaved and acted as an important contributor to trait improvements in the breeding process of temperate GJ varieties.

Here, we analysed the 58 long-read assemblies, including 12 time-ordering *de novo* assembled high-quality genomes for diverse GJ varieties bred from 1882 to 2011 that played important roles during the history of GJ rice breeding, and 46 existing long-read assemblies (Qin *et al*., 2021; Stein *et al*., 2018; Zhang *et al*., 2022) to answer the question of how SVs behaved and contributed to the GJ breeding by analysing the distribution and effects of SVs in these genomes and by identifying critical CNVs during temperate GJ rice breeding that could not be detected by short-read sequences. Beyond showcasing the power of high-quality genome assemblies for plant genomics and functional genomics research, the newly identified SVs will facilitate the genetic improvement of GJ through both marker-assisted selection and genomic selection.

Results

De novo assembly and annotation of 12 GJ rice genomes

We chose six interrelated modern Chinese temperate GJ varieties and six related Japanese GJ varieties to investigate genome enhancement in temperate GJ varieties during modern breeding. After their release, all 12 varieties were extensively grown from 1882 to 2011 in Japan and China. As core parents in breeding efforts, they have made significant contributions to the enhancement of temperate GJ rice types (Figure 1a,b). The Nanopore libraries of the 12 varieties were constructed and sequenced individually with an average coverage depth of 117× (Table S1). Illumina sequencing (HiSeq) with an average coverage depth of 65× (Table S2) and chromosome conformation capture (Hi-C) sequencing with an average depth of 372× were also performed (Table S3). Then, we generated the *de novo* genome assembly for each variety. The Hi-C data were used to further correct the assembled genome; the scaffolds were clustered, ordered and oriented onto chromosomes (Figure S1). The contig N50 sizes of the 12 genome assemblies ranged from 8.77 to 15.84 Mb with a mean of 13.68 Mb (Table S4). Finally, the 12 assembled genome sizes ranged from 379.07 to 385.87 Mb with a mean of 380.41 Mb. For each variety, an average of 99.54% of contigs were anchored to the chromosomes. The Illumina readings from each variety were then re-mapped onto the assembled genomes. The mapping ratio reached 98.69%, indicating that each constructed genome was nearly complete. We identified, on average, 40 079 protein-coding genes for each assembly. The completeness estimated by Benchmarking Universal Single-Copy Orthologs (BUSCO; Simão *et al.*, 2015) was 97.9% (Table S4).

Genomic analysis of 58 long-read *de novo* assemblies

We then collected 55 *de novo* assembled genomes of wild-type, XI, GJ and circum-Aus group (cA), which encompasses the Aus, Boro and Rayada ecotypes from Bangladesh and Indica, and circum-Basmati group (cB), which comprises the Basmati and Sadri aromatic varieties published based on the long-read sequencing techniques (Qin *et al.*, 2021; Stein *et al.*, 2018; Zhang *et al.*, 2022). After discarding duplicated assemblies, 45 collected assemblies and our 12 assembled genomes were compared to the genome of Nip (Kawahara *et al.*, 2013) using MUMmer (v 4.0) (Marçais *et al.*, 2018; Table S4). We observed a total of 156 319 SVs >50 bp relative to Nip, which could be converted into five types, including 41 331 insertions (INS), 57 184 deletions (DEL), 567 inversions (INV), 17 281 translocations (TRA) and 39 956 other types SVs (NOTAL: not aligned region, HDR: highly diverged region, TDM: tandem repeat; Figure 2a, Table S5 and Figure S2). The phylogenetic analysis and principal component analysis (PCA) of these 58 assemblies based on the 156 319 SVs showed that the 12 assemblies were clustered together with seven temperate GJ assemblies (KY131, Kosh, ZH11, 02428, DHX2, TG22 and Nip; Figure 2b,c). Population structure analysis indicated that the slight introgression from XI, cA and cB might occur in the genome of several GJ genomes, such as 02428, DHX2 and YF47 (Figure 2d).

Characterization and SVs related to temperate GJ breeding histories

Subsequently, we concentrated on the GJ and attempted to identify SVs associated with temperate GJ breeding histories. Even though the 17 temperate GJ genomes generally have

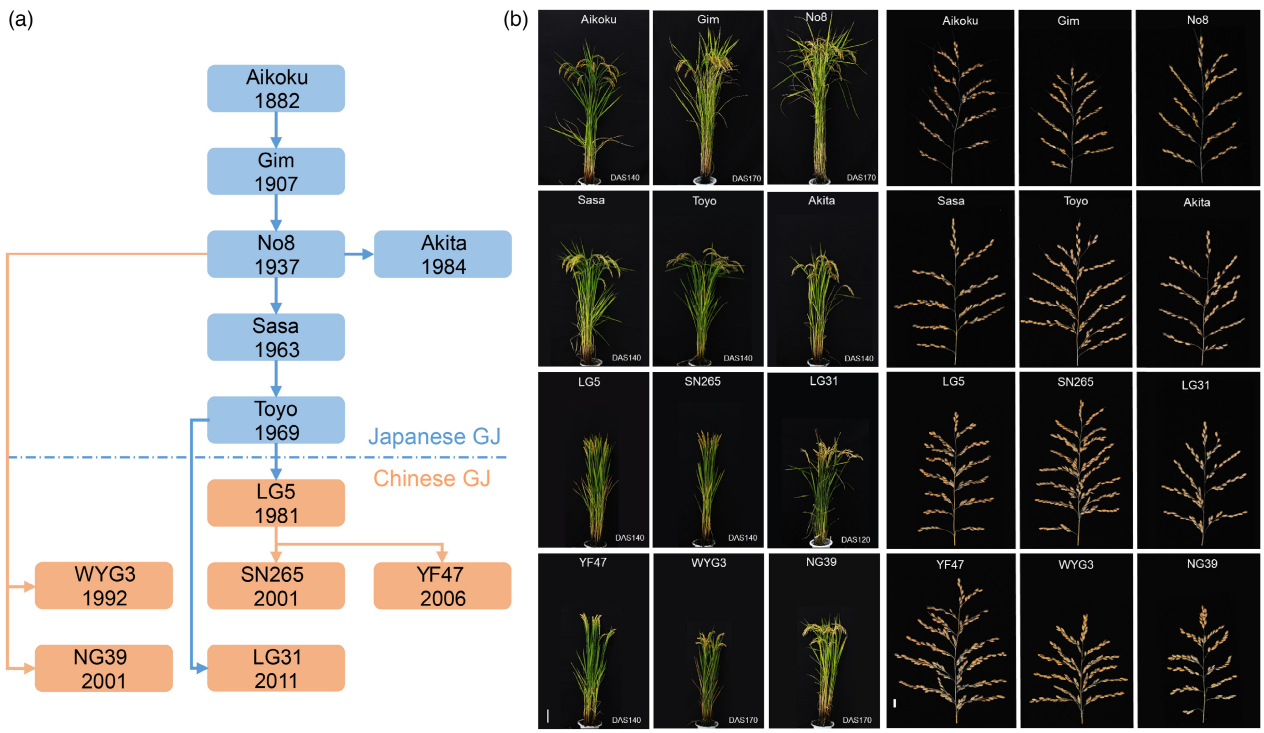


Figure 1 Agronomic phenotypes and pedigree relationship of the genome for 12 GJ varieties. (a) The pedigree relationship among the 12 temperate GJ assemblies. The colours orange and blue stand for Chinese GJ and Japanese GJ respectively. (b) The plant and panicle architecture of 12 temperate GJ varieties assembled in this study.

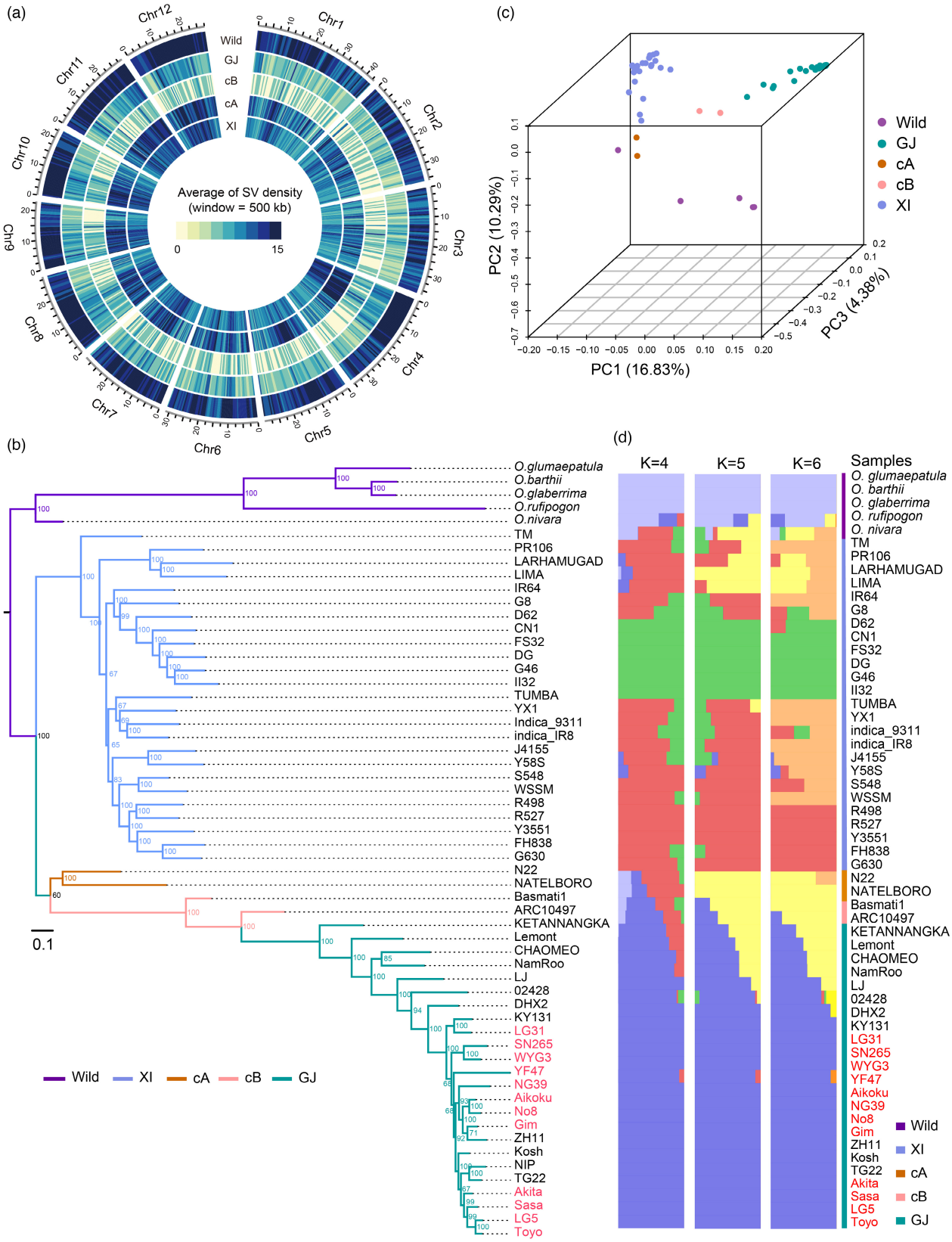


Figure 2 Population structure of 58 long-read assemblies. (a) The average SVs density of 58 assemblies. (b) Phylogenetic tree of 58 accessions including 12 assemblies in this study and 46 existing assemblies. The assemblies with red colour represent the 12 assemblies in the current study. Scale bar = 0.1. (c) Principal component analysis (PCA) plot for 58 *de novo* assemblies. (d) STRUCTURE analysis of 58 accessions with different numbers of clusters K = 4–6. The assemblies with red colour represent the 12 assemblies in the current study.

conservative gene-order syntenic relationships at the chromosome level, we found a total of 52 446 SVs, including 17 618 INS, 18 254 DEL, 205 INV, 5140 TRA and 11 229 other types of SVs (NOTAL: not aligned region and HDR: highly diverged region; Figure 3a). Our data indicated that 21.4% of the detected SVs overlapped with 2 kb upstream regions of rice genes (Figure 3b). Most of the five types of SVs among 18 assemblies were 50–500 bp (Figure S3). Notably, these SVs are unevenly distributed across different chromosomes and are more abundantly present on chromosomes 1, 4, 6, 7, 11 and 12 (Figure 3c, Figures S4 and S5). Figure 3d shows many large SVs in some assemblies, revealing historical events during these varieties' breeding. For example, the inversion across the centromere of Chromosome 6 in 02428 and ZH11 was reported as the most remarkable difference between Nip and R498 (Du *et al*., 2017), indicating that the large-fragment introgression from XI might occur in this region of 02428 and ZH11. To investigate the behaviours of SVs during previous breeding procedures, we classified the discovered SVs into four distinct sets based on their existence in the 17 varieties issued at four different times and attempted to monitor the flow of the SVs throughout the breeding history (Figure 3e). We defined the SVs only existing in one group as specific SVs, and the SVs not only existing in the first batch (shown in pink in Figure 3e) but also inherited into at least one group as common SVs. The result showed that 51.47% of SVs were specific SVs (Figure 3f). The common SVs included 16 114 (30.72%) SVs derived from Japanese GJ varieties prior to1980 and inherited by more recent breeding lines (Figure 3f). We found that the proportion of INS and INV was higher in common SVs compared to that of specific SVs (Figure 3g). The more abundance of INS was expected because of their possible fewer degrees of deleterious effects on plant growth and development, and the INV represses the recombination which might easy to be inherited (Crow *et al*., 2020; Kapun and Flatt, 2019). Moreover, common SVs preferred to be located in the intergenic region (Figure 3h), supporting that a vast majority of SVs in the coding region are deleterious, thus being discarded in breeding (Hämälä *et al*., 2021).

SVs impact gene expression profiles and selection

We evaluated the SV distribution in terms of their relative locations to genes to investigate the potential relevance of the discovered SVs. We determined that 41.5% of the SVs observed overlapped with genic sites. Furthermore, 21.4% of SVs in genic areas matched the 2 kb upstream sequences of gene coding regions, and 48.4% were associated with transposable elements. These amounts remained consistent across all 17 genomes (Figure 3b, Figure S6). The GO enrichment analysis and Genes and Genomes (KEGG) pathway analysis of the SVs located in the gene promoter, UTR and CDS regions showed that the SVs-related genes were enriched in peptidyl-threonine phosphorylation and the pathways related to fatty acid elongation (Figure S7). These findings suggested that SV-related genes may play significant functions in the chloroplast membrane system. We subsequently compared the expression levels of the SV-related and normal genes using the RNA-sequencing data and found that a consistently higher portion of the SV genes had significantly low expression levels than the normal genes in all varieties (Figure 4a). This result indicated that genes impacted directly by SVs tend to have reduced expression. We conducted an in-depth analysis to demonstrate the difference in gene expression characteristics between different types of SVs. The results showed that INS exhibited fewer degrees of effects on gene expression than DEL and INV (Figure S8). However, a 288 bp insertion upstream of *low-temperature growth 1* (*LTG1*) was found in six varieties. This insertion was associated with increased expression of *LTG1* in these varieties (Figure 4b). Rice cultivars harbouring dominant *LTG1* are more tolerant to low temperatures than those with the other type of alleles (Lu *et al*., 2014), suggesting a crucial role of *LTG1* in regulating the chilling tolerance of rice. We measured days to heading (DTH) of the 12 varieties plus Nip under a high temperature (HT) of 28–33 °C and low temperature (LT) of 20–25 °C on a short day (10 h light/14 h dark) incubator conditions. When cultivated under LT circumstances compared to HT settings, those cultivars lacking SVs in *LTG1* showed more than 12 days of delayed DTH. Varieties with SVs in *LTG1*, however, showed delays of fewer than 8 days (Figure 4c). Furthermore, plants of two independent CRISPR/Cas9-based knockout lines (*ltg1-cr1* and *ltg1-cr2*) showed delayed DTH under LT conditions compared to the WT (Sasa; Figures 4d,e), confirming that the insertion in the promoter region of *LTG1* lowering sensitivity to low temperature.

Gene CNVs characterization of GJ genome

Gene CNVs are widely distributed in plant genomes and are known to influence crop evolution and domestication, yet resolving gene CNVs is still difficult (Lye and Purugganan, 2019). We took advantage of our high-quality assemblies to assess the CNVs in the temperate GJ genomes and examined their possible effects on important agronomic traits. The whole-genome comparisons revealed a total of 9628 genes with CNVs in the 18 temperate GJ assemblies, which could be involved in the hormone response, disease resistance process, stress response, photosynthesis, etc. (Figure S9). We chose a representative selection of 225 genes with known functions from previous research to better understand the functionality of genes with CNVs in the temperate GJ genomes (Wei *et al*., 2021). Of the 225 genes, 64 (28.4%) genes showed CNVs among the 18 assemblies (Figure 5a). These 64 CNVs were associated with a wide range of functionalities and affect many important rice traits, including 19 genes associated with disease resistance to blast, bacterial blight and brown planthoppers; six genes affecting heading date, four genes involved in hybrid sterility and three genes associated with chilling tolerance, etc. (Table S6). Among 64 CNVs, 30 CNVs could be detected in both Japanese varieties and modern Chinese varieties, indicating that these 30 CNVs originated from Japanese varieties (Table S7). Figure 5b shows an example of *GNP1*-encoding GA20ox1 in gibberellin biosynthesis and affecting grain number per panicle in rice (Wu *et al*., 2016). Notably, the presence of CNVs for *GNP1* has not been previously reported. However, we discovered that *GNP1* contains two to three copies of chromosome 3 in 5 of the 18 assemblies (Kosh, Toyo, LG5, SN265 and YF47), while the other 13 types only had one copy. Several variations that were allelic to the cluster of *Pigm* and possessed CNVs in the 10.34–10.49 Mb area of Nip chromosome 6 were also discovered (Figure 5c). *Pigm* gives long-lasting resistance to the fungus *M. oryzae*, and a large diversity of CNVs at the *Pigm* locus was found among various genotypes. *Pigm* encodes 13 nucleotide-binding leucine-rich repeats (NLR) receptors (R1–R13) and is found in the germplasm GM4 (Deng *et al*., 2017). Notably, there are many tandem copies of *Pigm* (*R2*) genes that exist in ZH11, which is consistent with a previous study (Xie *et al*., 2019). Besides ZH11, DHX2 also contained at least two copies of *R2* (Figure 5c).

Gene CNVs associated with agronomic traits

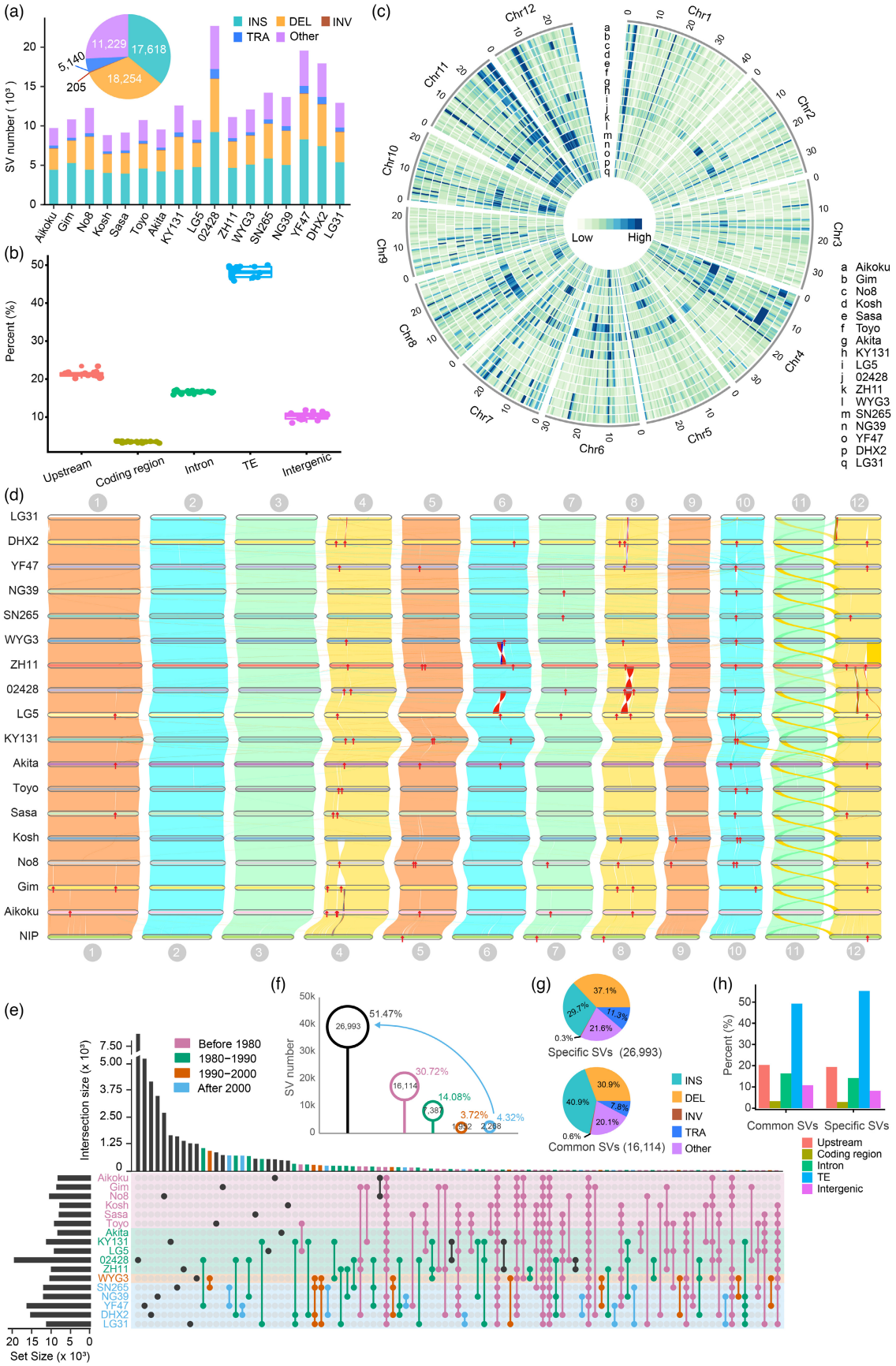
To verify the functional importance of CNVs of *GNP1*, we compared the Japanese temperate GJ variety Toyo harbouring three copies of *GNP1* with ZH11 which has one copy of *GNP1* (Figure 6a). The outcome revealed that *GNP1* in Toyo had considerably greater relative DNA content and expression levels than ZH11 (Figure 6b,c). Moreover, we found that the expression level of *GNP1* was well correlated with the copy number among 18 GJ varieties (Figure S10). To further demonstrate the function of the increased expression of *GNP1*, we investigated the *GNP1* over-expression transgenic lines under the genetic background of ZH11 (CK). The expression levels of *p35S::GNP1-1* and *p35S::GNP1-2* were significantly higher than that of CK (Figure 6d). Compared to CK, the grain number per panicle of both *p35S::GNP1-1* and *p35S::GNP1-2* increased significantly, accompanied by increased plant height (Figure 6e–i). Given that increased *GNP1* expression causes increased grain number per panicle, the CNVs of *GNP1* are likely associated with grain number variation and the duplication of *GNP1* could potentially improve the grain yield of GJ varieties. The SVs identification also proved that the genomes of LG5, Kosh, SN265 and YF47 had multiple copies of *GNP1* (Figure 6j). Among these, Kosh was brought to China and has been widely planted there, Toyo was introduced to China and functioned as a backbone parent for rice breeding for a considerable time and LG5, SN265 and YF47 demonstrated benefits in yield performance. Therefore, we proposed that Toyo and Kosh were chosen from a wide range of Japanese varieties for introduction to China because they had an advantage in grain number per panicle due to numerous copies of *GNP1*, which was a key breeding target in China in the 1980s. And then, the multiple CNVs of *GNP1* were inherited by later-bred varieties to increase the grain number per panicle in northern China. Additionally, it was discovered that the frequency of modern-temperate GJ varieties having multiple copies of *GNP1* had been progressively rising since the 1980s utilizing the resequencing data of the 74 GJ varieties (Table S8 and Figure S11) commonly grown in the main GJ cultivation area of China (Figure 6k). And in the LN province, 60% of types carry multiple copies of *GNP1* (Figure 6k). Among the 74 varieties, the grain number per panicle of varieties harbouring multiple copies of *GNP1* was significantly greater than that of those harbouring a single copy (Figure 6l), suggesting an increased copy number of *GNP1* introduced from the Japanese variety, Kosh and Sasa, may have significantly contributed to the improved productivity of many modern Chinese GJ varieties, particularly those varieties in the LN province (Figure S12).

The selection and introgression of the SVs in GJ

We used a collection primarily made up of 1275 rice genotypes of widely cultivated cultivars and parental hybrid rice lines from China (Li *et al.*, 2020) to perform a genome scan using a cross-population composite likelihood ratio (XP-CLR) and diversity reduction index (DRI) approach to detect putatively selected regions from GJ to gain insight into the impact of the SVs during breeding. The result revealed that a total of 24 878 SVs overlapped with selective sweeps, which involved 4089 genes (Figure 7a), and the distributions of 24 878 SVs were highlighted in Figure 7b. The majority of these selected genes had SVs located in the promoter region, indicating that the SVs mainly affect the expression level of selected genes in GJ breeding (Figure 7a). These genes were enriched in peptidyl-threonine phosphorylation and the pathways involved in fatty acid elongation, according to the results of the Genes and Genomes (KEGG) pathway analysis and GO enrichment analysis (Figure S13). Several SV genes were reported related to yield, grain quality, hybrid sterility and biotic and abiotic stresses, such as *SaF*, *SaM* (Long *et al.*, 2008), *NRT1.1B* (Hu *et al.*, 2015), *NAL1* (Fujita *et al.*, 2013) and *Xa21*(Song *et al.*, 1995), and were divergently selected during XI and GJ breeding (Figure 7b, Table S9). Interestingly, a substantial selective sweep signal between the GJ and XI subspecies was visible in the massive inversion (4.53 Mb) between the centromere of chromosome 6 of 02428 and ZH1 (Figure 3d). Given that the allele frequency of this inversion in GJ is just 4.35%, compared to 76% in XI and 20% in the wild, an introgression event might have occurred during GJ breeding.

We then compared the 18 temperate GJ assemblies to XI, cA, cB and wild rice to trace the origin of the SVs. The result showed that the origin of 51 419 SVs (98.0% of 52 446 SVs) could be traced (Figure 8a). For example, an SV at the intron of *Chalk5* was detected in wild rice, cA, cB and 23 out of 24 XI varieties, but was found in only two temperate GJ varieties which were bred after the 2000s (02428 and YF47; Figure 8b). These data indicated that the SVs in *Chalk5* were the product of an XI to GJ introgression. YF47 is a high-yielding GJ cultivar, although it has a moderate grain appearance quality, particularly in chalkiness. Given that *Chalk5* encodes a vacuolar H^+ translocation pyrophosphatase influencing rice grain chalkiness (Li *et al.*, 2014), it appears that the introgression of SV at the

Figure 3 Structural variant (SV) characterization of GJ rice genomes. (a) The number of SVs in each assembly includes five types of SV. DEL, deletion; INS, insertion; INV, inversion; TRA, translocation. Other types included NOTAL (not aligned region), TDM (tandem repeat) and HDR (highly diverged regions). (b) The percentage of the detected SVs overlapped with different genomic regions in the 17 *Geng* assemblies. The mean percentage values of elements (2 kb upstream, coding region, intron, transposable elements and intergenic regions) are 21.4%, 3.5%, 16.6%, 48.4% and 10.1% respectively. (c) The SV distribution among 17 assemblies relative to Nip. A circular show of the detected SVs among the 17 GJ genomes with a sliding window size of 500 kb. (d) The landscape of some large-size SVs among 17 GJ varieties and Nip. Red arrows direct the locus of SVs with insertions and deletions. Dark-coloured bands display examples of large structural variations in inversion and translocation. (e) Presence of SVs among different breeding stages. SVs only existed in one group and were defined as specific SVs; the SVs not only existed in the first batch (shown in pink) but are also inherited into at least one group that was defined as common SVs. (f) Summary of transmitted SVs during past breeding. The values in differently coloured circles represent the number of SVs for corresponding SVs-deriving sets including black (group-specific SVs), purple (before 1980 inherited by following released groups), brown (1980–1990 inherited by following released groups), green (1990–2000 group inherited by following released groups) and blue (after 2000 inherited by internal varieties). (g) The pie plot shows the proportion changes in SV types between specific SVs and common SVs. The insertion (INS) and inversion (INV) rates are increased in common SVs. (h) Profiles of SV locus among different genetic elements. The proportion of SVs in the intergenic region is increased in common SVs.



Chalk5 locus affected the chalkiness-related traits in YF47. We hypothesized that the SVs at *Chalk5* were introduced to YF47 from XI along with certain superior alleles during breeding selection due to genetic drag because grain chalkiness is a highly undesirable characteristic. Then, we scanned the 100 kb surrounding *Chalk5* and discovered *GS5*, a locus that modulates grain size (Li *et al*., 2011; Figure S14). The XP-CLR analysis result revealed that *GS5* and *Chalk5* are located in a single selective sweep (Figure 8c), indicating that the superior *GS5* allele from XI for larger grains and the inferior XI allele at *Chalk5* for more severe chalkiness were introduced together during the breeding attempts pursuing high yielding in north China. Fortunately, the linkage disequilibrium (LD) block analysis using the SVs of 58 genome assemblies (Figure 8d) and a collection comprised of 1275 rice accessions (Li *et al*., 2020; Figure S15) inferred that *GS5* and *Chalk5* were not tightly linked, suggesting that the linkage between *GS5* and *Chalk5* could be broken by the cross. We accelerated this process by knocking out the inferior XI allele of *Chalk5* in YF47 using the CRISPR/Cas9 gene-editing technology, as clearly demonstrated in two independent T_2 transgenic lines (Figure 8e). The transgenic lines exhibited identical plant morphology, grain size and panicle length as expected, but considerably improved chalkiness-related parameters as compared to YF47 (Figure 8f–i). Except for the transgenic plants' chalkiness level and head rice ratio, which were greatly enhanced over YF47, there was no change in yield components between the transgenic lines and YF47 (Figure S16).

Discussion

The temperate GJ rice varieties, a significant subpopulation of *O. sativa*, are currently cultivated on more than 14.78 million ha[2] of rice fields globally, particularly in northeast China, Japan and Korea, making a significant contribution to global food security (Tang and Chen, 2021). The modern-temperate GJ rice varieties in China are also highly productive with an average yield exceeding that of the XI hybrid rice cultivars in China from extensive breeding efforts to improve productivity during the past century (Fei *et al*., 2020). Thus, there is growing curiosity over how the temperate GJ population's genetic composition has changed due to previous breeding operations. Due to widespread genome structural diversity, the short-read sequencing data cannot capture the entire breeding signature. By focusing on SVs among the high-quality genomes of a carefully chosen selection of 18 temperate GJ varieties, we attempted to answer this question in this study. Our findings thus shed vital light on SVs and how past breeding efforts had affected how genetic diversity was organized within the temperate GJ population. The most important result of this study was the discovery of CNVs for 9628 genes of diverse functions in only 18 temperate GJ genomes, suggesting the presence of rich gene CNVs as an important feature of the total genetic diversity in the populations of *O. sativa*, which remains to be fully characterized in future using the long-read sequencing technologies. The findings also revealed genomic and genetic evidence of SV originating from intricate breeding histories intended to increase yield potential, stress tolerance and quality.

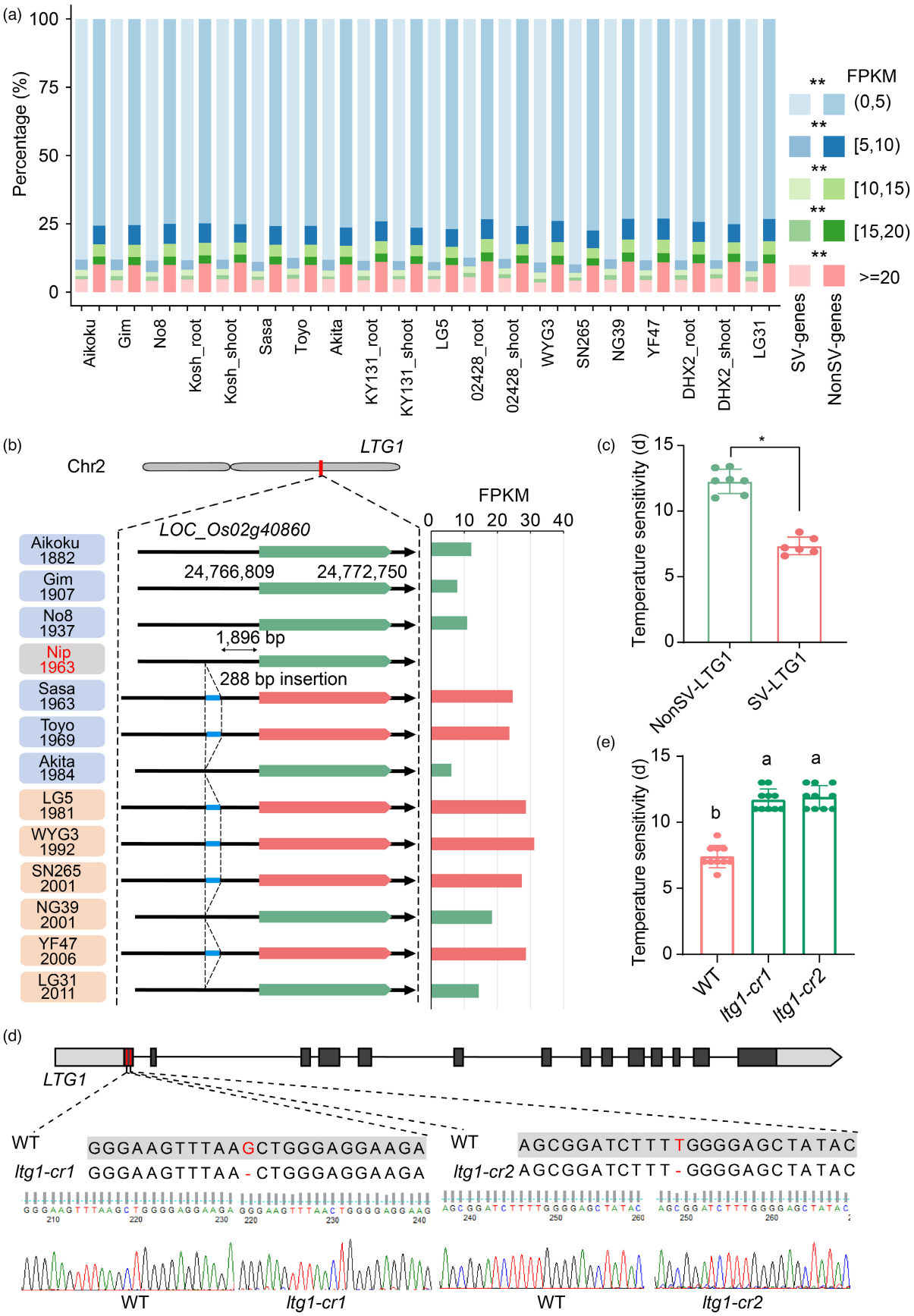
Importantly, because it is still difficult to resolve gene CNVs, few CNVs were found and very few of them were used in breeding practices. Our high-quality assembly could aid breeders and researchers in choosing superior backbone parents with beneficial CNVs for rice breeding. This would greatly speed up the breeding process. Additionally, attempts to introduce advantageous features from wild and XI subspecies to create GJ cultivars with superior yield and quality performance will be motivated by the information on SVs introgression offered in the current study.

Methods

Plant materials

We specifically selected 12 varieties that have greatly contributed to rice breeding research and massive planting, including six Japanese and six Chinese temperate GJ varieties. Gimbozu (Gim) was derived from the Aikoku, as the lone plant that remained erect after a storm caused all of the other plants in the field to lodge. The popularity of Aikoku and Gim cultivars with Japanese farmers in the first half of the 20th century led to the use of these cultivars throughout the country and the breeding of many distinct but closely related strains and landraces (Naito *et al*., 2006). Norin8 (No8) was derived from Gim, and a lot of famous varieties such as Koshihikari (Kosh), Akitakomachi (Akita) and Zhonghua11 (ZH11) were derived from No8. Sasanishiki (Sasa), developed in 1963, which was cultivated in the Tohoku region in Japan and gained status as high-quality rice. In 1990, 207 439 ha were devoted to its cultivation (11.3% of the total paddy field in Japan) and it became the second most popular rice cultivar (Nagano *et al*., 2013). Toyonishiki (Toyo) was introduced to China and served as a backbone parent for a long time. Liaogeng5 (LG5) was the first related commercial variety with erect panicle architecture. Shennong265 (SN265) was the first released 'super rice variety' by the Chinese Ministry of Agriculture. The total promotion area of both Longgeng31 (LG31) and Wuyugeng3 (WYG3) exceeds 100 million mu (6.67 million hectares). Nangeng39 (NG39) and Yanfeng47 (YF47) were in large-scale promotion in Jiangsu Province and Liaoning Province respectively. The release years of these varieties ranged from 1882 to 2011, spanning an interval of approximately 130 years. In addition, 74 temperate GJ varieties that have been widely cultivated in China since the 1980s were selected for the present study. These varieties were cultivated in fields, controlled greenhouses and incubators at the Rice Research Institute of Shenyang Agricultural University (LN N41°, E123°). The cultivation methods and field management were described in our previous report (Li *et al*., 2018a). We harvested a total of 20 plants from the middle

Figure 4 SVs impact gene expression profiles. (a) The proportion of SV genes and non-SV genes were associated significantly ($P < 0.01$) with altered expression of related genes. The differences in per cent values between SV genes and non-SV genes were assessed using Student's *t*-tests for five continuous expression ranges respectively. *Indicated a significance level at $P < 0.01$. (b) The SVs upstream of *LTG1* cause expression variants among GJ varieties. The blue line indicates the 288 bp insertion in the promoter of *LTG1*. (c) The temperature sensitivity (the difference in days to heading between plants under high and low temperatures) of non-SV-*LTG1* and SV-*LTG1* varieties. Data are mean ± SEM ($n = 7$ for non-SV-*LTG1*, and $n = 6$ for SV-*LTG1*), and *indicates significance at the $P < 0.05$ level. (d) Diagram and sequence of *LTG1* CRISPR knockout lines (*ltg1-cr*1 and *ltg1-cr2*). The red line indicates the position of the sgRNA target site. (e) The temperature sensitivity of WT and CRISPR knockout lines (*ltg1-cr*1 and *ltg1-cr2*). Data are mean ± SEM ($n = 10$), and different letters indicate significant differences ($P < 0.05$, one-way ANOVA, Tukey's HSD test).


(a)
Percentage (%)
100
75
50
25
0
FPKM
(0,5)
[5,10)
[10,15)
[15,20)
>=20
**
Aikoku
Gim
No8
Kosh_root
Kosh_shoot
Sasa
Toyo
Akita
KY131_root
KY131_shoot
LG5
02428_root
02428_shoot
WYG3
SN265
NG39
YF47
DHX2_root
DHX2_shoot
LG31
SV-genes
NonSV-genes
(b)
Chr2
LTG1
FPKM
0 10 20 30 40
LOC_Os02g40860
24,766,809
24,772,750
1,896 bp
288 bp insertion
Aikoku 1882
Gim 1907
No8 1937
Nip 1963
Sasa 1963
Toyo 1969
Akita 1984
LG5 1981
WYG3 1992
SN265 2001
NG39 2001
YF47 2006
LG31 2011
(c)
Temperature sensitivity (d)
15
10
5
0
*
NonSV-LTG1
SV-LTG1
(e)
Temperature sensitivity (d)
15
10
5
0
a
a
b
WT
ltg1-cr1
ltg1-cr2
(d)
LTG1
WT
GGGAAGTTTAAGCTGGGAGGAAGA
ltg1-cr1
GGGAAGTTTAA-CTGGGAGGAAGA
WT
AGCGGATCTTTTGGGGAGCTATAC
ltg1-cr2
AGCGGATCTTT-GGGGAGCTATAC
WT
ltg1-cr1
WT
ltg1-cr2

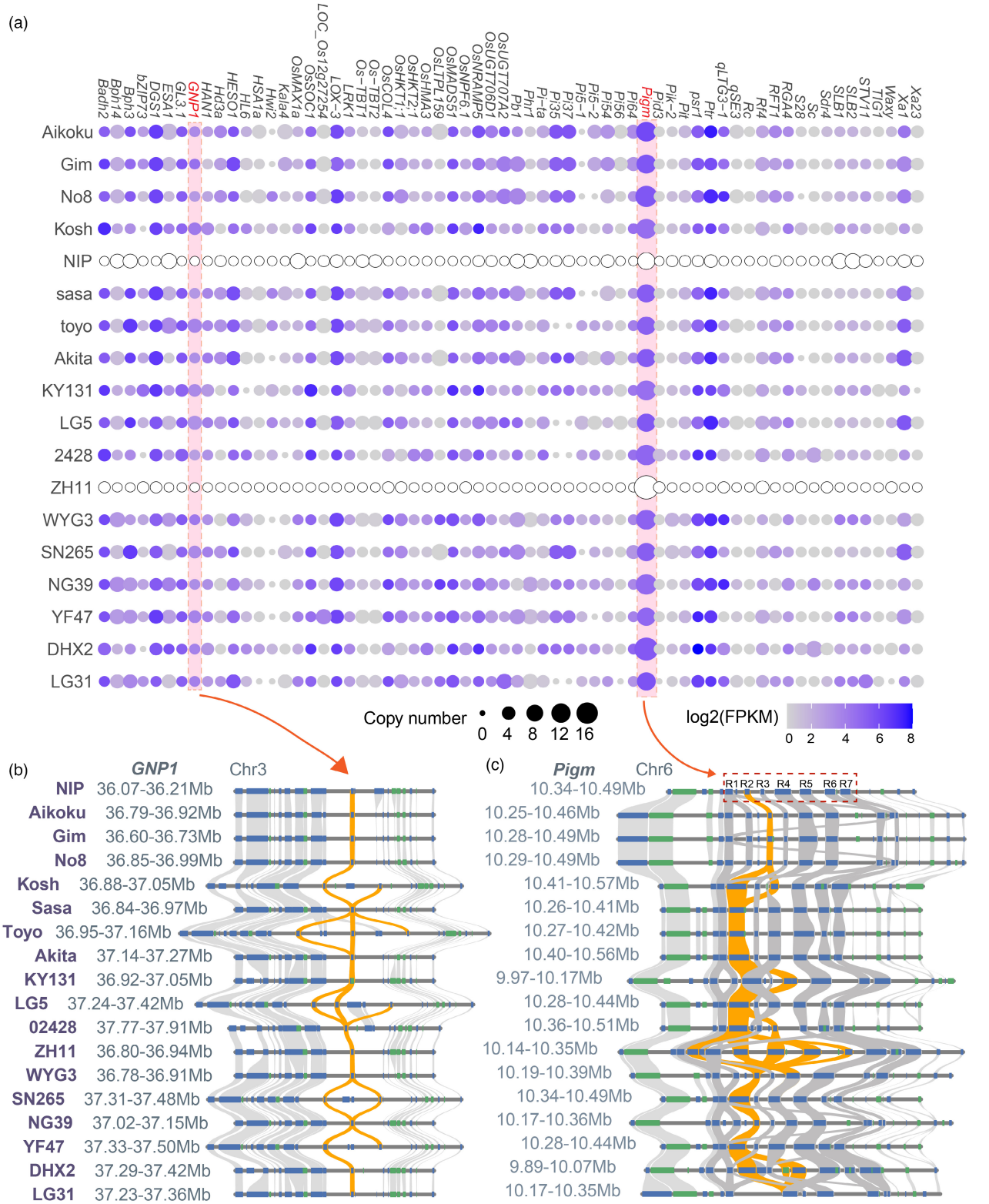


Figure 5 Characteristics of gene CNVs related to important agronomic traits. (a) The functional genes with CNV mutations. Circle size represents the number of gene copies potentially generated by a tandem duplicated mechanism. Colours from light to dark imply the global expression level [log2 (FPKM)] of genes ranging from low to high. (b) Local syntenic relation of *GNP1* implying breeding selection of different CNVs among 18 GJ varieties. The blue rectangle represents the forward strand gene in the chromosome, and the green rectangle means the reverse strand gene. Orange-linked bands highlight homologue gene pairs having different copy numbers in this region. (c) Local syntenic relation of *Pigm* implying breeding selection of different copy number variation among 18 GJ varieties. The colours are the same as (b). The red dashed rectangle represents the *Pigm* cluster (*R1–R13*) in Nip. Dark grey bands linked homologue R genes among assemblies. The orange band tracks the evolutionary pattern of R2 (*LOC_Os06g17900*) along with released GJ varieties.

rows 45 days after heading for each line. The measurement of the agronomic traits was conducted as described in our previous study (Li *et al.*, 2018b). Due to the low temperature in October, certain varieties developed in Jiangsu province could not mature in Shenyang's natural environment. As a result, we moved the varieties that did not mature from the field to the greenhouse in October to ensure their maturity. In temperature treatment, plants were maintained in the incubator at various temperatures and humidity levels, including 70% and 300 μmol/m^2/s fluorescent lamps. All of the samples were analysed with two biological replicates.

Illumina and nanopore sequencing

Illumina sequencing and nanopore sequencing were performed at BIOMARKER (Beijing, China). High-molecular-weight DNA was extracted from 3-week-old seedlings. One hundred nanograms (ng) of genomic DNA were used to prepare the library. Briefly, gDNA was sonicated to a fragment size of 500 bp by an ultrasonicator and the library was prepared by using an NEB Ultra DNA library prep kit (NEB, MA, USA) according to the manufacturer's instructions. Sequencing of the library was performed on a HiSeq X Ten system using the run configuration 2 × 350 bp. The sequencing reads were generated from the paired end. The library was trimmed using fastq_quality_trimmer in the FASTX Toolkit (ver. 0.0.11) with default parameters, and all trimmed reads were assembled into scaffolds using ALLPATHS-LG (ver. 44849). The assembly ambiguity information was removed using the efasta2fasta script. The gaps in the resulting scaffolds were filled using GapFiller (ver. 2.1.1) with default parameters. Nanopore sequencing of 2 μg of gDNA was repaired using NEB Next FFPE DNA Repair Mix kit (M6630) and subsequently processed using the ONT Template prep kit (SQK-LSK109, UK) according to the manufacturer's instructions. The large segments library was premixed with loading beads and then pipetted into a previously used and washed R9 flow cell. According to the manufacturer's instructions (EXP-FLP001.PRO.6, UK), the library was sequenced on the ONT PromethION platform with the corresponding R9 cell and ONT-sequencing reagents kit. We extracted the genomic DNA of 74 cultivars from fresh frozen leaves using the CTAB method. According to the manufacturer's instructions, the samples were sequenced on the Illumina HiSeq 2500, and the sequencing libraries were constructed. We aligned the sequencing reads to the genome of Nipponbare using BWA software (Li and Durbin, 2009). In total, 1638.72 Gb of clean data were generated across all 74 cultivars, with approximately 53-fold depth for each cultivar.

Twelve time-ordering GJ genomes assembly

De novo genome assembly was performed using a combination of three strategies: initial WTDBG2 (https://github.com/ruanjue/wtdbg2) assembly, followed by Smartdenovo (https://github.com/ruanjue/smartdenovo) assembly. We polished the assemblies using three rounds of racon software (https://github.com/marbl/canu, v1.; Vaser *et al.*, 2017), followed by three rounds of polishing with Pilon software (https://github.com/PacificBiosciences/FALCON, v0.3.0; Walker *et al.*, 2014). The assembled results were assessed by evaluation of the ratio to the Illumina sequencing reads and the evaluation of BUSCO integrity. BUSCO v 2.0 was conducted against the metazoan database to validate the genome completeness and gene set completeness of the draft genome sequences.

Chromosome construction using Hi-C links

About 1.5 g of 3-week-old seedlings were used for the Hi-C library construction. Hi-C libraries were created using a previously described method (Mascher *et al.*, 2017). Libraries (based on *Hind*III) with fragments ranging from 300 to 700 bp size were constructed and sequenced on the Illumina X-TEN platform (Illumina). Mapping of Hi-C reads and assignment to restriction fragments were performed as described previously (Burton *et al.*, 2013). We performed duplicate removal, sorting and quality evaluation using HiC-Pro v2.10.0 (Servant *et al.*, 2015) with the command of 'mapped_2hic_fragments.py -v -S -s 100 -l 1000 -a -f -r -o'. The raw counts of the Hi-C links were aggregated in 100-kb bins and normalized separately for intra- and inter-chromosomal contacts using HiC-Pro. The corrected contigs were assembled into 12 chromosome-level scaffolds by LACHESIS (Burton *et al.*, 2013). Adjacent contigs were linked together by filling the gap with 'N'. Finally, 12 high-quality pseudochromosome-level genomes were built for representative GJ accessions.

Gene annotation

The RNAs of 12 temperate GJ varieties were isolated from the mixed tissues (fresh leaves, roots and culm) following the manufacturer's protocol provided in the TaKaRa MiniBEST Universal RNA Extraction kit. We then performed the sequencing on the Illumina HiSeq 2500 platform according to the manufacturer's instructions. RNA-seq data (total of 8 Gb) for each variety was obtained. We used PILER-DF v2.4, RepeatScout v1.0.5, LTR_FINDER v1.05 and MITE-Hunter to construct a primary repeat sequence database based on *ab initio* prediction theory and structural prediction (Edgar and Myers, 2005; Han and Wessler, 2010; Price *et al.*, 2005; Xu and Wang, 2007). The primary database based on PASTE Classifier was classified and then combined with the Repbase database. To conduct the final prediction, we used the combined data to form the final repeat sequence database using Repeat Masker v4.0.6 (Jurka *et al.*, 2005; Tarailograovac and Chen, 2004; Wicker *et al.*, 2009). To perform protein-coding gene prediction, the repeat elements from the genome assembly were masked and excluded. Then, we performed the gene annotation through three prediction steps: (i) *ab initio* prediction by Augustus v2.4, Genscan, GlimmerHMM v3.0.4, GeneID v1.4 and SNAP (version 2006-07-28); (ii) homologous species prediction based on *Oryza sativa*, *Zea mays*, *Arabidopsis thaliana*, *Sorghum bicolor* and *Setaria italica* using GeMoMa v1.3.1; and (iii) unigene prediction based on full-length transcriptome data assembly with no reference genome conducted through PASA v2.0.2 (Blanco *et al.*, 2007; Campbell *et al.*, 2006; Jens *et al.*, 2016; Korf, 2004; Majoros *et al.*, 2004; Stanke and Waack, 2003). We integrated the three predictions through EVM v1.1.1, and performed final modifications by PASA v2.0.2 (Haas *et al.*, 2008). We subsequently identified non-coding RNAs (microRNAs, rRNAs and tRNAs) using different strategies based on their unique structural features. We used the Rfam, miRBase and tRNAscan-SE v1.3.1 databases to predict rRNA, microRNA and tRNA respectively (Griffiths-Jones *et al.*, 2005; Nawrocki and Eddy, 2013). The pseudogenes were predicted through scanning for homologous genes and excluding genuine genes using GenBlastA v1.0.4 (Rong *et al.*, 2009). The candidate genes with frameshift mutations and premature stop codons were selected as the final pseudogene predictions by GeneWise v2.4.1 (Birney *et al.*, 2004).

To annotate gene functions, the predicted genes were blasted to the GO and KEGG databases by BLAST v2.2.31 (-evalue 1e-5; Altschul *et al.*, 1990; Boeckmann *et al.*, 2003; Marchlerbauer *et al.*, 2011; Ogata *et al.*, 2000; Tatusov *et al.*, 2001). In addition, the motifs were annotated according to the sequence alignments with the HAMAP, PRINTS, Pfam, ProDom, SUPERFAMILY, TIGRFAMs, CATH-Gene3D, PANTHER and PIRSF databases by InterProScan software (Zdobnov and Apweiler, 2001).

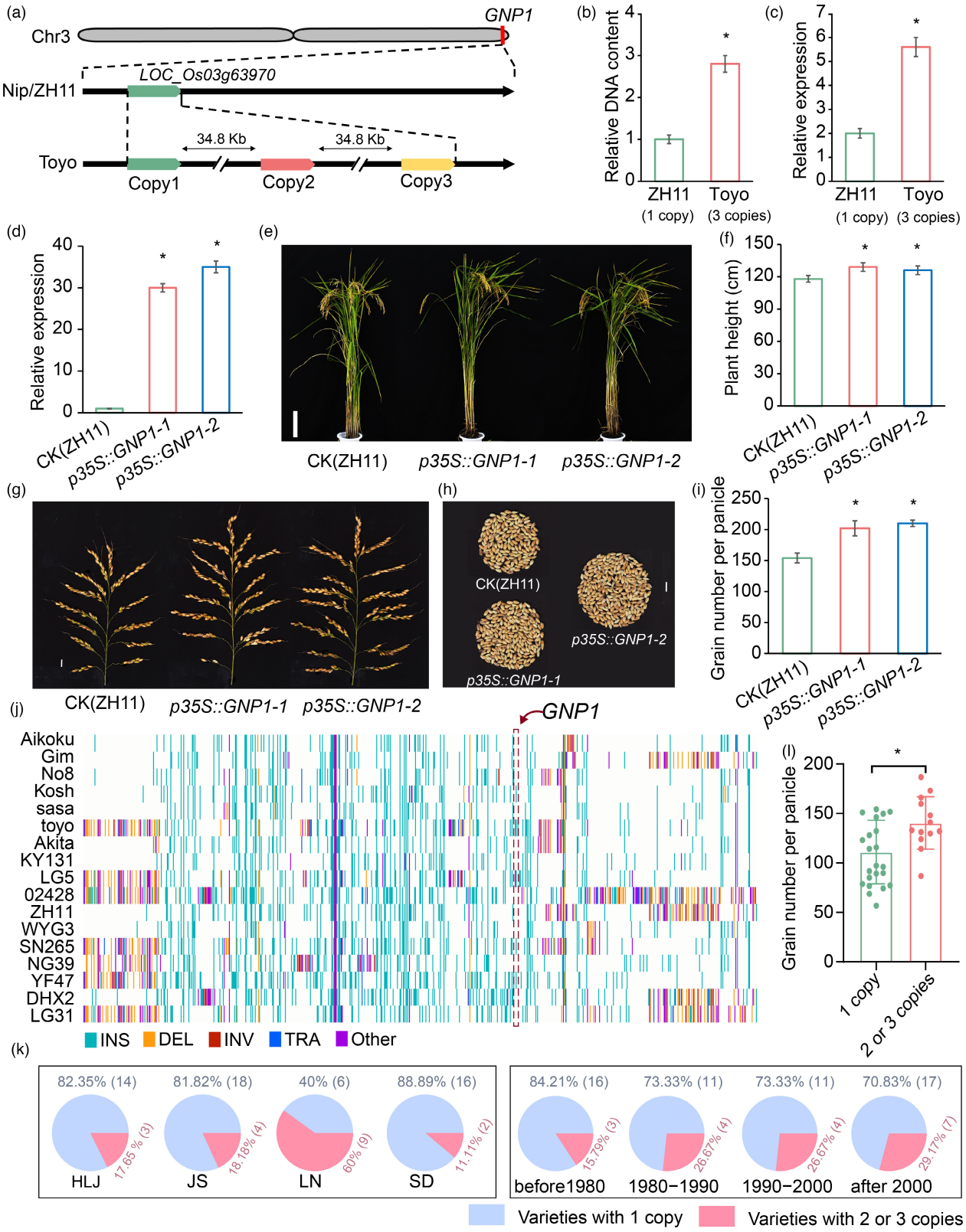


Figure 6 Gene copy number variants (CNVs) are associated with variations in production. (a) Schematic illustrating a single copy of *GNP1* in Nip and ZH11 and three copies of *GNP1* in Toyo. (b) DNA qPCR validation of the three *GNP1* copies. *Indicates significance at the $P < 0.05$ level. (c) The expression of *GNP1* in Toyo with three copies is significantly higher than in Nip with a single copy of *GNP1*. *Indicates significance at the $P < 0.05$ level. (d) The expression level of *GNP1* in Zh11 (CK) and two independent over-expression transgenic lines. *Indicates significance at the $P < 0.05$ level. (e) The Zh11 (CK) plant architecture and two independent over-expression transgenic lines. Bar = 20 cm. (f) The Zh11 (CK) plant height and two independent over-expression transgenic lines. Data are mean ± SEM ($n = 10$), and *indicates significance at the $P < 0.05$ level. (g) Zh11 (CK) panicle size and two independent over-expression transgenic lines. Bar = 1 cm. (h) The grains are derived from one Zh11 (CK) panicle and two independent over-expression transgenic lines. Bar = 1 cm. (i) The grain number per panicle of Zh11 (CK) and two independent over-expression transgenic lines. Data are mean ± SEM ($n = 10$), and *indicates significance at the $P < 0.05$ level. (j) The SVs around *GNP1* in the 18 assemblies. (k) The distribution of multiple copies of *GNP1* among 74 GJ varieties. (l) The grain number per panicle of varieties harbouring multiple *GNP1* copies and varieties harbouring a single copy of *GNP1*. *Indicates significance at the $P < 0.05$ level.

Genome-wide syntenic relationship analysis

The longest transcript was selected to represent the corresponding protein-coding gene. The all-to-all BLASTP program was used to identify homologous pairs with parameters 'E_VALUE=1e-05, MAX GAPS=25, and MATCH_SIZE=5' (Camacho *et al.*, 2009). Synteny blocks between each pair of rice varieties were called using McScanX v1.1 with the default parameter (Wang *et al.*, 2012). Only synteny blocks having more than five homologous gene pairs were considered conserved syntenic blocks. The large gene-based structure variations including inversions, deletions, insertions and translocations were manually highlighted by dark colours.

Structural variation identification

We collected 55 *de novo* genome assemblies based on the long-read sequencing technologies which had been published so far (Qin *et al.*, 2021; Stein *et al.*, 2018; Zhang *et al.*, 2022). After discarding duplicated assemblies, 45 existing assemblies and 12 assemblies from the present study were aligned to the genome of Nip (MSU7) using MUMmer (v 4.0; Marçais *et al.*, 2018). The resulting filtered delta files were used to detect structural variations using the SyRI pipeline with default parameters (Goel *et al.*, 2019). The SV identification and classification detail were shown in the previous study (Qin *et al.*, 2021). We defined SV genes if it has an SV locus in 2 kb upstream or exon. Other genes were considered non-SV genes.

Inference of gene CNVs and trait variations

For each rice of 18 varieties (including NIP), we detected tandem duplicated genes by DupGen_finder with the default parameter (Qiao *et al.*, 2019). Compared to Nip, if one duplicated locus (including tandem and proximal duplicate types) has a different copy number in at least 1 of the other 17 GJ varieties, this locus was defined as gene-CNV locus. We obtained 225 function-known genes (associated with traits) of rice from the previous study (Wei *et al.*, 2021). We identified 9268 gene-CNVs loci. Of these, 64 gene CNVs are potentially associated with important traits. We displayed these CNVs and expression levels according to the time-ordering breeding varieties ranging from 1882 to 2011. The expression level of CNVs was tested using qPCR. The DNA was extracted from the fresh leaves of varieties using the DNeasy Plant Mini kit (Qiagen, Duesseldorf, Germany). The RNA extraction was carried out using TRIzol (Invitrogen, CA, USA). qPCR [2 × SYBR Green qPCR Master (Mimake)] is used to verify DNA and RNA expression content for genes. ACTIN was used as the internal control.

SVs-based phylogenetic and population structure analysis

After filtering the low-quality structure variations (SVs) using metrics of 'minor allele frequency (MAF) >0.05' from the raw SVs dataset, we retained 156 319 high-confidence SVs for further analysis. To investigate the phylogenetic relationships of the 58 rice samples, including five wild, 24 GJ, two cA, two cB and 25 XI accessions, a maximum-likelihood phylogenetic tree was constructed using IQ-TREE v1.6.11 (Minh *et al.*, 2020) with the optimal models and standard bootstrap for 1000 replicates. The resulting tree was visualized with FigTree v1.4.4 (Rambaut, 2009) with an outgroup of wild rice. The population genetic structure was examined using the program ADMIXTURE (v1.23; Alexander *et al.*, 2009) with K values (the putative number of populations) from 2 to 10. The $K = 4$–6 values were chosen to display the genetic admixtures of rice populations.

Principal component and linkage disequilibrium analysis

Principal component analysis (PCA) was carried out using the smartPCA program from the EIGENSOFT package v.6.0.1 (https://github.com/DReichLab/EIG) based on 156 319 SVs. The first three principal components were used to separate the cultivar and wild progenitor samples. Linkage disequilibrium between pairs of SVs and SNPson each chromosome, respectively, was assessed as the correlation coefficient (r^2) using PLINK v1.07 (Purcell *et al.*, 2007). The haplotype blocks were plotted for interesting regions using Haploview v4.2 (Barrett *et al.*, 2005).

Selective-sweep analysis

The selective sweeps potentially related to XI and GJ divergence were investigated using DRI ($\pi_{XI} \times \pi_{GJ}^{-1}$), DRI ($\pi_{GJ} \times \pi_{XI}^{-1}$) and XP-CLR following methods in the previous studies (He *et al.*, 2019; Li *et al.*, 2021). The XP-CLR scores between GJ and XI cultivars were calculated using the XP-CLR package (Chen *et al.*, 2010) with sliding windows of 100 kb that had a 10-kb step between adjacent windows. The nucleotide diversity (π) was calculated by using PopGenome package v2.2.4 (Pfeifer *et al.*, 2014). The SNP list of 1275 rice accessions was used to identify selective-sweep regions for GJ and XI populations (Li *et al.*, 2020). The top 5% outliers of regions were assigned as candidate selection sweep signals. Adjacent signals that overlapped with each other were merged into a single selective sweep and SV genes in these regions with more than 30% allele frequency divergence between XI and GJ populations were considered candidate selective SV genes.

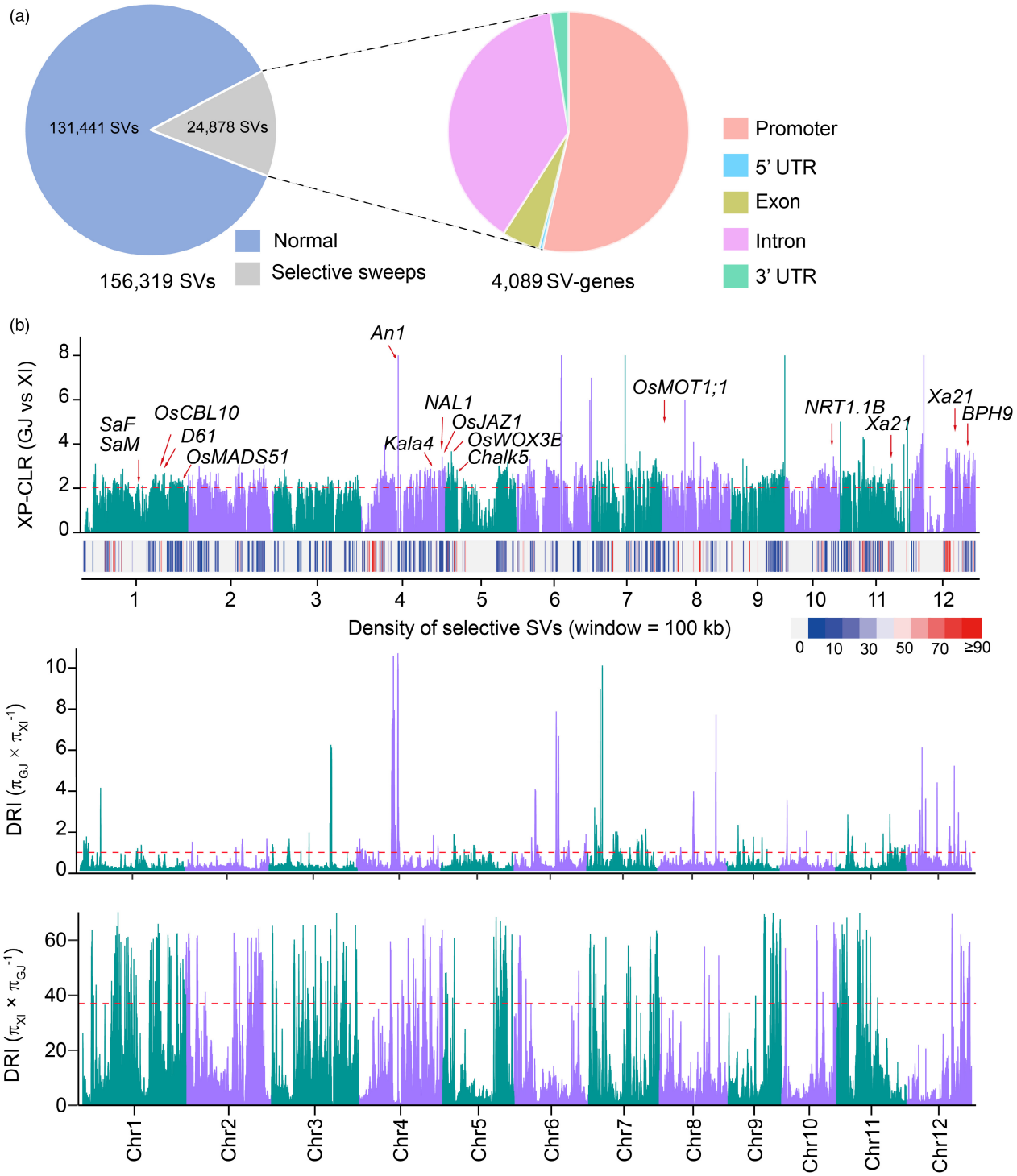


Figure 7 The selection of the SVs in GJ. (a) A total of 24 878 SVs were overlapped with select sweeps, which involved 4089 genes. (b) Selection sweeps uncovered by joint cross-population composite likelihood ratio (XP-CLR) and diversity reduction index (DRI) approaches for the GJ population. Genes or QTLs related to yield, grain quality, hybrid sterility and biotic and abiotic stresses in the selection sweeps are indicated (Table S9).

Transcriptome analysis

Besides 12 pooling transcriptome datasets from corresponding *de novo* rice varieties, we also obtained shoot and root RNAseq datasets of other four temperate GJ (02428, DHX2, KY131 and Kosh) from Genome Sequence Archive (https://ngdc.cncb.ac.cn/gsa/browse/CRA004087). The RNAseq reads were processed to obtain clean reads using the FASTP v0.21.1 (Chen *et al.*, 2018). The clean reads were mapped to the NIP reference genome by HISAT2 v 2.0.4 (Kim *et al.*, 2015). The gene expression level was calculated by StringTie v2.1.4 using default parameters (Pertea *et al.*, 2015). The gene expression level was normalized by reads

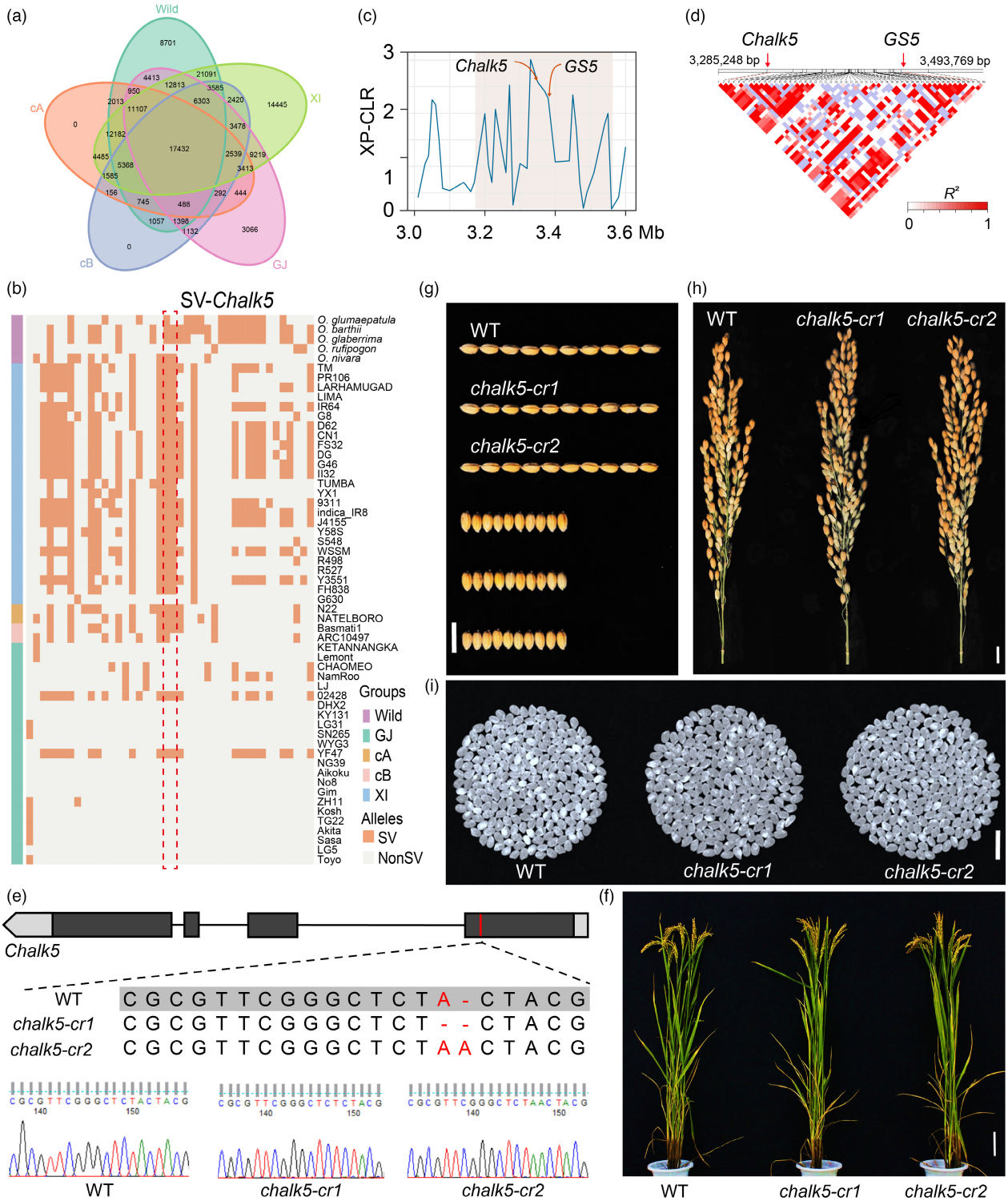


Figure 8 The introgression of the SVs and inferior allele editing breeding. (a) Venn diagrams showing the number of the traced SVs from wild, cA, cB, *japonica/geng* (GJ) and *indica/xian* (XI). (b) A heat map showing the introgression of SVs around *Chalk5*. (c) The enlarged image of XP-CLR score around *Chalk5* and *GS5*. (d) Linkage disequilibrium plot for SVs. (e) Diagram and sequence of *Chalk5* CRISPR knockout lines (*chalk5-cr*1 and *chalk5-cr2*). The red line indicates the position of the sgRNA target site. (f) The plant architecture of WT and *chalk5-cr*1 and *chalk5-cr2*. Bar = 10 cm. (g) The grain shape of WT and *chalk5-cr*1 and *chalk5-cr2*. Bar = 1 cm. (h) The panicle of WT and *chalk5-cr*1 and *chalk5-cr2*. Bar = 1 cm. (i) The chalkiness trait of WT and *chalk5-cr*1 and *chalk5-cr2*. Bar = 1 cm.

per million per kilo bases (FPKM). Genes with expression levels greater than 0.1 FPKM were considered expressed genes. To assess the association of SVs and gene expression, we computed the frequency of expressed genes on different expression level ranges (FPKM: [0–5], [5–10], [10–15], [15–20] and [>20]) for the SV genes and non-SV genes. Then, we tested the difference of

proportion values for each range of gene expression levels between SV genes and non-SV genes among all transcriptome experiments using the Student's *t*-tests function in R.

Vector construction and plant transformation

To conduct the CRISPR/Cas9 gene editing, the vector construction was performed as described by Li *et al*. (2017). We designed the specific single-guide RNA (sgRNA) sequences targeting the *Chalk5* and *LTG1* genes. The specificity of the targeting sequence was confirmed by BLAST searching against the Nip genome (Hsu *et al*., 2013). The rice transformation was conducted as described elsewhere (Nishimura *et al*., 2006). We extracted the genomic DNA from transformants, and the genomic DNA was sequenced for mutant identification. The PCR products (200–500 bp) were sequenced and identified using the degenerate sequence decoding method (Ma *et al*., 2015).

Acknowledgements

We thank Hongxuan Lin at the Shanghai Institute for Biological Sciences of the Chinese Academic of Sciences for providing the seeds of *GNP1* over-expression transgenic plants. We thank Peng Qin at Sichuan Agricultural University for providing the seeds of 02428 and DHX2. Funding for this work was provided by the National Natural Science Foundation of China (32071982).

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Y.W., S.W., Y.Z., R.J., W.H. and H.Z. conducted the SV and CNV identification. F.L. and L.W. performed the CRISPR/Cas9 gene editing. F.L., L.W., Q.S. and N.X. conducted the phenotypic analysis. L.F., H.C., Z.Y., J.L., K.J. and X.W. collected samples for RNA sequencing and conducted expression validation. Q.X., A.G., F.Z., Z.X and W.C. designed and supervised this project. Q.X. wrote the manuscript with input from F.Z., J.W., J.S., L.T. and H.X. All authors read, edited and approved the manuscript.

Data availability statement

All data released with this study can be freely used. Genome sequencing data of 12 GJ varieties, 74 GJ varieties and 12 assemblies in this study have been deposited at the Genome Warehouse (GWH; https://bigd.big.ac.cn/gwh/) under PRJCA011169. The seeds of GJ varieties are available from the corresponding author upon reasonable request.

References

Alexander, D.H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and genomewise. *Genome Res.* **14**, 988–995.

Blanco, E., Parra, G. and Guigó, R. (2007) Using geneid to identify genes. *Curr. Protoc. Bioinform.* **18**, 4–3.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J. *et al*. (2003) The Swiss-Prot knowledgebase and its supplement TREMBL in 2003. *Nucleic Acid Res.* **31**, 365–370.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421.

Campbell, M.A., Haas, B.J., Hamilton, J.P., Mount, S.M. and Buell, C.R. (2006) Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics*, **7**, 327.

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

Chen, S., Zhou, Y., Chen, Y. and Jia, G. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Chen, Z., Li, X., Lu, H., Gao, Q., Du, H., Peng, H., Qin, P. *et al*. (2020) Genomic atlases of introgression and differentiation reveal breeding footprints in Chinese cultivated rice. *J. Genet. Genomics*, **47**, 637–649.

Cook, D.E., Lee, T.G., Guo, X., Melito, S., Wang, K., Bayless, A.M., Wang, J. *et al*. (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*, **338**, 1206–1209.

Crow, T., Ta, J., Nojoomi, S., Aguilar-Rangel, M.R., Torres Rodríguez, J.V., Gates, D., Rellán-Álvarez, R. *et al*. (2020) Gene regulatory effects of a large chromosomal inversion in highland maize. *PLoS Genet.* **16**, e1009213.

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

Deng, Y., Zhai, K., Xie, Z., Yang, D., Zhu, X., Liu, J., Wang, X. *et al*. (2017) Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. *Science*, **355**, 962–965.

Du, H., Ying, Y., Ma, Y., Qiang, G., Cao, Y., Zhuo, C., Ma, B. *et al*. (2017) Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* **8**, 15324.

Edgar, R.C. and Myers, E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics*, **21**(Suppl 1), i152.

Fei, C., Xu, Q., Xu, Z. and Chen, W. (2020) Effect of rice breeding process on improvement of yield and quality in China. *Rice Sci.* **27**, 363–367.

Fujita, D., Trijatmiko, K.R., Tagle, A.G., Sapasap, M.V., Koide, Y., Sasaki, K., Tsakirpaloglou, N. *et al*. (2013) NAL1 allele from a rice landrace greatly increases yield in modern indica cultivars. *Proc. Natl. Acad. Sci. USA*, **110**, 20431–20436.

Goel, M., Sun, H., Jiao, W.B. and Schneeberger, K. (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, 121–124.

Haas, B.J., Salzberg, S.L., Wei, Z., Pertea, M., Allen, J.E., Orvis, J., White, O. *et al*. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7.

Hämälä, T., Wafula, E.K., Guiltinan, M.J., Ralph, P.E., dePamphilis, C.W. and Tiffin, P. (2021) Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proc. Natl. Acad. Sci. USA*, **118**, e2102914118.

Han, Y. and Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199.

He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrest, K. *et al*. (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.* **51**, 896–904.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F. *et al*. (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, **26**, 121–135.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y. *et al*. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832.

Hu, B., Wang, W., Ou, S., Tang, J., Li, H., Che, R., Zhang, Z. *et al*. (2015) Variation in NRT1.1B contributes to nitrate-use divergence between rice subspecies. *Nat. Genet.* **47**, 834–838.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C. *et al*. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967.

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W. *et al*. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.* **44**, 32.

Hufford, M.B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.M., Cartwright, R.A., Elshire, R.J. *et al*. (2012) Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811.

Jens, K., Michael, W., Erickson, J.L., Schattat, M.H., Jan, G. and Frank, H. (2016) Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467.

Kapun, M. and Flatt, T. (2019) The adaptive significance of chromosomal inversion polymorphisms in *Drosophila melanogaster*. *Mol. Ecol.* **28**, 1263–1282.

Kawahara, Y., Bastide, M.D.L., Hamilton, J.P., Kanamori, H., Mccombie, W.R., Shu, O., Schwartz, D.C. *et al*. (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.

Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinform.* **5**, 59.

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

Li, Y., Fan, C., Xing, Y., Jiang, Y., Luo, L., Sun, L., Shao, D. *et al*. (2011) Natural variation in GS5 plays an important role in regulating grain size and yield in rice. *Nat. Genet.* **43**, 1266–1269.

Li, Y., Fan, C., Xing, Y., Yun, P., Luo, L., Yan, B., Peng, B. *et al*. (2014) Chalk5 encodes a vacuolar H^+-translocating pyrophosphatase influencing grain chalkiness in rice. *Nat Genet.* **46**, 398–404.

Li, W., Zhu, Z., Chern, M., Yin, J., Yang, C., Ran, L., Cheng, M. *et al*. (2017) A natural allele of a transcription factor in rice confers broad-spectrum blast resistance. *Cell*, **170**, 114–126.

Li, X., Wu, L., Geng, X., Xia, X., Wang, X., Xu, Z. and Xu, Q. (2018a) Deciphering the environmental impacts on rice quality for different rice cultivated areas. *Rice*, **11**, 7.

Li, X., Wu, L., Wang, J., Sun, J., Xia, X., Geng, X., Wang, X. *et al*. (2018b) Genome sequencing of rice subspecies and genetic analysis of recombinant lines reveals regional yield- and quality-associated loci. *BMC Biol.* **16**, 102.

Li, X., Chen, Z., Zhang, G., Lu, H., Qin, P., Qi, M., Yu, Y. *et al*. (2020) 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.

Li, G., Wang, L., Yang, J., He, H., Jin, H., Li, X., Ren, T. *et al*. (2021) A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. *Nat. Genet.* **53**, 574–584.

Long, Y., Zhao, L., Niu, B., Su, J., Wu, H., Chen, Y., Zhang, Q. *et al*. (2008) Hybrid male sterility in rice controlled by interaction between divergent alleles of two adjacent genes. *Proc. Natl. Acad. Sci. USA*, **105**, 18871–18876.

Lu, G., Wu, F.Q., Wu, W., Wang, H.J., Zheng, X.M., Zhang, Y., Chen, X. *et al*. (2014) Rice LTG1 is involved in adaptive growth and fitness under low ambient temperature. *Plant J.* **78**, 468–480.

Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T., Li, Y. *et al*. (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* **6**, 6914.

Lye, Z.N. and Purugganan, M.D. (2019) Copy number variation in domestication. *Trends Plant Sci.* **24**, 352–365.

Ma, X., Zhang, Q., Zhu, Q., Liu, W., Chen, Y., Qiu, R., Wang, B. *et al*. (2015) A robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Mol. Plant*, **8**, 1274–1284.

Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018) MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944.

Marchlerbauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., Deweesescott, C., Fong, J.H. *et al*. (2011) CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, 225–229.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., Radchuk, V. *et al*. (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. and Lanfear, R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534.

Muthayya, S., Sugimoto, J.D., Montgomery, S. and Maberly, G.F. (2014) An overview of global rice production, supply, trade, and consumption. *Ann. N. Y. Acad. Sci.* **1324**, 7–14.

Nagano, K., Sasaki, K. and Endo, T. (2013) Breeding of new rice cultivar 'Tohoku 194' with 'Sasanishiki'-type good eating quality of cooked rice. *Breed Sci.* **63**, 233–237.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T. *et al*. (2006) Dramatic amplification of a rice transposable element during recent domestication. *Proc. Natl. Acad. Sci. USA*, **103**, 17620–17625.

Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Nishimura, A., Aichi, I. and Matsuoka, M. (2006) A protocol for Agrobacterium-mediated transformation in rice. *Nat. Protoc.* **1**, 2796–2802.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E. and Lercher, M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936.

Price, A.L., Jones, N.C. and Pevzner, P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21**(Suppl 1), i351.

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

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S. *et al*. (2019) Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q. *et al*. (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, e3516.

Rambaut, A. (2009) *FigTree, a graphical viewer of phylogenetic trees*.

Rong, S., Chu, J.S.C., Ke, W., Jian, P. and Chen, N. (2009) genBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149.

Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E. *et al*. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259.

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., Lin, H. *et al*. (2022) A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896.

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S. and Yano, M. (2008) Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**, 1023–1028.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Song, W.Y., Wang, G.L., Chen, L.L., Kim, H.S., Pi, L.Y., Holsten, T., Gardner, J. *et al*. (1995) A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. *Science*, **270**, 1804–1806.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**, 215–225.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C., Chougule, K. *et al*. (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. *Nat. Genet.* **50**, 285–296.

Tang, L. and Chen, W. (2021) Development trend and prospect of Geng rice in northeast China (in Chinese). *China Rice*, **27**, 1–4.

Tarailograovac, M. and Chen, N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **5**, 4–10.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B. *et al*. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28.

Vaser, R., Sović, I., Nagarajan, N. and Šikić, M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A. *et al*. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE*, **9**, e112963.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Tae-Ho, L. *et al*. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M. *et al*. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

Wei, X., Qiu, J., Yong, K., Fan, J., Zhang, Q., Hua, H., Liu, J. *et al*. (2021) A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat. Genet.* **53**, 243–253.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A. *et al*. (2009) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **10**, 276.

Wu, Y., Wang, Y., Mi, X.F., Shan, J.X., Li, X.M., Xu, J.L. and Lin, H.X. (2016) The QTL GNP1 encodes GA20ox1, which increases grain number and yield by Increasing cytokinin activity in rice panicle meristems. *PLoS Genet.* **12**, e1006386.

Xie, W., Wang, G., Yuan, M., Yao, W., Lyu, K., Zhao, H., Yang, M. *et al*. (2015) Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc. Natl. Acad. Sci. USA*, **112**, E5411–E5419.

Xie, Z., Yan, B., Shou, J., Tang, J., Wang, X., Zhai, K., Liu, J. *et al*. (2019) A nucleotide-binding site-leucine-rich repeat receptor pair confers broad-spectrum disease resistance through physical association in rice. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20180308.

Xu, Z. and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268.

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A.M. *et al*. (2006) Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature*, **442**, 705–708.

Zdobnov, E.M. and Apweiler, R. (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.

Zhang, F., Wang, C., Li, M., Cui, Y., Shi, Y., Wu, Z., Hu, Z. *et al*. (2021) The landscape of gene-CDS-haplotype diversity in rice: properties, population organization, footprints of domestication and breeding, and implications for genetic improvement. *Mol. Plant*, **14**, 787–804.

Zhang, F., Xue, H., Dong, X., Li, M., Zheng, X., Li, Z., Xu, J. *et al*. (2022) Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.* **32**, 853–863.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 The summary of 12 GJ for nanopore sequencing.
Table S2 The summary of 12 GJ for Illumina sequencing.
Table S3 The summary of 12 GJ for chromosome conformation capture (Hi-C) sequencing.
Table S4 Statistics of genomic assembly and annotation for 58 long-read assemblies.
Table S5 Summary of structural variations among wild, GJ, circum-Aus group (cA), circum-Basmati group (cB) and XI groups.
Table S6 The functional gene with CNVs.
Table S7 The copy numbers of functional genes in both Japanese and Chinese GJ varieties.
Table S8 The GJ varieties used in this study.
Table S9 The reported gene is located in the select sweeps.
Figure S1 Hi-C interaction matrices show the pairwise correlations between ordered scaffolds along the 12 pseudomolecules.
Figure S2 The aggregate SV distribution of SVs among 58 assemblies.
Figure S3 Histogram of the size distribution of insertions, deletions, inversions, translocations and other types among 18 assemblies.
Figure S4 The distribution of total INS, DEL, INV and TRA of 18 assemblies.
Figure S5 The example of the distribution of INS, DEL, INV and TRA, and other 18 assemblies on chromosomes 2, 3 and 4.
Figure S6 The percentage of SVs overlap with different genomic regions among 18 assemblies.
Figure S7 Gene ontology analyses and KEGG analyses for SVs-related genes.
Figure S8 An in-depth analysis to demonstrate the difference in gene expression characteristics between different types.
Figure S9 The GO terms of CNV genes are potentially related to important agronomic traits.
Figure S10 Correlation between *GNP1* expression levels and the copy number of *GNP1*.
Figure S11 Geographic distribution of the 74 GJ varieties used in this study.
Figure S12 The 2020 rice production in northern China.
Figure S13 Gene ontology analyses and KEGG analyses for SVs-related genes located in the selected sweeps.
Figure S14 The fragment contains *GS5* and *Chalk5* on chromosome 5.
Figure S15 The LD block analysis for a collection comprised of 1275 rice accessions.
Figure S16 The yield and quality traits of *Chalk5* CRISPR knockout lines, and different letters indicate significant differences ($P < 0.05$, one-way ANOVA, Tukey's HSD test).