frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 18 November 2022
DOI 10.3389/fpls.2022.1065449

OPEN ACCESS

EDITED BY
Zhenbin Hu,
Saint Louis University, United States

REVIEWED BY
Jindong Liu,
Institute of Crop Sciences (CAAS), China
Yuanda Lv,
Jiangsu Academy of Agricultural Sciences (JAAS), China
Jingguang Chen,
Sun Yat-sen University, China
Jianping Yu,
Beijing University of Agriculture, China

*CORRESPONDENCE
Longbiao Guo
guolongbiao@caas.cn
Lianguang Shang
hanglianguang@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Functional and Applied Plant Genomics, a section of the journal Frontiers in Plant Science

RECEIVED 09 October 2022
ACCEPTED 03 November 2022
PUBLISHED 18 November 2022

CITATION
Ma J, Wei H, Yu X, Lv Y, Zhang Y, Qian Q, Shang L and Guo L (2022) Compared analysis with a high-quality genome of weedy rice reveals the evolutionary game of de-domestication. *Front. Plant Sci.* 13:1065449. doi: 10.3389/fpls.2022.1065449

Compared analysis with a high-quality genome of weedy rice reveals the evolutionary game of de-domestication

Jie Ma[1,2†], Hua Wei[2†], Xiaoman Yu[1,2†], Yang Lv[1], Yu Zhang[1], Qian Qian[1,2], Lianguang Shang[2*] and Longbiao Guo[1*]

[1]State Key Lab for Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou, China, [2]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong, China

The weedy rice (*Oryza sativa* f. *spontanea*) harbors large numbers of excellent traits and genetic diversities, which serves as a valuable germplasm resource and has been considered as a typical material for research about de-domestication. However, there are relatively few reference genomes on weedy rice that severely limit exploiting these genetic resources and revealing more details about de-domestication events. In this study, a high-quality genome (~376.4 Mb) of weedy rice A02 was assembled based on Nanopore ultra-long platform with a coverage depth of about 79.3× and 35,423 genes were predicted. Compared to Nipponbare genome, 5,574 structural variations (SVs) were found in A02. Based on super pan-genome graph, population SVs of 238 weedy rice and cultivated rice accessions were identified using public resequencing data. Furthermore, the de-domestication sites of weedy rice and domestication sites of wild rice were analyzed and compared based on SVs and single-nucleotide polymorphisms (SNPs). Interestingly, an average of 2,198 genes about de-domestication could only be found by F_{ST} analysis based on SVs (SV-F_{ST}) while not by F_{ST} analysis based on SNPs (SNP-F_{ST}) in divergent region. Additionally, there was a low overlap between domestication and de-domestication intervals, which demonstrated that two different mechanisms existed in these events. Our finding could facilitate pinpointing of the evolutionary events that had shaped the genomic architecture of wild, cultivated, and weedy rice, and provide a good foundation for cloning of the superior alleles for breeding.

KEYWORDS

weedy rice, genome assembly, structural variation, pan-genome graph, de-domestication

Introduction

Different from cultivated rice, weedy rice (*Oryza sativa* f. *spontanea*) possesses stronger seed dormancy, reproductive ability, and higher phenotypic plasticity. Similar with wild rice (*Oryza rufipogon*), weedy rice also has features such as a red pericarp, a black hull, and seed shattering. Furthermore, weedy rice has stronger tolerance to biotic and abiotic stress than cultivated rice. On the whole, weedy rice is a gold mine, which contains numerous useful genetic resources for rice functional genomic studies and breeding (Wu et al., 2022). Since weedy rice has these reproductive advantages, many quantitative trait locus (QTLs) mapping traits for seed shattering and dormancy have been identified during decades past, such as shattering loci of *sh4* and *qSH1* (Qi et al., 2015) and seed dormancy and red pericarp major loci of *qSD7-1*/*qPC7* (Gu et al., 2004; Gu et al., 2011). Pan-genome of rice not only presents more useful genetic information for molecular breeding by design but also provides insights into the evolutionary events (Shang et al., 2022). The pan-genome of cultivated rice has been constructed, while weedy rice was missed in the currently available rice pan-genome (Wang et al., 2018; Zhao et al., 2018; Qin et al., 2021; Zhang et al., 2022). Therefore, few clues were provided to further understand the molecular mechanism of cultivated rice de-domestication.

De-domestication is an interesting phenomenon in both plants and animal, which denotes that the domesticated crops and livestock reacquire components of wild like traits to form independent reproducing population (Ellstrand et al., 2010; Gering et al., 2019). If de-domestication constantly appears in crop fields, it usually leads to weeds, such as weedy rice (Qiu et al., 2020), weedy barely (Zeng et al., 2018) and weedy sunflower (Presotto et al., 2011). In rice, multiple studies identified that many loci (*sh4*, *qSH1*, *SSH1*, *OSH15* and *GRF4*) were involved in shaping the loss of seed shattering, among which *sh4* and *qSH1* were the major loci (Li et al., 2006; Sun et al., 2016; Yoon et al., 2017; Jiang et al., 2019). For example, the cultivated rice contains the G-to-T mutation of *sh4* underlying the loss of seed shattering. Under natural selection, de-domestication could provide valuable genetic resources for crop breeding. To elucidate the mysterious of de-domestication will expand our understanding of the evolutionary process of crops.

Previous studies showed that an increasing number of de-domestication events have been revealed in crops, such as *Oryza sativa* (Li et al., 2017; Qiu et al., 2017; Sun et al., 2019; Qiu et al., 2020), *Hordeum agriocrithon* (Zeng et al., 2018), *Triticum aestivum* (Guo et al., 2020), *Sorghum bicolor* (Morrell et al., 2005), *Secale cereale* (Morrell et al., 2005), *Olea europaea* (Mekuria et al., 2002) and so on. However, of these events only three have been confirmed by genomic studies including rice, barley, and wheat. Therefore, the large scale and scope of genomic studies in crops will be helpful for confirming the de-domestication events and revealing more details about this event.

High-quality pan-genome can represent the full genetic information of a population and provide a new foundation for exploitation of genetic resources for crop improvement, which has been used to deeply analyze the large number of genetic variations, functional genes, species origin and domestication in rice (Wang et al., 2018; Zhao et al., 2018; Liu et al., 2020). Structural variation (SV), presence/absence variation (PAV) and gene copy number variation (gCNV) are the main causes of individual genomic differences and the core of pan-genome research. Among them, SV plays crucial role in crop evolution, domestication, and improvement. For example, Kou et al. (2020) used SV to study domestication by contrasting between Asian rice (*Oryza sativa*) and its wild relative *O. rufipogon* and found that SVs contributed to the domestication in rice. Recent studies also revealed the diverse genetic mechanism of rice de-domestication through the whole-genome sequencing of 524 global weedy rice and comparative analysis with accordingly cultivated rice. They mainly conducted single nucleotide polymorphism (SNP) across 1,003 samples genomic analysis (Qiu et al., 2020). However, the function of SV in de-domestication of rice has yet to be verified. Additionally, crop-weed genome comparison can provide new insights on the genetic mechanism of cultivated rice de-domestication process. In this study, Nanopore sequencing and assembly were performed on *japonica* weedy rice A02 and obtained a high-quality genetic map. Additionally, we used a variety of data source including super pan-genome graph to detect SVs in cultivated and weedy rice and evaluated the function and accuracy of SV which was performed as a tool to study de-domestication for weedy rice. And the de-domestication results were used to compare with domestication events of ordinary wild rice and cultivated rice populations.

Materials and methods

Plant materials and data preparation

High-latitude weedy rice A02 from Liaoning province was selected in our research and was grown in the greenhouse of China National Rice Research Institute. The Illumina resequencing data of 238 weedy rice and cultivated rice accessions used in this part were derived from public data (Qiu et al., 2017), which include 155 weedy rice samples from four representative provinces of Liaoning (LN), Ningxia (NX), Jiangsu (JS) and Guangdong (GD) in China, 76 local cultivated rice and 7 weedy rice accessions from the United States and South Korea. And the Illumina sequencing data of 26 Asian wild rice samples were collected from Shang et al. (2022). The super

pan-genome involved in structural variation detection was constructed by Nanopore sequencing data of rice germplasm materials (Shang et al., 2022) based on Nipponbare reference genome (MSUv7) (Kawahara et al., 2013).

Illumina sequencing

gDNA of A02 for short-read length sequencing was extracted from leaves of two-week-old seedlings with the CTAB method. The index libraries were constructed by using the New England Biolabs (NEB) Next® Ultra™ DNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) according to the manufacturer's instructions. After quality assessment, at least 0.2 μg gDNA was randomly fragmented by sonication. After size grading by electrophoresis, approximately 350 bp DNA fragments were purified with the AMPure XP system (Beckman Coulter, Beverly, USA) for library construction. It was then sequenced on the Xten platform (Illumina, San Diego, CA, USA).

Nanopore ultra-long sequencing

For the ultra-long Nanopore library, approximately 8-10 μg DNA (>50 Kb) of A02 was selected using the SageHLS HMW library system (Sage Science, USA) and processed using Ligation Sequencing Kit 1D (SQK-LSK109, Oxford Nanopore Technologies, UK). About 800 ng DNA library was constructed and sequenced on PromethION (Oxford Nanopore) to obtain the original sequencing data following the manufacturer's instructions.

Transcriptome sequencing

The RNA of A02 was extracted from young leaves of one-week-old seedlings with a TRIzol kit. The index library was constructed by TruSeq RNA Library Preparation Kit (Illumina, USA). And RNA was sequenced using Illumina sequencing platform NovaSeq 6000.

De novo genome assembly and evaluation

Based on ultra-long Nanopore sequencing platform, the qualified Nanopore ultra-long reads (NULRs) (quality value >7) of the weedy rice A02 were obtained. Then NULRs were assembled by NextDenovo v2.4 (https://github.com/Nextomics/NextDenovo). After preliminary assembly, the genome sequences were further corrected three times with 22.8 Gb (60.8×) of Illumina paired-end reads and the NULRs using NextPolish v1.0 (https://github.com/Nextomics/NextPolish). BUSCO v9 (Simão et al., 2015) was used to evaluate the assembled sequences, which means using single copy Embryophyta genes in a specific library to predict the genes status of existing sequences in the genome.

Chromosome assembly

Ragtag v2.1.0 (https://github.com/malonge/RagTag) was used to link the contigs into the closet chromosomes according to the coverage of contigs with Nipponbare genome (Kawahara et al., 2013) as reference and the chromosome-level assemblies were generated.

Gene annotation

For assembled A02 genome, the RepeatMasker v4.0.7 (www.repeatmasker.org) was used to mask the repetitive sequences. And the coding regions in the repeat-masked genome were predicted using Augustus v3.0.3 (Keller et al., 2011), SNAP v2006-07-28 (Leskovec and Sosic, 2016) and Fgenesh (http://www.softberry.com/). Proteins from four plant genomes (*Arabidopsis thaliana*, *Brachypodium distachyon*, *Os* and *Sorghum bicolor*) were downloaded from Phytozome (https://phytozome-next.jgi.doe.gov). These protein sequences were aligned to the assembly using tBLASTN v2.9.0+ (Camacho et al., 2009) with an *E*-value cutoff of 1e-5. Genewise v2.4.2 (Birney et al., 2004) was used to refine the alignment with parameters '-gff -quiet -silent -sum'. Raw RNA-seq reads were qualified by Trimmomatic v0.36 (Bolger et al., 2014) with parameters 'ILLUMINACLIP : TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36'. The clean RNA-seq data were mapped to the assembly using HISAT2 v2.1.0 (Kim et al., 2019). StringTie2 v2.1.4 (Kovaka et al., 2019) was used to assemble the transcripts into gene models. All gene models were integrated by EvidenceModeler v1.1.1 (Haas et al., 2008) to obtain the gene annotation results.

Transposable element annotation

The whole-genome TEs were annotated with the TE library in the Extensive *de novo* TE Annotator (EDTA, v1.9.6) (Ou et al., 2022) package.

SNP and SV calling

Raw DNA-seq reads from A02 gDNA sample were trimmed by Trimmomatic (Bolger et al., 2014) with parameters 'ILLUMINACLIP:2:30:10 MINLEN:75 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20'. Clean reads were

mapped to the Nipponbare reference genome (Kawahara et al., 2013) by Burrows-Wheeler Aligner (BWA, v0.7.17-r1188) (Li and Durbin, 2009). SAMtools v1.8 (Danecek et al., 2021) and BCFtools v1.8 (Danecek et al., 2021) were used to call and filter (DP < 3 and quality score < 30) SNP.

To call SVs, NULRs of A02 were aligned to the Nipponbare genome (Kawahara et al., 2013) using minimap2 v2.17-r974-dirty (Li, 2018) and NGMLR v0.2.7 (Sedlazeck et al., 2018). Sniffles v1.0.11 (Sedlazeck et al., 2018) was used to call SVs with parameters '-I 50 -genotype'.

SV calling based on Illumina sequencing data: raw Illumina resequencing data of 238 accessions from public data (Qiu et al., 2017) were first trimmed by Trimmomatic (Bolger et al., 2014) with parameters 'ILLUMINACLIP:2:40:15 MINLEN:100 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:15'. After quality control, the clean reads were mapped to the super pan-genome (Shang et al., 2022) to call SV with the short reads comparison tools Giraffe (Siren et al., 2021) in vg toolkit v1.38.0 (Hickey et al., 2020). SURVIVOR v1.0.7 (Jeffares et al., 2017) was used to merge SVs called by each accession into a population genotype.

Genetic differentiation (F_{ST}) analysis

In the study of weedy rice genome variation, Qiu et al. (2017) divided weedy rice and local cultivated rice in China into four group: Guangdong (GD), Jiangsu (JS), Liaoning (LN) and Ningxia (NX) according to the geographical location. In order to avoid the noise effect caused by population structure mixing, we selected four subgroups (NX1, LN1, JS1 and GD1) with clear population structure published by Qiu et al. (2017) to analyze the genetic differentiation of weedy rice and cultivated rice. We performed F_{ST} analysis based on SVs (SV-F_{ST}) and SNPs (SNP-F_{ST}), respectively. VCFtools v0.1.13 (Danecek et al., 2011) was used to calculate SV-F_{ST} and SNP-F_{ST} between weedy rice and local cultivated rice in four regions, and between Asian wild rice and cultivated rice as well. SNP-F_{ST} was performed in a 100 Kb window size with a step size of 10 Kb, and the parameters of SV-F_{ST}, referring to Kou et al. (2020), were set in a 20 Kb window size with a step size of 10 Kb. Through the above SNP-F_{ST} and SV-F_{ST}, the windows with top 5% F_{ST} values were identified as highly differentiated intervals. Genes in these series of intervals were then annotated based on the Nipponbare reference genome (Kawahara et al., 2013).

Results

Assembly and validation of high-quality genome sequences of weedy rice

To develop assemblies, 29.9 Gb (~79.3× coverage) of NULRs were generated for weedy rice A02 using Nanopore ultra-long platform in which the N50 and the longest reads can reach 52.9 Kb and 520.6 Kb respectively, of which the average read lengths are 30.7 Kb (Supplementary Figure S1 and Supplementary Table S1). To improve accuracy, we used Illumina sequencing data to reduce the single-base and InDel error rate. After quality control of the Illumina reads data was completed, weedy rice retained 22.8 Gb of data with a coverage depth of about 60.8×. The Q20 of the data reached more than 97.0%. The assembled weedy rice genome was 376.4 Mb based on NULRs and contain 19 contigs, of which the contig N50 length was 29.39 Mb (Table 1). The 376.4 Mb sequences of A02 were anchored to 12 chromosomes, in which the BUSCO value reached more than 98% (Figure 1A and Table 1). The above data suggested that a high-quality genome of weedy rice was achieved.

Annotation and comparison of the genome sequences for weedy rice

To determine the content of TEs, genome-wide annotation of the transposon sequence about weedy rice A02 was performed based on the TE library in the extensive *de novo* TE Annotator (Ou et al., 2022) and 183.9 Mb of TEs were identified, accounting for 48.95% of the entire genome. The most abundant class of TEs were the long terminal repeat (LTR), accounting for 22.75% of the genome (Table 2). Additionally, 35,423 genes were predicted in the weedy rice A02 genome, with an average gene length of 2.85 Kb. Subsequently, the genome sequence of the weedy rice at the chromosome level was aligned with the Nipponbare sequence using MUMmer 4.0.0 (Marçais et al., 2018), which suggested that the weedy rice had good collinearity compared with Nipponbare (Figure 1B). To better compare the variation of weedy rice A02, the genome-wide collinearity analysis of weedy rice A02 was carried out with the reference of Nipponbare genome sequence, and the results could show the SVs of A02 relative to Nipponbare (Figure 1C), which would provide a well-grounded basis for resolving the mechanisms associated with phenotypic variation in weedy rice.

TABLE 1 The statistics of Nanopore ultra-long reads assembly results of A02.

Sequencing type	NULRs
Total size of assembled genome (Mb)	376.4
Number of contigs	19
Largest contig (Mb)	43.54
Contig N50	29.39
Single base calling error (%)	0.02
BUSCOs (%)	98.00

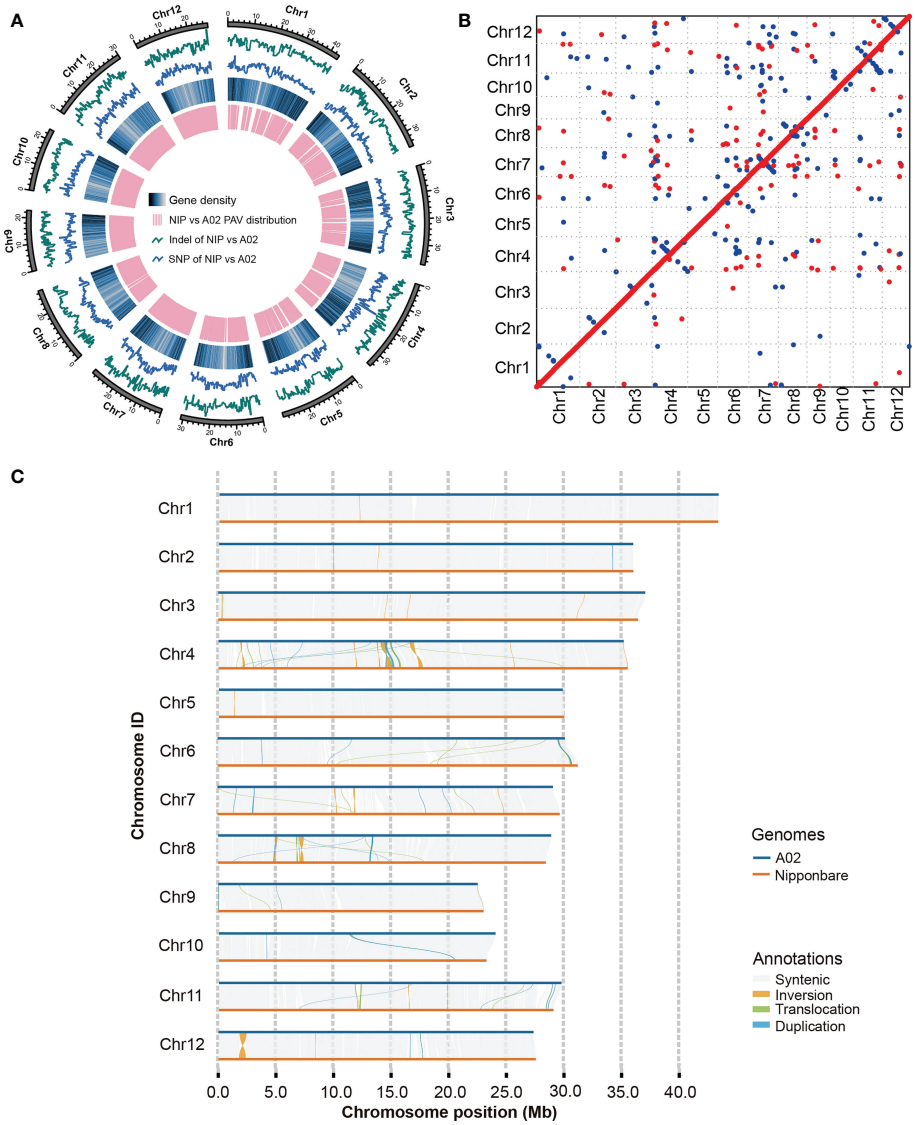

FIGURE 1
Assembly and analysis of A02 genome. **(A)** Landscape of A02 reference genome. Tracks from the outer to inner circles indicate: 1, indels between A02 and the Nipponbare reference genome; 2, SNPs in A02 with respect to Nipponbare; 3, gene density; 4, highlights of the PAV distribution between A02 and Nipponbare. NIP refers to Nipponbare. **(B)** The collinearity relationship between A02 genome and Nipponbare. The X-axis represents the genome of Nipponbare, and the Y-axis indicates the genome of weedy rice A02. **(C)** The collinearity of weedy rice A02 and Nipponbare.

SV identification and analysis of de-domestication based on super pan-genome

A super pan-genomic landscape of rice revealed extensive SVs, which enabled the accurate identification and characterization of their inter- and intraspecific diversity (Shang et al., 2022). Previous studies have showed that SVs underlie important crop improvement and domestication traits. To explore the role of SVs in de-domestication process, we calculated the SVs distribution of all weedy and cultivated rice genome based on super pan-genome graph. The results showed that the number of SVs was significantly different among materials, which was mainly related to the distance of evolutionary relationship between these materials and Nipponbare. For example, *indica* and *aus* subgroups generally contained more SVs than *japonica* subgroup with an average number of 20,327, while the *japonica* had an average of only 7,397 (Figure 2A). Next, we used the published data to count SV numbers for weedy rice subpopulations (NX1, LN1, JS1, and GD1) and the well-defined local cultivated rice (NX_C, LN_C, JS_C, and

TABLE 2 The statistics of transposable elements annotation results in A02.

	Type of TEs	Number	Length (bp)	Proportion (%)
LTR	*copia*	11,521	13,390,207	3.56
	gypsy	46,404	70,063,097	18.65
	TRIM	2,887	577,556	0.15
	Unknown	1,377	1,450,919	0.39
TIR	CACTA	18,813	15,024,542	4.00
	Mutator	53,486	21,063,723	5.61
	PIF/Harbinger	43,381	9,896,554	2.63
	Tc1/*mariner*	54,224	16,098,042	4.29
	*h*AT	16,315	5,568,481	1.48
	Unknown	6,610	2,068,969	0.55
nonLTR	LINEs	6,968	3,861,853	1.03
	SINEs	6,420	1,228,455	0.33
	Unknown	302	544,273	0.14
nonTIR	*Helitron*	48,357	23,077,004	6.14
Total		317,065	183,913,675	48.95

GD_C), respectively, which have been reported by Qiu et al. (2017). These data showed that the distribution of SVs in most groups was relatively uniform, except for JS_C (Figure 2B), which contains both *japonica* and *indica* materials. Additionally, we performed SV frequency statistics in 162 weedy and 76 cultivated rice and found that an average of 13.8% SVs only appeared in a single material and existed at private frequencies (Figures 2C–E), which further illustrated that SVs were generally harmful for plant growth and development.

To analyze the role of SVs in the process of rice de-domestication, we selected four subgroups with clear population structure (NX1, LN1, JS1 and GD1) to study population genetic differentiation. Through the SV-F_{ST} analysis of weedy and local cultivated rice populations, the highly differentiated ranges with top 5% SV-F_{ST} values were annotated in combination with the Nipponbare reference genome (Kawahara et al., 2013), among which these genes were mostly associated with heading date, stress or plant

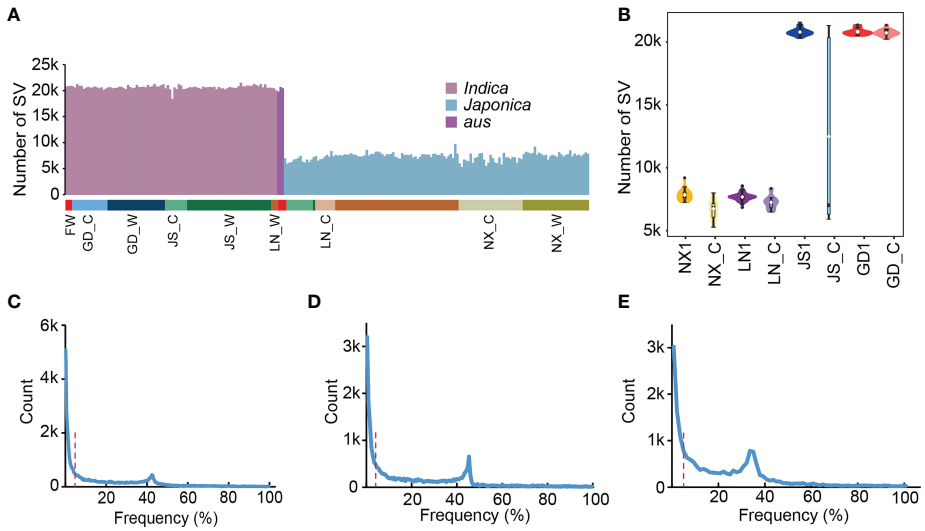

FIGURE 2
The statistics of SVs' number and frequency. **(A)** The statistics of SVs' number in 238 rice materials, which contains 162 weedy and 76 locally cultivated rice. LN for Liaoning, NX for Ningxia, JS for Jiangsu, GD for Guangdong and FW for weedy rice sampled outside of China. **(B)** The number of SVs in the four weedy rice subpopulations (NX1, LN1, JS1 and GD1) and the corresponding local cultivated rice. **(C–E)** Frequency of SVs in 238 rice materials **(C)**, 162 weedy rice **(D)**, and 76 local cultivated rice **(E)**. The red dashed line represents a frequency of 5%.

growth and development (Figures 3A–D). Subsequently, we used SNPs data that published by Qiu et al. (2017) to conduct SNP-F_{ST} analysis, and screened out highly differentiated windows in the same way. Then, the similarities and differences of differentiation intervals based on SNP and SV detection were further analyzed (Figures 3E–H). These data showed that the overlapped gene accounted for about 50% genes detected through SV-F_{ST} in four different areas (Figures 3E–H), indicating that SV-F_{ST} could verify the differentiation range obtained by SNP-F_{ST}. Additionally, SV-F_{ST} also detected many highly differentiated genes that could not be found by SNP-F_{ST}. To explore whether there are some common de-domestications related intervals in the four different regions, we investigated the simultaneous intersection of the highly differentiated intervals detected by SV-F_{ST} analysis and found a few identical windows (only five common genes) in different regional groups (Figure 3I). This indicated that different genes were selected to cope with the selection pressure and adapt to the local environment in the de-domestication process of weedy rice. Among them, the *OsHAP2E* gene was detected in all four regions (Figure 3I), which was reported to confer salt and drought tolerance, phytophthora ramorum and phytophthora albicanan resistance, improve the capacity of photosynthetic, and increase tiller number in rice (Alam et al., 2015). These characteristics were closely related to weedy rice, suggesting *OsHAP2E* might be involved in the de-domestication process to some extent. Moreover, a highly differentiated SV locus was identified, which was a 5,755 bp PAV, and a functional gene *NAL1* (*narrow leaf 1*) was detected in this genomic region in Nipponbare but not in A02 (Figure 3J). Previous studies have

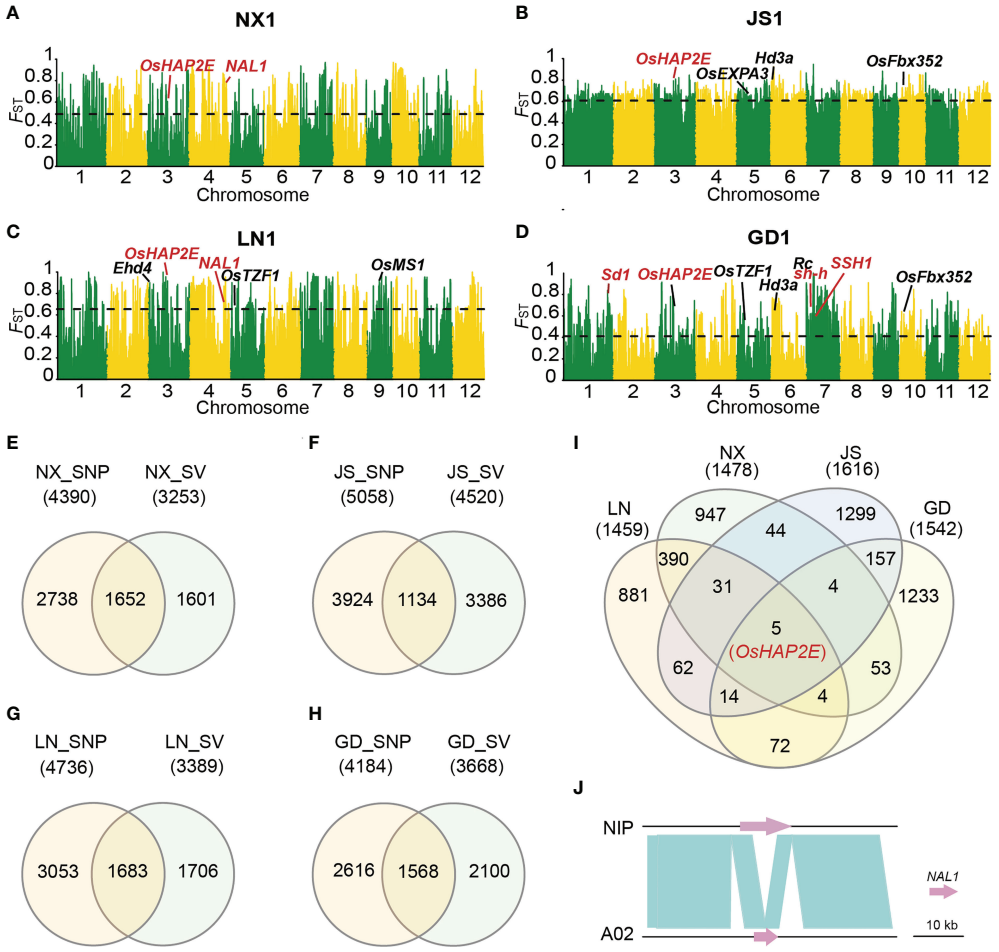

FIGURE 3
Genomic differentiation between weedy and cultivated rice from different locations. **(A-D)** SV-F_{ST} between weedy rice NX1 **(A)**, JS1 **(B)**, LN1 **(C)**, GD1 **(D)** and their corresponding cultivated rice. Genes emphasized in the text were marked in red. Horizontal dashed lines correspond to the top 5% threshold. **(E-H)** The intersection of divergent gene regions identified through SV-F_{ST} and SNP-F_{ST} in NX **(E)**, JS **(F)**, LN **(G)**, GD **(H)**. **(I)** The intersection of divergent windows identified in four regions through SV-F_{ST}. **(J)** The variation of the *NAL1* gene between Nipponbare and A02. Most weedy rice lost a 5755 bp fragment in the region of the *NAL1* gene compared to Nipponbare.

shown that *NAL1* positively regulate plant leaf width and plant height (Xu et al., 2015). Interestingly, this fragment was present in most of the cultivated rice while absent in weedy rice, indicating that it might be related to the characteristics of weedy rice. For example, in NX, this fragment was missed in 21 of 22 weedy rice, while only absent in one accession of the cultivated rice. Similar results were obtained in LN, where the PAV was absent in 38 of 40 weedy rice but present in all cultivated rice, which indicated a high degree of differentiation between weedy and cultivated rice.

Comparison of domestication and de-domestication events

To explore the similarities and differences between the mechanisms of the domestication and de-domestication processes, we further performed SV-F_{ST} analysis on 26 wild rice and cultivated rice in NX, LN, JS and GD to obtain the domestication related sites, respectively (Figures 4A–D). Subsequently, we compared these intervals between domestication and de-domestication, which showed these genes had a low overlap between domestication and de-domestication intervals in highly differentiated regions (Figures 4E–H). Unusually, these intervals had a high overlap that accounted for 20.8% of total (705/3389) in the LN area (Figure 4G). Thus, the mechanisms of domestication and de-domestication had some similarities. For example, *SSH1* (*suppression of shattering 1*), controlling shattering and grain size, was present in both domestication and de-domestication process detected by SV-F_{ST} in GD (Figure 4D). In the de-domestication region, the *sd1* (*Semi-dwarf* 1) gene was only detected in GD (Figure 3D), while *sd1* played an important role in domestication in four areas (Figures 4A–D). And large numbers of genes were only found in significantly differentiated regions related to the domestication process, such as *SHAT1* (*Shattering abortion1*) and *DEP1* (*Dense panicle1*), which could not be detected in de-domestication region (Figures 3A–D). On the contrary, shattering gene *sh-h* was only detected in the de-domestication related region (Figure 3D). Taken together, the de-domestication process from cultivated to weedy rice might have different genetic mechanisms with the domestication process from wild to cultivated rice. Our findings highlighted the underexplored role of SVs in de-domestication process and their widespread importance and utility in crop improvement.

Discussion

Weedy rice (*Oryza sativa* f. *spontanea*) has many excellent traits, such as strong growth potential, environmental adaptability and multiple stresses resistance. High-quality genome sequences are the bases for mining functional genes, evolutionary origins, and genetic breeding. Here, we assembled a high-quality reference genome sequences of weedy rice A02 using Nanopore ultra-long platform. Compared with next-generation sequencing, Nanopore sequencing has greatly improved the read length, but there are still many assembly errors in long repetitive regions. Nanopore ultra-long sequencing is one of the best methods to solve this problem, which makes it possible to telomere-to-telomere assembly genome rapidly (Miga et al., 2020). NULRs have longer N50 and can effectively span large repeats in the genome, even the centromeric regions. The combination of these two sequencing platforms facilitated the assembly of the A02 high-quality genome, which had good collinearity with the reference genome of Nipponbare (Figure 1B). Through transposon annotation, we found that the weedy rice A02 contained 183.9 Mb transposons, accounting for 48.95% of the whole genome, and 35,423 protein-coding genes were predicted. Our study therefore provided genes pools for gene excavation to underlie the research about de-domestication of weedy rice.

Except for SNPs, SVs are the major source of genetic variation and tend to have a great impact on gene expression and phenotype in plants (Alonge et al., 2020). Consequently, using super pan-genome as a reference genome, SV-based and SNP-based genetic differentiation analysis were performed on 162 weedy rice populations mainly from GD, JS, LN and NX and 76 local cultivated rice populations, and found that SVs existed mainly at low frequencies in weedy rice populations compared with SNPs (Figure 3). By further analyzing the de-domestication loci in the four regions, we found that many loci could be detected through SV-F_{ST} but not SNP-F_{ST}, and most of these de-domestication related SV loci were unique to each region. Additionally, we compared SV loci involved in the de-domestication and domestication events, and found that the shared differentiation loci only accounted for 20% (Figure 4), suggesting that the mechanism of de-domestication process of weedy rice might be different from that of the domestication process.

For another, SV-F_{ST} can detect many highly differentiated genes of weedy rice while the locally cultivated rice that cannot be found by SNP-F_{ST}. For example, the *SSH1* can be specifically detected in the highly differentiated interval of weedy rice and cultivated rice in GD with SV-F_{ST}. Previous studies have showed that *SSH1* controls the grain size and seed shattering by positively regulating the expression of two rice *REPLUMLESS* orthologs *qSH1* and *SH5*, which indicated that *SSH1* is valuable for improving rice seed shattering and grain yield (Jiang et al., 2019). Noticeably, a highly differentiated SV site, PAV of 5,755 bp, was present in most cultivated rice while lost in weedy rice, which was present on the functional gene *NAL1* that regulate leaf width and plant height (Xu et al., 2015) (Figure 3J). Genomic loci present in the highly differentiated regions might be crucial for the origin and adaption of weedy rice. At present, the excavation of genes controlling excellent traits such as stress resistance in

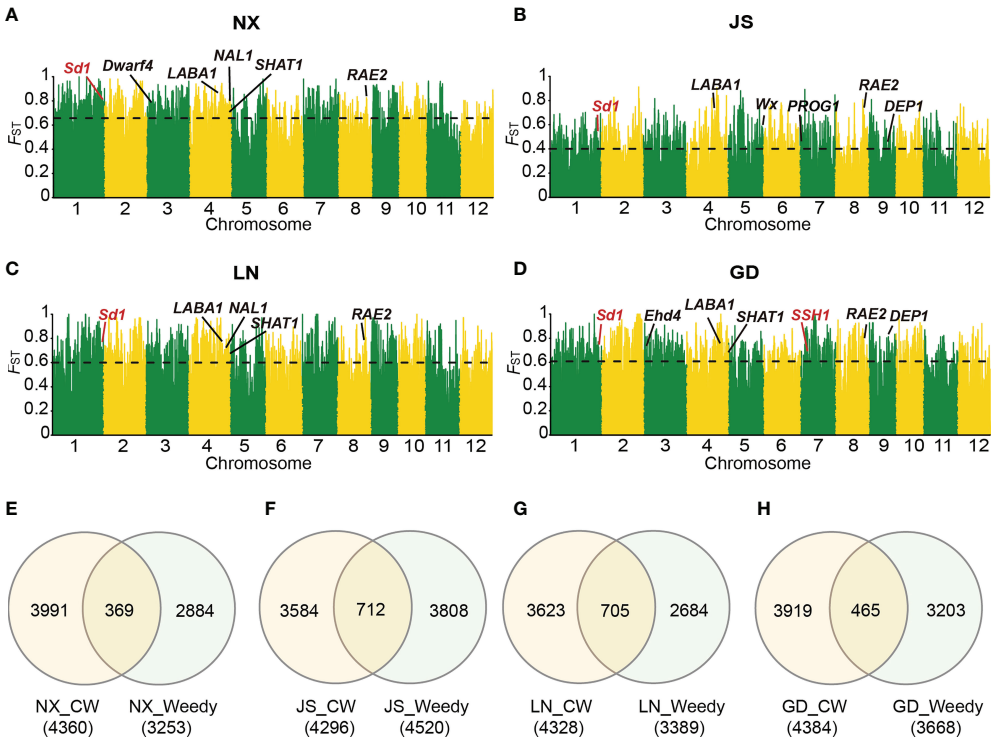

FIGURE 4
Genomic differentiation between Asian wild rice and cultivated rice from different locations. **(A-D)** SV-F_{ST} between Asian wild rice and cultivated rice from NX **(A)**, JS **(B)**, LN **(C)** and GD **(D)**, respectively. Genes emphasized in the text were marked in red. Horizontal dashed lines correspond to the top 5% threshold. **(E-H)** The intersection of divergent gene regions identified through SV-F_{ST} between Asian wild rice and cultivated rice and that between weedy and cultivated rice in four regions NX **(E)**, JS **(F)**, LN **(G)** and GD **(H)**, respectively. CW means cultivated rice.

weedy rice is still insufficient. Our research provided a solid foundation to further disclosing the genetic variation mechanisms of weedy rice and the cloning of beneficial genes for breeding.

Conclusions

Herein, we assembled a high-quality weedy rice genome of A02 using Nanopore ultra-long platform, which provided a good foundation for understanding the molecular mechanism of cultivated rice de-domestication process and cloning of the superior alleles for breeding. Based on the super pan-genome and comparative genomic analysis, we found that two different mechanisms existed in domestication and de-domestication events, and different genes were selected to cope with the selection pressure and adapt to the local environment in the de-domestication process.

Data availability statement

The datasets presented in this study can be found in online repositories. The names ofthe repository/repositories and accession number(s) can be found below: https://ngdc.cncb.ac.cn/gsub/, PRJCA011659. The SV information between the A02 genome and Nipponbare genome and the data of the SV genotyping information based on the super pan-genome are available in https://zenodo.org/record/7265880#.Y187b-RBxD9.

Author contributions

LG, LS, and QQ designed the experiment. JM, XY, YL, and YZ performed analysis and interpretation of the data. JM and HW drafted the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from The National Natural Science Foundation of China (Grant No.32101718 and No. 3221101587); Hainan Yazhou Bay Seed Laboratory Project (B21HJ0223).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1065449/full#supplementary-material

References

Alam, M. M., Tanaka, T., Nakamura, H., Ichikawa, H., Kobayashi, K., Yaeno, T., et al. (2015). Overexpression of a rice heme activator protein gene (OsHAP2E) confers resistance to pathogens, salinity and drought, and increases photosynthesis and tiller number. *Plant Biotechnol. J.* 13 (1), 85–96. doi: 10.1111/pbi.12239

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182 (1), 145–161.e123. doi: 10.1016/j.cell.2020.05.021

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14 (5), 988–995. doi: 10.1101/gr.1865504

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10, 421. doi: 10.1186/1471-2105-10-421

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi: 10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience* 10 (2), giab008. doi: 10.1093/gigascience/giab008

Ellstrand, N. C., Heredia, S. M., Leak-Garcia, J. A., Heraty, J. M., Burger, J. C., Yao, L., et al. (2010). Crops gone wild: evolution of weeds and invasives from domesticated ancestors. *Evol. Appl.* 3 (5-6), 494–504. doi: 10.1111/j.1752-4571.2010.00140.x

Gering, E., Incorvaia, D., Henriksen, R., Conner, J., Getty, T., and Wright, D. (2019). Getting back to nature: Feralization in animals and plants. *Trends Ecol. Evol.* 34 (12), 1137–1151. doi: 10.1016/j.tree.2019.07.018

Gu, X. Y., Foley, M. E., Horvath, D. P., Anderson, J. V., Feng, J., Zhang, L., et al. (2011). Association between seed dormancy and pericarp color is controlled by a pleiotropic gene that regulates abscisic acid and flavonoid synthesis in weedy red rice. *Genetics* 189 (4), 1515–1524. doi: 10.1534/genetics.111.131169

Gu, X. Y., Kianian, S. F., and Foley, M. E. (2004). Multiple loci and epistases control genetic variation for seed dormancy in weedy rice (Oryza sativa). *Genetics* 166 (3), 1503–1516. doi: 10.1534/genetics.166.3.1503

Guo, W., Xin, M., Wang, Z., Yao, Y., Hu, Z., Song, W., et al. (2020). Origin and adaptation to high altitude of Tibetan semi-wild wheat. *Nat. Commun.* 11 (1), 5085. doi: 10.1038/s41467-020-18738-5

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9 (1), R7. doi: 10.1186/gb-2008-9-1-r7

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Siren, J., Eizenga, J., et al. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* 21 (1), 35. doi: 10.1186/s13059-020-1941-7

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., et al. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* 8, 14061. doi: 10.1038/ncomms14061

Jiang, L., Ma, X., Zhao, S., Tang, Y., Liu, F., Gu, P., et al. (2019). The APETALA2-like transcription factor SUPERNUMERARY BRACT controls rice seed shattering and seed size. *Plant Cell* 31 (1), 17–36. doi: 10.1105/tpc.18.00304

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., et al. (2013). Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. *Rice (N Y)* 6 (1), 4. doi: 10.1186/1939-8433-6-4

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27 (6), 757–763. doi: 10.1093/bioinformatics/btr010

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37 (8), 907–915. doi: 10.1038/s41587-019-0201-4

Kou, Y. X., Liao, Y., Toivainen, T., Lv, Y. D., Tian, X. M., Emerson, J. J., et al. (2020). Evolutionary genomics of structural variation in Asian rice (Oryza sativa) domestication. *Mol. Biol. Evol.* 37 (12), 3507–3524. doi: 10.1093/molbev/msaa185

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 20 (1), 278. doi: 10.1186/s13059-019-1910-1

Leskovec, J., and and Sosic, R. (2016). SNAP: A general-purpose network analysis and graph-mining library. *Trans. Intell. Syst. Technol.* 8 (1). doi: 10.1145/2898361

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34 (18), 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, L. F., Li, Y. L., Jia, Y. L., Caicedo, A. L., and Olsen, K. M. (2017). Signatures of adaptation in the weedy rice genome. *Nat. Genet.* 49 (5), 811. doi: 10.1038/ng.3825

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* 182 (1), 162–176.e113. doi: 10.1016/j.cell.2020.05.023

Li, C., Zhou, A., and Sang, T. (2006). Rice domestication by reducing shattering. *Science* 3115769, 1936–1939. doi: 10.1126/science.1123604

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 (1), e1005944. doi: 10.1371/journal.pcbi.1005944

Mekuria, G. T., Collins, G., and Sedgley, M. (2002). Genetic diversity within an isolated olive (Olea europaea l.) population in relation to feral spread. *Sci. Hortic.* 94 (1), 91–105. doi: 10.1016/S0304-4238(01)00375-2

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585 (7823), 79–84. doi: 10.1038/s41586-020-2547-7

Morrell, P. L., Williams-Coplin, T. D., Lattu, A. L., Bowers, J. E., Chandler, J. M., and Paterson, A. H. (2005). Crop-to-weed introgression has impacted allelic composition of

johnsongrass populations with and without recent exposure to cultivated sorghum. *Mol. Ecol.* 14 (7), 2143–2154. doi: 10.1111/j.1365-294X.2005.02579.x

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., et al. (2022). Author correction: Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 23 (1), 76. doi: 10.1186/s13059-022-02645-7

Presotto, A., Fernández-Moroni, I., Poverene, M., and Cantamutto, M. (2011). Sunflower crop-wild hybrids: Identification and risks. *Crop Prot* 30 (6), 611–616. doi: 10.1016/j.cropro.2011.02.022

Qi, X., Liu, Y., Vigueira, C. C., Young, N. D., Caicedo, A. L., Jia, Y., et al. (2015). More than one way to evolve a weed: parallel evolution of US weedy rice through independent genetic mechanisms. *Mol. Ecol.* 24 (13), 3329–3344. doi: 10.1111/mec.13256

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184 (13), 3542–3558.e3516. doi: 10.1016/j.cell.2021.04.046

Qiu, J., Jia, L., Wu, D. Y., Weng, X. F., Chen, L. J., Sun, J., et al. (2020). Diverse genetic mechanisms underlie worldwide convergent rice feralization. *Genome Biol.* 21 (1), 70. doi: 10.1186/s13059-020-01980-x

Qiu, J., Zhou, Y. J., Mao, L. F., Ye, C. Y., Wang, W. D., Zhang, J. P., et al. (2017). Genomic variation associated with local adaptation of weedy rice during de-domestication. *Nat. Commun.* 8, 15323. doi: 10.1038/ncomms15323

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., et al. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* 15 (6), 461–468. doi: 10.1038/s41592-018-0001-7

Shang, L., Li, X., He, H., Yuan, Q., Song, Y., Wei, Z., et al. (2022). A super pan-genomic landscape of rice. *Cell Res.* 32 (10), 878–896. doi: 10.1038/s41422-022-00685-z

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 (19), 3210–3212. doi: 10.1093/bioinformatics/btv351

Siren, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 374 (6574), 1461. doi: 10.1126/science.abg8871

Sun, J., Ma, D. R., Tang, L., Zhao, M. H., Zhang, G. C., Wang, W. J., et al. (2019). Population genomic analysis and *De novo* assembly reveal the origin of weedy rice as an evolutionary game. *Mol. Plant* 12 (5), 632–647. doi: 10.1016/j.molp.2019.01.019

Sun, P., Zhang, W., Wang, Y., He, Q., Shu, F., Liu, H., et al. (2016). OsGRF4 controls grain shape, panicle length and seed shattering in rice. *J. Integr. Plant Biol.* 58 (10), 836–847. doi: 10.1111/jipb.12473

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557 (7703), 43–49. doi: 10.1038/s41586-018-0063-9

Wu, D. Y., Qiu, J., Sun, J., Song, B. K., Olsen, K. M., and Fan, L. J. (2022). Weedy rice, a hidden gold mine in the paddy field. *Mol. Plant* 15 (4), 566–568. doi: 10.1016/j.molp.2022.01.008

Xu, J. L., Wang, Y., Zhang, F., Wu, Y., Zheng, T. Q., Wang, Y. H., et al. (2015). SS1 (NAL1)- and SS2-mediated genetic networks underlying source-sink and yield traits in rice (Oryza sativa l.). *PloS One* 10 (7), e0132060. doi: 10.1371/journal.pone.0132060

Yoon, J., Cho, L. H., Antt, H. W., Koh, H. J., and An, G. (2017). KNOX protein OSH15 induces grain shattering by repressing lignin biosynthesis genes. *Plant Physiol.* 174 (1), 312–325. doi: 10.1104/pp.17.00298

Zeng, X., Guo, Y., Xu, Q., Mascher, M., Guo, G., Li, S., et al. (2018). Origin and evolution of qingke barley in Tibet. *Nat. Commun.* 9 (1), 5433. doi: 10.1038/s41467-018-07920-5

Zhang, F., Xue, H., Dong, X., Li, M., Zheng, X., Li, Z., et al. (2022). Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.* 32 (5), 853–863. doi: 10.1101/gr.276015.121

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50 (2), 278–284. doi: 10.1038/s41588-018-0041-z