Chromosome-level genome assembly of Dongxiang wild rice (*Oryza rufipogon*) provides insights into resistance to disease and freezing

Zhiqun Que, Qineng Lu and Chunxiu Shen*

Jiangxi Key Laboratory of Crop Growth and Development Regulation, College of Life Sciences, Resources and Environment Sciences, Yichun University, Yichun, China

Dongxiang wild rice (DXWR, *Oryza rufipogon* Griff.) belongs to common wild rice *O. rufipogon*, which is the well-known ancestral progenitor of cultivated rice, possessing important gene resources for rice breeding. However, the distribution of DXWR is decreasing rapidly, and no reference genome has been published to date. In this study, we constructed a chromosome-level reference genome of DXWR by Oxford Nanopore Technology (ONT) and High-through chromosome conformation capture (Hi-C). A total of 58.41 Gb clean data from ONT were *de novo* assembled into 231 contigs with the total length of 413.46 Mb and N50 length of 5.18 Mb. These contigs were clustered and ordered into 12 pseudo-chromosomes covering about 97.39% assembly with Hi-C data, with a scaffold N50 length of 33.47 Mb. Moreover, 54.10% of the genome sequences were identified as repeat sequences. 33,862 (94.21%) genes were functionally annotated from a total of predicted 35,942 protein-coding sequences. Compared with other species of *Oryza* genus, the genes related to disease and cold resistance in DXWR had undergone a large-scale expansion, which may be one of the reasons for the stronger disease resistance and cold resistance of DXWR. Comparative transcriptome analysis also determined a list of differentially expressed genes under normal and cold treatment, which supported DXWR as a cold-tolerant variety. The collinearity between DXWR and cultivated rice was high, but there were still some significant structural variations, including a specific inversion on chromosome 11, which may be related to the differentiation of DXWR. The high-quality chromosome-level reference genome of DXWR assembled in this study will become a valuable resource for rice molecular breeding and genetic research in the future.

KEYWORDS

Oryza rufipogon, wild rice, *de novo* genome assembly, stress resistance, structural variations

Introduction

Rice is a main crop consumed by half the world's population (Elert, 2014). As the population explosion intensifies and the climate change challenges, the contradiction between large population and food supply is severe day by day, so improving crop yield is urgently needed. In the world, Asian rice *O. sativa* and African rice *Oryza glabarrima* were the common cultivated rice. During the long-term human selection, desirable agronomic traits including bigger seeds, lodging-resistant, and high yield were bred. However, the continuous selection also resulted in decreasing genetic diversity (Sun et al., 2001b; Meyer and Purugganan, 2013), which may be the inadequacy of facing climate change. Fortunately, there were more than twenty wild rice distributed in the world, which displayed abundant geographical, morphological, physiological and genetic diversity (Vaughan et al., 2008).

With the application of sequencing technologies, genome sequences of many species have been decoded. The cultivated rice *O. sativa* was the first crop to assemble the genome sequence, which supplied a useful foundation for functional study and breeding improvement (Yu et al., 2002). Next, more and more cultivated and wild rice were assembled using short-read sequencing and long-read sequencing (Chen et al., 2013; Wang et al., 2014; Wang et al., 2018; Choi et al., 2020; Kou et al., 2020; Li et al., 2021; Panibe et al., 2021; Qin et al., 2021; Zhang et al., 2022), which was helpful for comparative genomics, evolution analysis and variation identification. Genomic variations, including single nucleotide polymorphisms (SNPs, = 1 bp), insertion-deletion (InDels, ≤ 50 bp), and structural variations (SVs > 50 bp), are an important source of genetic diversity in species (Huang et al., 2012; Hollister, 2014; Sun et al., 2018b). Among them, structural variation, including insertion, deletion, inversion, copy number variation and some more complex variants, is a major driving force for the evolution of species, and plays a crucial biological role in rice phenotypic variation (Sun et al., 2018b; Yang et al., 2019a; Ho et al., 2020; Hollox et al., 2022).

The *Oryza* genus includes cultivated rice and wild rice species, Cultivated rice (two species *O. sativa* and *O. glaberrima*) is domesticated from wild rice species (22 species, including *O. rufipogon*, etc.), and during the domestication process, the diversity of cultivated rice morphological traits is reduced by 40% compared to wild rice species (Xie et al., 2009a; Wang et al., 2014). In addition, the domestication process of rice leads to the loss of several genes associated with biotic and abiotic stress (Xie et al., 2009a). In recent years, many rice disease resistance and stress tolerance genes have been found in wild rice, including genes associated with cold tolerance, insect resistance and other resistance genes (Sun et al., 2001a; Zhang et al., 2006; Huang et al., 2012; Mao et al., 2015; Zhao et al., 2016). Wild rice resources in China are very abundant and widely distributed in central and southern China. Dongxiang wild rice (hereafter DXWR), a wild rice *O. rufipogon* discovered in 1978 in Dongxiang County, Jiangxi Province of China, is thought to be the northernmost distribution (28°14′N latitude and 116°30′E longitude) of any wild rice species (N 28°14′) (Xie et al., 2009b). DXWR is rich in genetic diversity and is a potential source of many genes associated with high yield, hardiness and drought resistance, disease and insect resistance, and cytoplasmic male sterility. The distribution of DXWR was sharply reduced in recent decades. Therefore, DXWR was classified by the Chinese government as the second class of wild relative to food crop for protection. Despite its importance, no reference genome for DXWR has been published to date.

In this study, a high-quality chromosome-level DXWR genome using Nanopore long-read sequencing technology and Hi-C technology was obtained. Comparative genomics and structural variation revealed the reason of DXWR characters, which showed strong resistance to disease and cold. Our reference genome will lay a solid foundation for the molecular breeding of cultivated rice, and develop improved phenotypic rice with high yield, disease resistance and stress resistance.

Materials and methods

Plant materials and high-throughput sequencing

DXWR sequenced in this study (*O. rufipogon* Griff.) was planted in the greenhouse of Jiangxi Key Laboratory of Crop Growth and Development Regulation with normal growth conditions. Young leaves (2 cm × 0.3 cm) were collected and immediately frozen in liquid nitrogen, then stored at −80°C. The total genomic DNA from young leaves was extracted using CTAB method (Doyle, 1987). For Illumina DNA paired end (PE) sequencing, library with insert size of 400 bp was constructed by Illumina TruSeq Nano DNA Library Prep Kit and sequenced on Navoseq 6000 instrument (Illumina, San Diego, United States). For Nanopore sequencing, approximately 10 µg of gDNA was size-selected (10–50 kb) and processed using the Ligation sequencing 1D kit (SQK-LSK109, ONT, United Kingdom) according to the manufacturer's instructions to construct a Nanopore library, and then the library was sequenced on a PromethION sequencer (ONT, United Kingdom) at the Genome Center of Nextomics (Wuhan, China). For Hi-C experiment, the young leaves were fixed with 1% formaldehyde to induce cross-linking (Sigma), and then were lysed and formed the cohesive ends by restriction endonuclease DPN II (NEB). The digested DNA was blunt-ended by filling nucleotides by Klenow enzyme (NEB) with biotin-14-dATP (Invitrogen), then ligated by T4 DNA ligase (NEB). After incubating overnight to reverse cross-links, the ligated DNA was sheared into 300- to 600-bp fragments. The

DNA fragments were blunt-end repaired and A-tailed, followed by purification through biotin–streptavidin-mediated pulldown. Finally, Hi-C library was sequenced on Illumina NovaSeq 6000 platform.

Estimation of genome size and genome assembly

The Illumina PE reads were filtered and used to estimate the genome size and heterozygosity. K-mers were counted with Jellyfish (Kingsford, 2011), and then analyzed with skew normal distribution model and negative binomial model by FindGSE (Sun et al., 2018a) and GenomeScope (Vurture et al., 2017), respectively.

The Nanopore reads with mean quality score more than seven were retained and corrected by NextDenovo (https://github.com/Nextomics/NextDenovo) with specific parameters (read_cutoff = 1k, seed_cutoff = 42k). The corrected reads were assembled into contigs by smartdenvo (-k 17 -J 4000 -d dmo) (Liu et al., 2020a). To acquire more accurate genome, three rounds of correction were performed to the assembled contigs using Racon (Vaser et al., 2017) with nanopore long reads and another four rounds to the corrected genome were applied using NextPolish (Hu et al., 2020) with Illumina short reads.

The qualified Hi-C reads were aligned to the draft genome obtained from the previous step using bowtie2 (v2.3.2) with end-to-end model (-very-sensitive -L 30) (Langmead and Salzberg, 2012). Only the reads that both ends could be uniquely mapped to the draft genome were used in further analysis. Then LACHESIS 14 (https://github.com/shendurelab/LACHESIS) (Burton et al., 2013) according to the agglomerative hierarchical clustering algorithm was used to cluster contigs. The cross-linked maps were visualized and manually checked using Juicebox. Benchmarking Universal Single-Copy Orthologs (BUSCO v3.0.1) (Seppey et al., 2019) and Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra et al., 2007) was used to evaluate the completeness of the assembled genome. Moreover, Illumina PE reads were mapped to the assembled genome using BWA to assess the accuracy. Then single-nucleotide polymorphisms (SNPs) and Indels were called and filtered using SAMtools and bcftools (Li et al., 2009).

Repeat and non-coding RNA identification

The repeat sequence can be classified as Tandem repeats (TRs) and transposable elements (TEs). TRs were annotated using GMATA (Wang and Wang, 2016) and Tandem Repeats Finder (TRF) (Benson, 1999). A repeat library for DXWR was constructed with the combination of TE.lib, RepMod.lib, and Repbase (Jurka et al., 2005). TE.lib was generated using MITE-hunter (Han and Wessler, 2010), LTR_finder (Xu and Wang,

2007), LTRharverst, and LTR_retriver (Ou and Jiang, 2018); RepMod.lib, a *de novo* repeat library, was generated using RepeatModeler. Then TEs were identified using RepeatMasker (v4.0.6) based on the combined repeat library (Bedell et al., 2000). Finally results of TRs and TEs were merged together.

Additionally, non-coding RNAs were also identified. snRNA and miRNA were obtained using Infernal based on the Rfam (v11.0) database (Griffiths-Jones et al., 2005). rRNA and tRNA were detected by BLAST and tRNAscan-SE (v1.3.1). Then the rRNA and subunits were predicted by RNAmmer (v1.2) (Lagesen et al., 2007).

Gene prediction and annotation

Gene models were constructed by three methods, *ab initio* prediction, homology-based prediction and RNA-seq-assisted prediction. For the *ab initio* prediction, Augustus (v3.3.1) was used for the *de novo*-based gene prediction with default parameters (Stanke et al., 2006). Meanwhile, proteins of five species (*Oryza brachyantha*, *O. rufipogon*, *Oryza longistaminata*, *Zea mays*, and *Setaria italica*) were used for homology-based prediction through GeMoMa (v1.5.3) with default parameters (Keilwagen et al., 2019). Then, PASA (v2.0.2) was used for the RNAseq-based method of gene prediction, and the RNAseq data was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73181, SRA number, SRP063832) (Haas et al., 2003). Finally, the results from the three approaches were integrated using EVidenceModeler (EVM v1.1.1) to get the raw gene set (Haas et al., 2008). To obtain precise gene set, genes including transposable elements were filtered with TransposonPSI software (http://transposonpsi.sourceforge.net).

Functional annotation of predicted genes was obtained using two strategies. Firstly, those predicted protein sequences were aligned to SwissProt protein database using Blastp with the best match parameter. The involved pathways of predicted sequences were extracted from the KEGG Automatic Annotation Server (v2.1). Then the annotation of motifs and domains were performed using InterProScan (v5.32–71.0) to search against opening databases including Pfam, ProDom, PRINTS, PANTHER, SMRT, and PROSITE (Blum et al., 2021).

Whole genome duplication analysis

To analyze the WGD events among DXWR, *O. brachyantha*, *Z. mays*, *O. sativa*, *Triticum aestivum* and *Arabidopsis thaliana*, their protein sequences were aligned against themselves with Blastp (E-value ≤ 1e−10) to acquire conserved paralogs in each species. Then, the respective collinear blocks of these species were identified with MCScanX. Finally, potential WGD events in each genome were evaluated based on their 4DTv (Kimura, 1980) and Ks distribution (Blanc and Wolfe, 2004).

Evolution analyses

The nucleotide and amino acid sequences of nine species (*O. brachyantha*, *O. rufipogon*, *Z. mays*, *S. italica*, *O. sativa*, *T. aestivum*, *Brachypodium distachyon*, *Oryza meyeriana var. Granulata* and *A. thaliana*) were downloaded. All-to-all BLASTP with an E-value threshold of 1e-5 was applied to determine the similarities between protein sequences for all the species, and genes were classified into orthologues, paralogues and single copy orthologues (only one gene in each species) using OrthoMCL (v2.0.9) (Li et al., 2003). For the genes of unique family, the GO and KEGG enrichment analysis were performed to reveal the function of these unique genes.

Molecular phylogenetic analysis was conducted using all the single copy orthologues genes, and each gene family for multiple sequence alignment used Mafft (Katoh and Standley, 2013) and curated the alignments with Gblocks v0.91b (Castresana, 2000). The phylogenetic tree was built based on the PROTGAMMAAUTO model and a bootstrap of 1,000 by RAxML (v 8.2.11) (Stamatakis, 2006). A. thaliana was set as the outgroup. MCMCTREE in PAML v4.9e was used to estimate the divergence times (Yang, 2007). Two fossil calibration times were obtained from the TimeTree database (http://www.timetree.org/), including divergence times of 147.97–172.96 Mya and 9. 89–21.38 Mya.

Comparative transcriptome analysis

Transcriptome datasets under normal and cold treatment were downloaded from NCBI (accessions:SRP026336) (Shen et al., 2014). NGSQC Toolkit (Patel and Jain, 2012) was used to remove the adapter sequences and low quality sequence reads. Clean reads were aligned with the reference genome DXWR by Hisat2 v2.0.5 (Kim et al., 2015). FeatureCount (Liao et al., 2014) was used to calculate the read count. Differentially expressed genes (DEGs) was identified with the thresholds $|log2(FoldChange)| > 1$ and padj <0.005 by edgeR.

SV calling and collinear analysis

To investigate the structure variations among the wide and cultivated rice, minimap2 (Li, 2021) was selected for alignment and Smartie-sv pipeline (https://github.com/zeeev/smartiesv) was used to call SVs (Kronenberg et al., 2018). Collinear analysis among different genomes was analyzed using MCScanX (Wang et al., 2012). For verification with single chromosome, MUMmer (Kurtz et al., 2004) was used to display the detail information.

Results

Genome sequencing and assembly

To obtain a high-quality genome assembly for DXWR (Figure 1), three methods including Illumina short read sequencing, Nanopore long read sequencing and Hi-C chromosome conformation capture were used. After filtering, a total of 18.65 Gb clean PE reads (~45x) were yielded for genome survey and correction, 58.41 Gb Nanopore long reads (~140x) with reads N50 of 33.36 kb were obtained for genome assembly, and 44.58 Gb clean Hi-C reads (~108x) were used for chromosome construction (Supplementary Table S1). The genome size was surveyed with FindGSE and GenomeScope based on Illumina PE reads, and the predicted size ranged from 374.82 to 421.16 Mb. The heterozygous and repeat rate were 0.5% and 35%, respectively (Supplementary Figure S1; Supplementary Table S2).

For genome assembly, the corrected nanopore reads were assembled into 233 contigs with 414.59Mb, and the contig N50 size was 5.17 Mb. Further, the polished genome sequences were aligned to the NT database, and two contigs with a total of 1,126,901 bp were filtered out as mitochondrion and chloroplast genomes. Therefore, the final assembly for nuclear genome was 413.46 Mb with 231 contigs and the contig N50 was 5.18 Mb (Table 1), which was very close to the predicted genome size.

After mapping, 62,511,242 valid paired reads were used for Hi-C scaffolding analysis, and the assembled 231 contigs were clustered into 12 groups, which were further ordered and oriented into chromosomes. And 97.39% (402,670,344 bp) of the total contig bases (413,462,892 bp) were reliably anchored to the 12 chromosomes (Figure 2A). Finally, the nuclear genome size was 413,480,492 bp with a scaffold N50 of 33.49 Mb (Figure 2B; Table 1).

To evaluate the completeness genome of the nuclear genome, BUSCO v3.0.1 was performed by using embryophyta_odb10 database with default parameters to search single-copy orthologs genes. Approximately 98.25% of the orthologs genes were found in the assembly (Figure 2C; Supplementary Table S3). Meanwhile, Core Eukaryotic Genes Mapping Approach (CEGMA) was also used, and a total of 245 core genes were identified, which was 98.79% of the eukaryotic core genes (the complete set was 248) (Figure 2D). These results showed that the genome assembly of DXWR was highly complete and robust. Moreover, the accuracy was checked by the SNPs and Indels which were detected from alignment with the nuclear genome. Finally, A total of 3,269 homozygous SNPs (0.000233% of the assembled genome) and 964 homozygous Indels (0.000343% of the assembled genome) were identified with more than ×10 sequencing depth. The accuracy of the assembled genome was up to 99.999%, which suggested the high accuracy of the assembly.
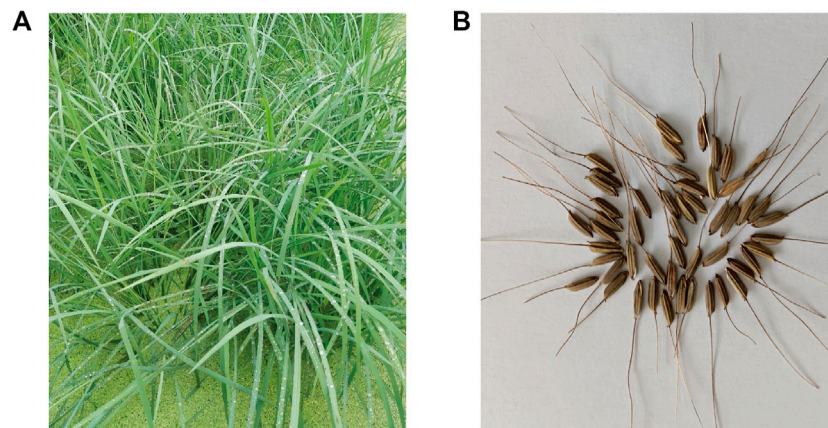
FIGURE 1
The plants **(A)** and seeds **(B)** of DXWR.

TABLE 1 Genome assembly statistics and post-processing of Dongxiang wild rice.

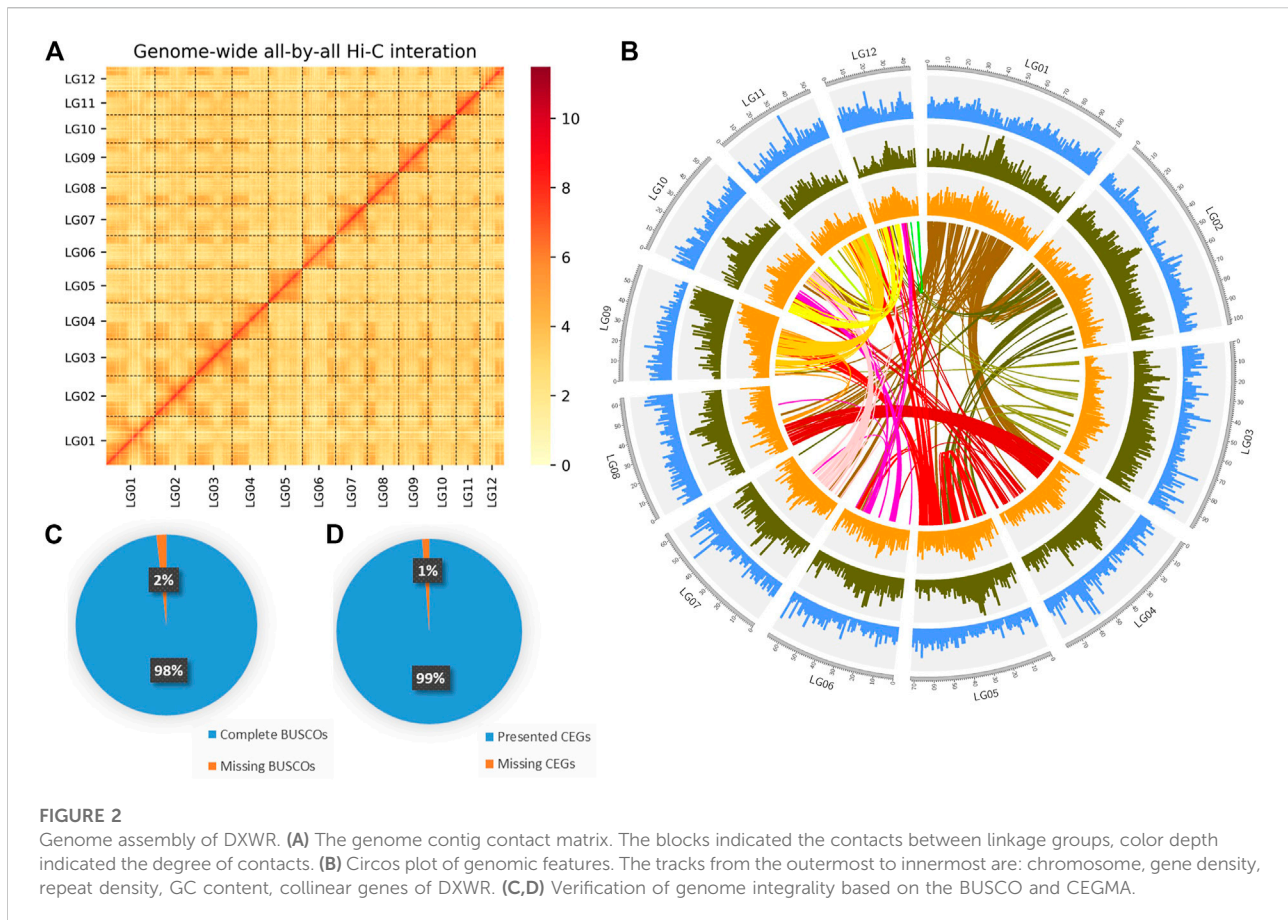| Genome assembly	Total assembly size of contigs/scaffolds (bp)	Number of contigs/scaffolds	N50 contig/scaffold length (bp)
Smartdenovo	405,743,318	233	5,062,292
Smartdenovo + Racon 3× + Nextpolish 4×	414,589,793	233	5,177,887
Nuclear genome	413,462,892	231	5,177,887
Nuclear genome + Hi-C	413,480,492	66	33,492,066

Genome annotation

Overall, 54.10% in the DXWR genome (223,689,067 bp) were identified as repetitive sequences. Among all the repeat elements, transposable elements (TEs) were the main types, accounting for 51.67% (213,636,330 bp). In terms of TEs, the dominant type was DNA transposon, but the longest total length was long terminal repeat (LTR), with the ratio up to 29.71% (122,836,029 bp) of the genome (Figures 3A,B; Supplementary Table S4). For non-coding RNAs, a total of 222 rRNA, 1,839 small RNA, and 683 tRNA were identified (Figure 3C; Supplementary Table S5). For protein-coding genes, the results of three methods were merged, and finally 35,942 protein-coding genes were predicted, with an average CDS length of 1,128.28 bp and an average exons number of 4.63 for each gene. To check the completeness of the genome annotation, BUSCO v3.0.1 was also used with default parameters. The results showed that 96.80% of the orthologs genes were found in DXWR annotated gene set (Supplementary Table S3), indicating most conserved genes of DXWR were predicted completely. To better understand the function of predicted genes, a variety of databases were used. In total, 33,862 (94.21%) genes were successfully assigned to at least one public functional database. Specifically, 90.94%, 65.84%, 42.18%, 40.3%, 26.46% of the total genes were mapped into the NR, Swissprot, KOG, GO and KEGG databases, respectively (Figure 3D; Supplementary Table S6). And 5,035 (14.01%) genes were annotated simultaneously by these five databases.

Comparative genomics and evolutionary analysis

The protein sequences of DXWR and nine species were selected for comparative genomics analysis, and 2,471 single copy orthologues genes were identified among them (Figure 4A). For unique genes, the number ranged from 1,396 to 16,480 in these species, and there were 4,741 unique genes in the genome of DXWR. Then GO and KEGG enrichment analysis was used to predict the function of these unique genes, and the results showed that hydrolase activity, polysaccharide binding and kinase activity were enriched terms in the molecular function category, whereas the terms including defense response and recognition of pollen in the biological process category were

FIGURE 2
Genome assembly of DXWR. **(A)** The genome contig contact matrix. The blocks indicated the contacts between linkage groups, color depth indicated the degree of contacts. **(B)** Circos plot of genomic features. The tracks from the outermost to innermost are: chromosome, gene density, repeat density, GC content, collinear genes of DXWR. **(C,D)** Verification of genome integrality based on the BUSCO and CEGMA.

highly enriched. Meanwhile, KEGG enrichment analysis showed that plant-pathogen interaction and plant hormone signal transduction pathways were the significantly enriched ones. It suggested that DXWR could resist against various biotic and abiotic stresses, which would be an invaluable gene pool for the genetic improvement of modern rice cultivars.
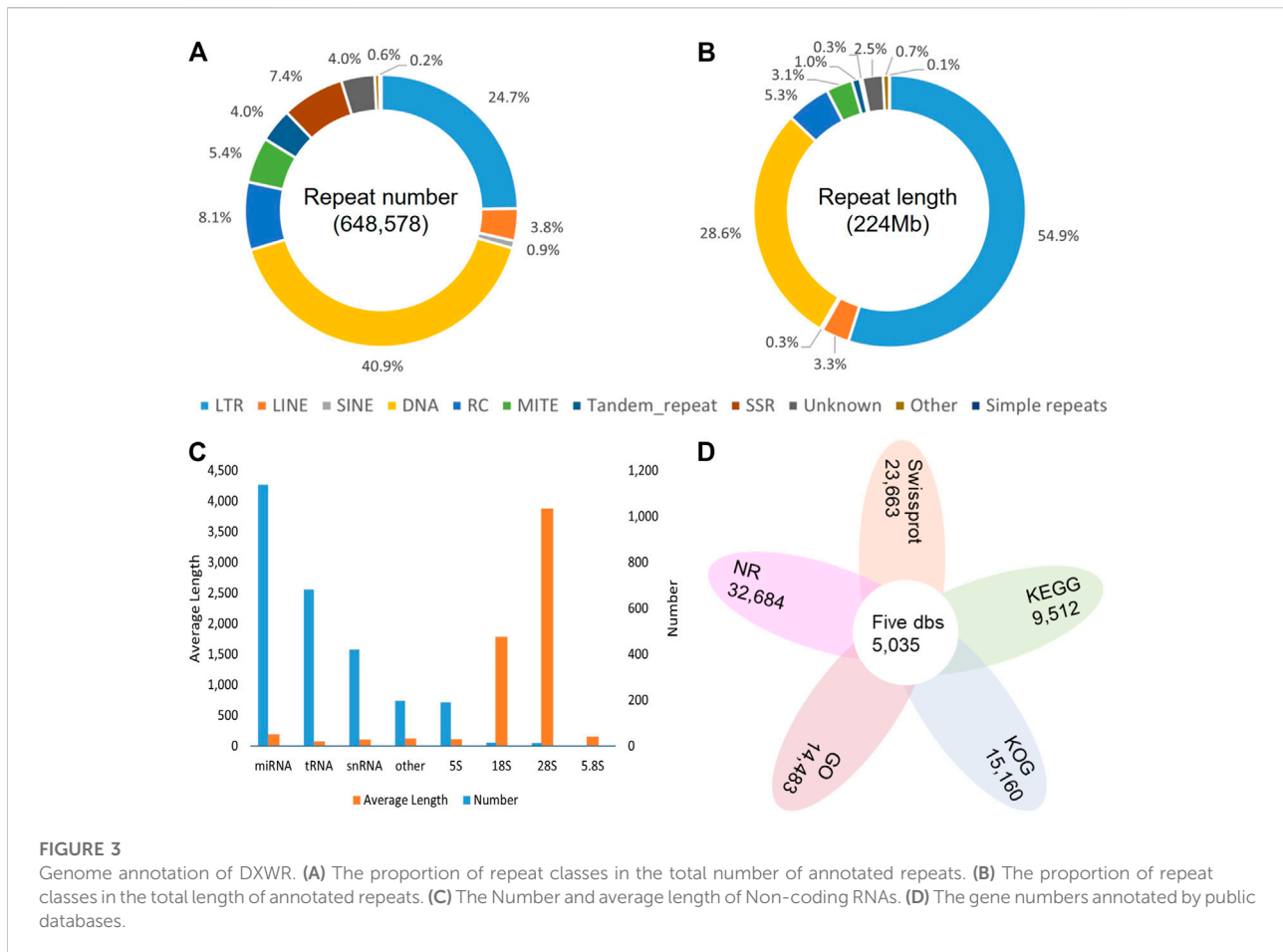
For evolution analysis, 2,471 single copy orthologues genes were used to construct the phylogenetic tree with *Arabidopsis* as the outgroup. It was obvious that *O. sativa* and W1943 were the most closely relative of DXWR. According to the gene families and phylogenetic tree (Figure 4B), 2,355 expanded gene families and 3,880 contracted gene families were identified in the genome of DXWR. The KEGG enrichment analysis of the expanded gene families revealed that plant-pathogen interaction, phenylpropanoid biosynthesis, starch and sucrose metabolism were the most enriched pathways (Figure 4C). As to contracted gene families, response to oxidative stress, peroxidase activity and aminoacyl-tRNA ligase activity terms were enriched.

Whole Genome Duplication is an important event in the history of biological evolution, which has great significance in the origin of species and the expansion of genomes, therefore, we investigated the WGD event in DXWR by comparing with other

five symbolic species. And the results of 4DTv and Ks distribution (Supplementary Table S7) suggested that DXWR experienced two recent WGD events, just like its close relative *O. brachyantha* and *O. sativa* (Figures 4D,E).

Comparative analysis of differentially expressed genes

To investigate expression changes under cold stress in DXWR, RNA-seq data from normal and cold treatment was compared. In total, 1801 DEGs were determined with the thresholds |log2(FoldChange)| > 1 and padj <0.005, among them, 1,072 DEGs were up-regulated and 729 DEGs were down-regulated. The KEGG enrichment of DEGs showed that they gathered in the process of photosynthesis, carbon metabolism, glyoxylate and dicarboxylate metabolism, biosynthesis of amino acids, pentose phosphate pathway, carotenoid biosynthesis and linoleic acid metabolism. Moreover, 43 DEGs were overlapped with contracted genes, while 340 DEGs were determined in expanded gene families of DXWR (Supplementary Table S8), which would be candidate

FIGURE 3
Genome annotation of DXWR. **(A)** The proportion of repeat classes in the total number of annotated repeats. **(B)** The proportion of repeat classes in the total length of annotated repeats. **(C)** The Number and average length of Non-coding RNAs. **(D)** The gene numbers annotated by public databases.

genes for DXWR to enhance the cold resistance. Actually, there were a number of genes that had been verified as cold-response genes, like potassium transporter 1, sugar transport protein MST6-like, calmodulin-like protein 5, glycerophosphodiester phosphodiesterase GDPDL3-like, polyol transporter 5, CBL-interacting protein kinase 5, protein NRT1/PTR FAMILY 6.2 and so on.

Structural variation and collinear analysis

SVs are a major source of genetic diversity (Hollister, 2014; Wang et al., 2018; Ho et al., 2020). In order to understand the structural variation of DXWR, we conducted a collinearity analysis of ten species of *Oryza* including DXWR (Figure 5A). The genome sequences of the *Oryza* genus had relatively high collinearity. Compared with the other nine *Oryza* species, DXWR and Nipponbare had a nearly 6 Mb inversion on chromosome 6 (Figure 5A).

Using wild rice DXWR as a reference genome, we aligned cultivated rice Nipponbare (*O. sativa*) and wild rice W1943 (*O. rufipogon*) genomes with DXWR genomes, respectively. As a

result, DXWR had more collinear regions and longer collinear lengths with W1943 (Figure 4B). Insertion and deletion accounted for the majority of structural variants, whereas inversion proportion was the least one (Figure 5B). At the same time, we found SNPs and SVs between DXWR and W1943 were less than those of DXWR and Nipponbare, which was consistent with the evolutionary relationship of the three, that is, DXWR and W1943 were more closely related, and more distantly related to cultivated rice (Figure 4B).

Chromosomal inversion is a structural variation that often contributes significantly to evolution, and its appearance is often related to biological processes such as biological adaptive phenotype and differentiation (Arostegui et al., 2019; Wellenreuther et al., 2019). Compared with the cultivated rice Nipponbare, it was discovered that there are 154 inversion regions in DXWR, of which 105 inversion fragments contained 1,061 genes (Supplementary Table S9). These genes were unevenly distributed on 12 chromosomes, among which chromosome 6 and chromosome 11 had the most genes, 284 and 268 respectively. GO enrichment analysis of these 1,061 genes revealed that these genes were associated with denfense response to fungus (Figure 5C), while KEGG enrichment analysis found
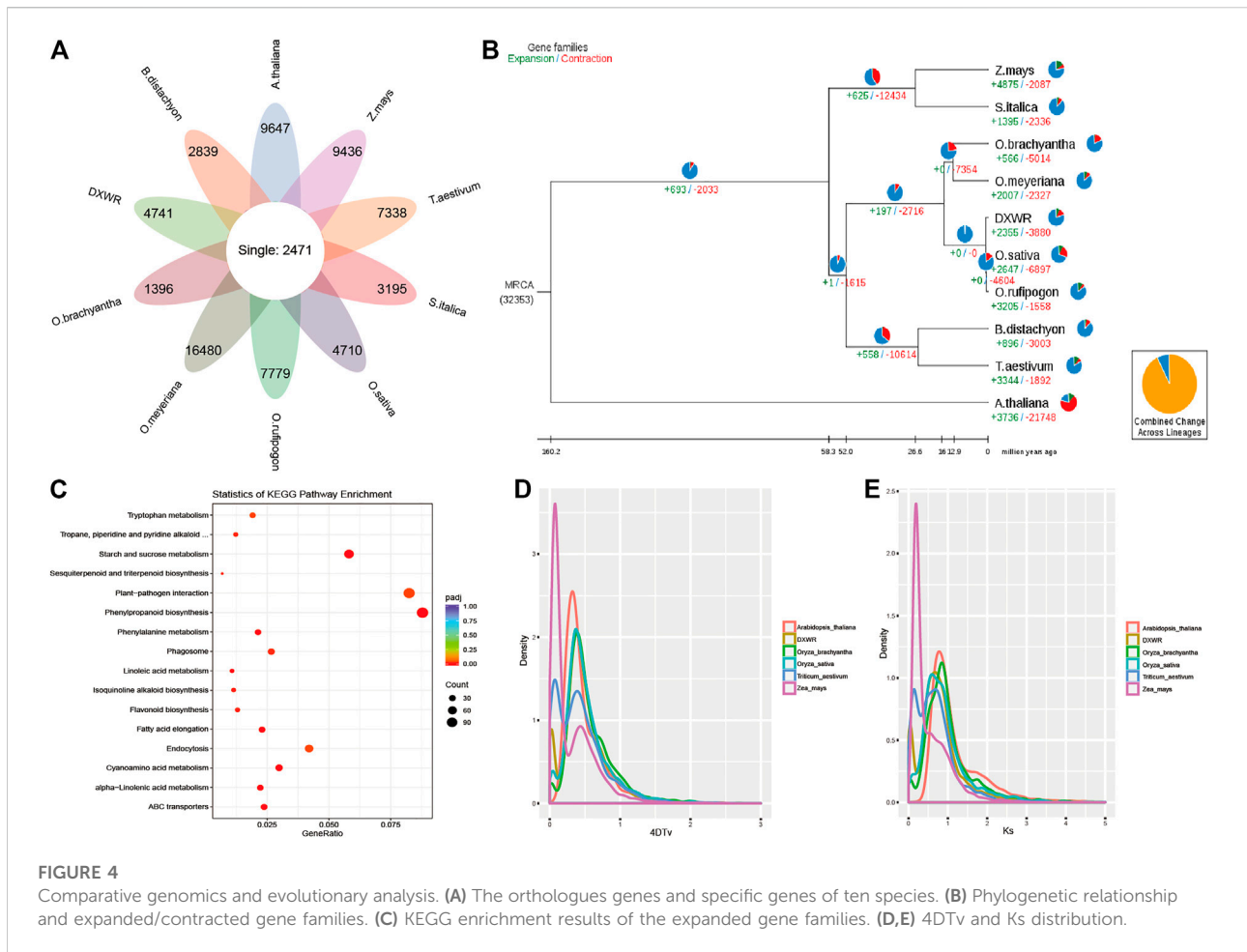
FIGURE 4
Comparative genomics and evolutionary analysis. **(A)** The orthologues genes and specific genes of ten species. **(B)** Phylogenetic relationship and expanded/contracted gene families. **(C)** KEGG enrichment results of the expanded gene families. **(D,E)** 4DTv and Ks distribution.

that these genes were connected to fatty acid metabolism, such as fatty acid degradation and Biosynthesis of unsaturated fatty acids (Figure 5D). This was similar to the enrichment analysis results of DXWR expansion genes, which were also related to plant-pathogen interactions and fatty acid metabolism.
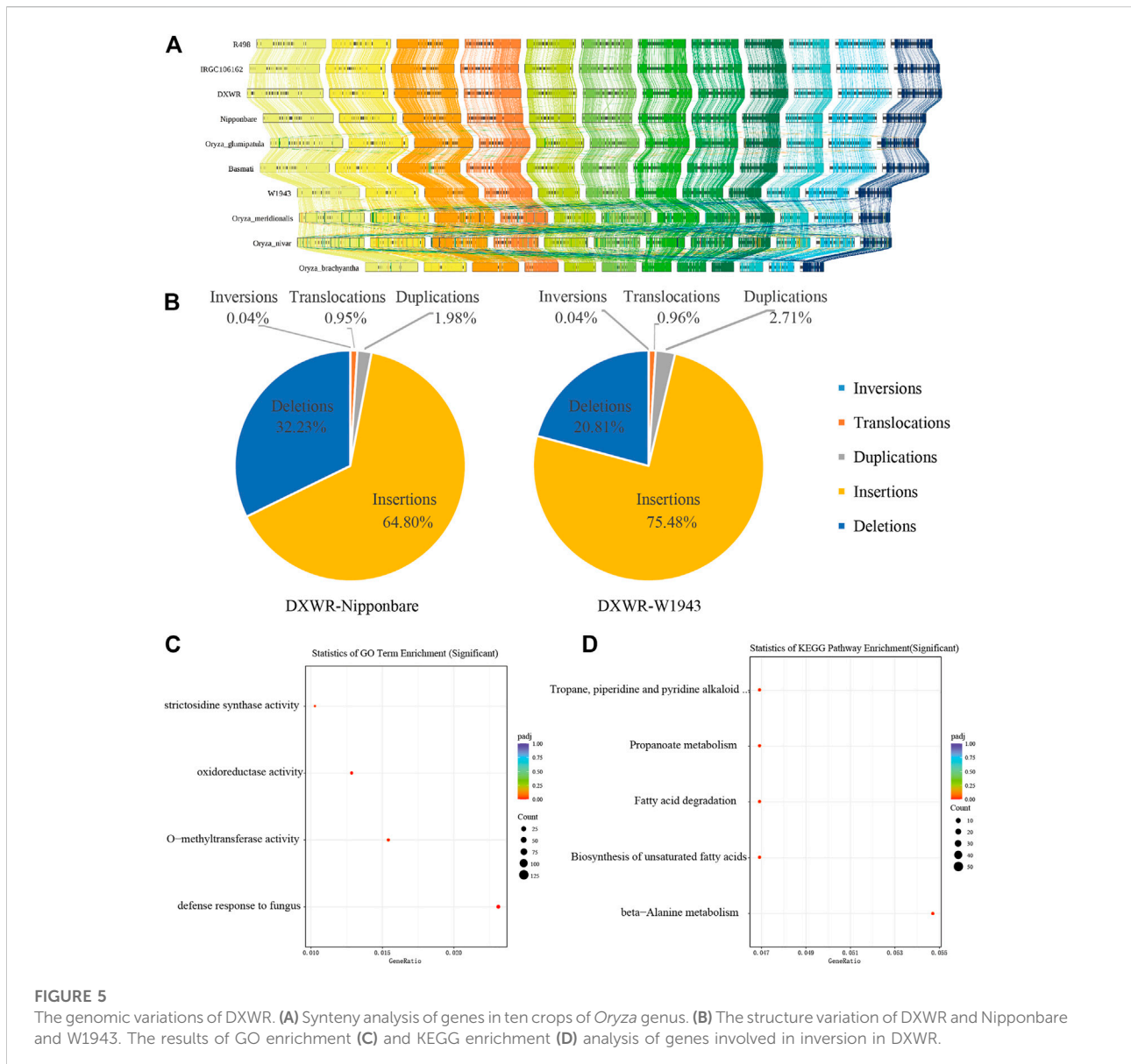
There are three inversions larger than 500 kb in chromosome 11 of DXWR (Chr11: 12,364,796–13,088,674, Chr11: 13,518,190–14,923,590, Chr11: 24,597,628–26,486,542). Compared with the other two wild rice (*O. rufipogon*) and other *Oryza* rice, these three inversions existed specifically in DXWR (Figure 5A), which may be one of the reasons for the differentiation of DXWR. We verified the authenticity of these three inversions using Hi-C data. The Hi-C data of DXWR was aligned to the genomes of W1943 and Nipponbare, respectively. The heat map signal can clearly find that the inversion occurred at the corresponding position of chromosome 11 of W1943 and Nipponbare, confirming the authenticity of the inversions of the corresponding position of chromosome 11 (Figure 6). The longest inversion (Chr11: 24,597,628–26,486,542) of the three inversions was 1,888,914 bp in length and involved a total of 176 genes. The enrichment analysis of these genes found that the main functions described defense

response to fungus, ADP binding and linoleic acid metabolism. There are six genes (evm.model.Contig 25.47–51 and evm.model.Contig 25.99) related to defense response to fungus in this inversion region, of which the first five are tandem genes, and these five genes are all related to defense response to bacterium. At the same time, unsaturated fatty acids can enhance the cold resistance of plants, so this inversion may be related to the disease and cold resistance of DXWR. Colinearity analysis of ten varieties from *Oryza* genera showed that the SV in Chr11 existed only in DXWR (Supplementary Figure S2). Then we compared this area in detail. We performed collinearity analysis on W1943 (Chr11: 20.88–23.27 Mb), DXWR (Chr11: 24.06–27.26 Mb) and IRGC106162 (Chr11: 24.14–23.26.83 Mb) and confirmed that this SV existed alone in DXWR and absent in W1943 and IRGC106162 (Supplementary Figure S3).

Discussion

Wild rice is usually regarded as abundant resource for holding genetic diversity, excellent agronomic traits and resistance against

FIGURE 5
The genomic variations of DXWR. **(A)** Synteny analysis of genes in ten crops of *Oryza* genus. **(B)** The structure variation of DXWR and Nipponbare and W1943. The results of GO enrichment **(C)** and KEGG enrichment **(D)** analysis of genes involved in inversion in DXWR.

stresses (Xu et al., 2011; Zhao et al., 2018). DXWR is a common wild rice located at the northernmost of *O. rufipogon* species. To better understand and exploit the characteristic of DXWR, sequencing the genome is a convenient and effective way, which will provide all the genetic information. In this study, a chromosome-level genome of DXWR was assembled using nanopore long reads and high-through chromatin conformation capture (Hi-C) technology, which could overcome issue of heterozygosity and high repeat rate. After primary assembly, more than 97% of the total contig bases (413,462,892 bp) were correctly anchored into 12 chromosomes. The final nuclear genome was 413.48 Mb with a scaffold N50 of 33.49 Mb, which was a little larger than other published rice genome, and not only DXWR, it was noted that generally the genome size of *O. rufipogon*

varieties was larger (Shang et al., 2022), which was caused by higher repeat rates, especially the transposon expansion. Moreover, both BUSCO and CEGMA assessment verified the completeness of the genome, and the high accuracy of the genome was also affirmed by alignment using short reads. For protein-coding genes, a high annotation rate was observed, up to 94.21% genes could be annotated by public functional databases. The high-quality genome and annotation could supply valuable resources for comparative genomics, evolution analysis and genetic breeding of rice.

Based on the high-quality genome assembly, DXWR was compared with other species to understand its character. According to the enrichment analysis of unique genes and expanded gene families, it was noted that plant-pathogen

FIGURE 6
The inversions between DXWR and Nipponbare and W1943 supported by Hi-C contact maps.
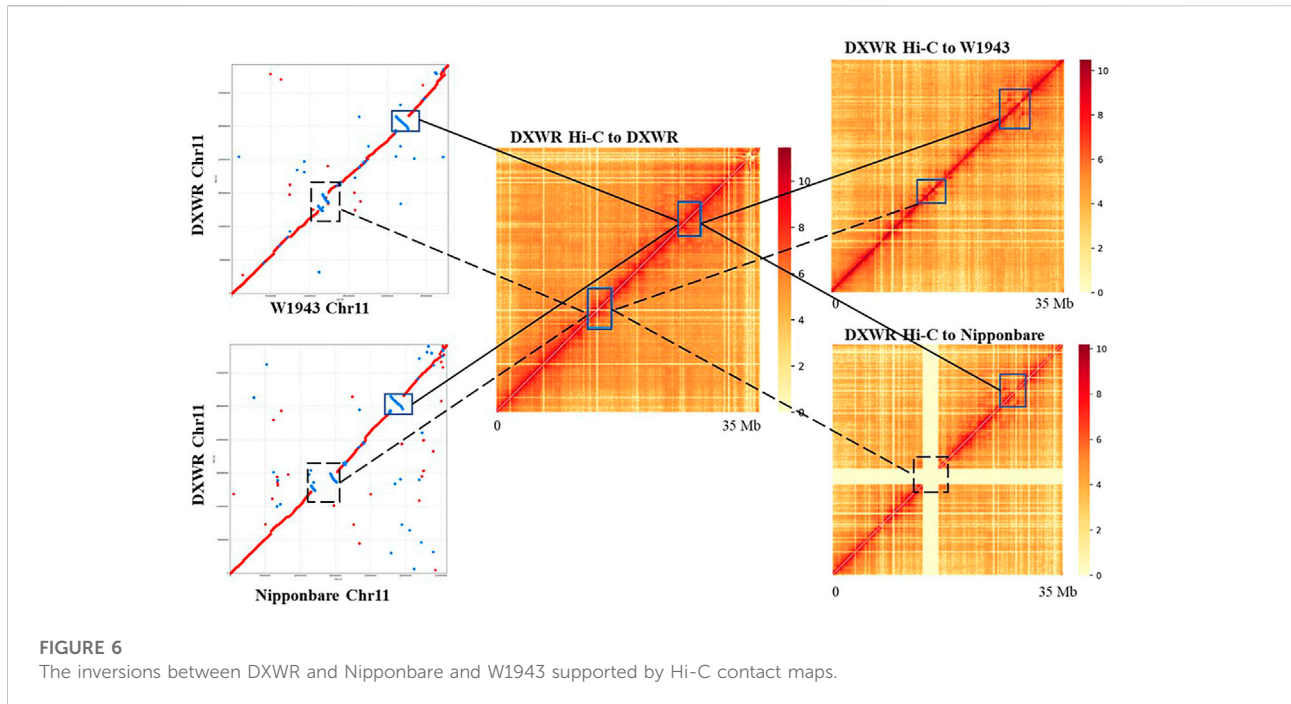
interaction, phenylpropanoid biosynthesis, starch and sucrose metabolism, and plant hormone signal transduction pathways were significantly enriched. The plant-pathogen interaction and defense response would help DXWR survive under biotic stress (Kushalappa and Gunnaiah, 2013). Phenylpropanoid which was a rich source of secondary metabolites in plants, together with plant hormone, played important roles in plant growth, development, and defense (Desmedt et al., 2021; Dong and Lin, 2021). Moreover, starch and sucrose metabolism was the vital pathway in the process of rice grain filling, which could finally affect the yield (Fan et al., 2019; Jiang et al., 2021; Mathan et al., 2021). All these information indicated that DXWR probably had superior disease-resistance and starch synthesis ability.

Like other stress-responsive traits, the cold resistance characteristics of plants are affected by multiple factors and are also controlled by genetics (Xu and Cai, 2014; Fang et al., 2017). As we know, cold resistance is related to the content of unsaturated fatty acids in plants. The increase of unsaturated fatty acids in membrane lipids will reduce the temperature of phase transition of membrane lipids, so the cold resistance of plants can be improved by increasing the degree of unsaturation of fatty acids (Khodakovskaya et al., 2006; Matteucci et al., 2011; Soria-Garci et al., 2019). Overexpression of the chloroplast omega-3 fatty acid desaturase gene (LeFAD7) in tomato (Lycopersicon esculentum Mill.) can increase the content of unsaturated fatty acids and enhance the resistance to low temperature stress (Liu et al., 2008). Overexpression chloroplast ω- 3 fatty acid desaturase gene can also enhance the cold tolerance of transgenic tobacco (Khodakovskaya et al., 2006). The increase of small molecular substances and soluble substances is one of the cold

resistance responses of plants, and plants with strong cold resistance will accumulate more soluble sugars (Wei et al., 2017; Zhao et al., 2019; Liu et al., 2021). These soluble sugars have a certain protective effect on preventing protein denaturation after dehydration, and intercellular sugars will alleviate low temperature damage by affecting the growth of ice crystals. Liang et al. (2021) reported that the increasing of soluble sugar in grapes can enhance the cold resistance of grapes. DXWR is the wild rice with the most northward distribution and the strongest cold resistance discovered so far. It can resist low temperature and cold wave for a long time in the seedling and heading stage. Analysis of its genome and comparative transcriptome under normal and cold treatment found that genes related to phenylpropanoid biosynthesis, linoleic acid metabolism and starch and sucrose metabolism were significantly expanded and differentially expressed in DXWR, which may promote the cold resistance of DXWR. The discovery of cold resistance-related genes in wild rice is of great significance for understanding the cold resistance mechanism of wild rice and cultivating strong cold resistance rice varieties.

SNPs have long been considered to be a significant component of genetic variation, but now there is increasing evidence that structural variation is also an important part of genetic variation (Alonge et al., 2020; Liu et al., 2020b; Qin et al., 2021). Structural variation may have the addition or deletion of DNA information, such as insertion, deletion, or duplication, but it may not lead to the increase or deletion of DNA information, such as inversion, transposition in an individual (Fu et al., 2016; Yang et al., 2019b). SVs are a key and pervasive force driving genetic diversity and contribute to important agronomic traits in

crops (Yang et al., 2019a; Yang et al., 2019b). In tomato, SV has been shown to relate with fruit size and flavor, disease response, and the plant's ability to detect pathogens (Jobson and Roberts, 2022). Copy number variants (CNVs), one important type of SV, has been shown to play an important role in the adaptive response of plants, by regulating development, and by increasing resistance to biotic and abiotic stresses (Francia et al., 2015). SVs in maize, sorghum and rice have been shown to be associated with plant disease resistance, and stress responsive genes in soybeans have also found related with SVs (Saxena et al., 2014). Published studies have shown that SVs play an important role in the domestication of rice from wild rice to cultivated rice, and peaks of SV divergence were concentrated in genes associated with domestication (Kou et al., 2020). The appearance of inversion is often related to evolution and biological processes, such as biological adaptive phenotypes and differentiation. In the chromosome 11 of DXWR, there was an inversion region close to 2 Mb, which is consistent with the results obtained by Mao et al. (2015) using SLAF marker. Compared with the other nine *Oryza* genus plants, this inversion was specific to DXWR and this region was related to 176 genes. The enrichment analysis showed that these genes were closely related to defense response to fungus and linoleic acid metabolism, which may be the reason for the differentiation and characteristics of DXWR.

Conclusion

In this research, a high-quality chromosome-level genome of wild rice DXWR was obtained using Nanopore sequencing and Hi-C technology, which could provide elaborate genomic information for evolution and genetic breeding. Here the comparative genomics and transcriptomics analysis had indicated that DXWR probably had superior disease and cold resistance. Moreover, when compared to other cultivated rice, DXWR exhibited certain distinct inversions. These inversions were also connected to defense response to biotic stress and may be responsible for the characteristics of DXWR. Overall, the genome assembly of DXWR would help us to better understand and utilize the characteristic of this wild rice, which is critical for future genetic breeding in this crop.

Data availability statement

This Whole Genome Shotgun project has been deposited at GenBank under the accession PRJNA641241. All sequencing data are available at NCBI Sequence Read Archive SRR12104676, SRR12102354 and SRR12102351. The genome assembly and annotation files are available at figshare https://doi.org/10.6084/m9.figshare.20072117.v1.

Author contributions

CS and ZQ designed the experiment and prepared the manuscript; CS acquired funding; CS, QL, and ZQ prepared the materials and performed the bioinformatics analysis. All authors have read and agreed to the published version of the manuscript.

Funding

Conflict of interest

The authors declare that this study received sequencing and technical support from Beijing GrandOmics Biotechnology Co., Ltd.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1029879/full#supplementary-material

SUPPLEMENTARY FIGURE S1
K-mer and coverage distribution of DXWR.

SUPPLEMENTARY FIGURE S2
The collinearity relationship between ten rice varieties in the Oryza genus.

SUPPLEMENTARY FIGURE S3
The collinearity relationship between W1943 (Chr11: 20.88−23.27 Mb), DXWR (Chr11: 24.06−27.26 Mb) and IRGC106162 (Chr11: 24.14−23.26.83 Mb).

References

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, 145–161. doi:10.1016/j.cell.2020.05.021

Arostegui, M. C., Quinn, T. P., Seeb, L. W., Seeb, J. E., and Mckinney, G. J. (2019). Retention of a chromosomal inversion from an anadromous ancestor provides the genetic basis for alternative freshwater ecotypes in rainbow trout. *Mol. Ecol.* 28, 1412–1427. doi:10.1111/mec.15037

Bedell, J. A., Korf, I., and Gish, W. (2000). MaskerAid: A performance enhancement to RepeatMasker. *Bioinformatics* 16, 1040–1041. doi:10.1093/bioinformatics/16.11.1040

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi:10.1093/nar/27.2.573

Blanc, G., and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678. doi:10.1105/tpc.021345

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi:10.1093/nar/gkaa977

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. doi:10.1038/nbt.2727

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi:10.1093/oxfordjournals.molbev.a026334

Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., et al. (2013). Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nat. Commun.* 4, 1595. doi:10.1038/ncomms2596

Choi, J. Y., Lye, Z. N., Groen, S. C., Dai, X., Rughani, P., Zaaijer, S., et al. (2020). Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biol.* 21, 21. doi:10.1186/s13059-020-1938-2

Desmedt, W., Jonckheere, W., Nguyen, V. H., Ameye, M., De Zutter, N., De Kock, K., et al. (2021). The phenylpropanoid pathway inhibitor piperonylic acid induces broad-spectrum pest and disease resistance in plants. *Plant Cell Environ.* 44, 3122–3139. doi:10.1111/pce.14119

Dong, N. Q., and Lin, H. X. (2021). Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. *J. Integr. Plant Biol.* 63, 180–209. doi:10.1111/jipb.13054

Doyle, J. J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19.

Elert, E. (2014). Rice by the numbers: A good grain. *Nature* 514, S50–S51. doi:10.1038/514a50a

Fan, C., Wang, G., Wang, Y., Zhang, R., Wang, Y., Feng, S., et al. (2019). Sucrose synthase enhances hull size and grain weight by regulating cell division and starch accumulation in transgenic rice. *Int. J. Mol. Sci.* 20, E4971. doi:10.3390/ijms20204971

Fang, C., Zhang, P., Jian, X., Chen, W., Lin, H., Li, Y., et al. (2017). Overexpression of *Lsi1* in cold-sensitive rice mediates transcriptional regulatory networks and enhances resistance to chilling stress. *Plant Sci.* 262, 115–126. doi:10.1016/j.plantsci.2017.06.002

Francia, E., Pecchioni, N., Policriti, A., and Scalabrin, S. (2015). "CNV and structural variation in plants: Prospects of NGS approaches," in *Advances in the understanding of biological Sciences using next generation sequencing (NGS) approaches*. Editors G. Sablok, S. Kumar, S. Ueno, J. Kuo, and C. Varotto (Cham: Springer International Publishing), 211–232.

Fu, D., Mason, A. S., Xiao, M., and Yan, H. (2016). Effects of genome structure variation, homeologous genes and repetitive DNA on polyploid crop research in the age of genomics. *Plant Sci.* 242, 37–46. doi:10.1016/j.plantsci.2015.09.017

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, D121–D124. doi:10.1093/nar/gki081

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr, Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi:10.1093/nar/gkg770

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, R7. doi:10.1186/gb-2008-9-1-r7

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. doi:10.1093/nar/gkq862

Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural variation in the sequencing era. *Nat. Rev. Genet.* 21, 171–189. doi:10.1038/s41576-019-0180-9

Hollister, J. D. (2014). Genomic variation in *Arabidopsis*: Tools and insights from next-generation sequencing. *Chromosome Res.* 22, 103–115. doi:10.1007/s10577-014-9420-1

Hollox, E. J., Zuccherato, L. W., and Tucci, S. (2022). Genome structural variation in human evolution. *Trends Genet.* 38, 45–58. doi:10.1016/j.tig.2021.06.015

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36, 2253–2255. doi:10.1093/bioinformatics/btz891

Huang, X., Kurata, N., Wei, X., Wang, Z. X., Wang, A., Zhao, Q., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497–501. doi:10.1038/nature11532

Jiang, Z., Chen, Q., Chen, L., Yang, H., Zhu, M., Ding, Y., et al. (2021). Efficiency of sucrose to starch metabolism is related to the initiation of inferior grain filling in large panicle rice. *Front. Plant Sci.* 12, 732867. doi:10.3389/fpls.2021.732867

Jobson, E., and Roberts, R. (2022). Genomic structural variation in tomato and its role in plant immunity. *Mol. Hortic.* 2, 7. doi:10.1186/s43897-022-00029-w

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. doi:10.1159/000084979

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010

Keilwagen, J., Hartung, F., and Grau, J. (2019). GeMoMa: Homology-Based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* 1962, 161–177. doi:10.1007/978-1-4939-9173-0_9

Khodakovskaya, M., Mcavoy, R., Peters, J., Wu, H., and Li, Y. (2006). Enhanced cold tolerance in transgenic tobacco expressing a chloroplast omega-3 fatty acid desaturase gene under the control of a cold-inducible promoter. *Planta* 223, 1090–1100. doi:10.1007/s00425-005-0161-4

Kim, D., Langmead, B., and Salzberg, S. L. (2015). Hisat: A fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi:10.1007/BF01731581

Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi:10.1093/bioinformatics/btr011

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

Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O. S., et al. (2018). High-resolution comparative analysis of great ape genomes. *Science* 360, eaar6343. doi:10.1126/science.aar6343

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi:10.1186/gb-2004-5-2-r12

Kushalappa, A. C., and Gunnaiah, R. (2013). Metabolo-proteomics to discover plant biotic stress resistance genes. *Trends Plant Sci.* 18, 522–531. doi:10.1016/j.tplants.2013.05.002

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., and Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35, 3100–3108. doi:10.1093/nar/gkm160

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 37, 4572–4574. doi:10.1093/bioinformatics/btab705

Li, K., Jiang, W., Hui, Y., Kong, M., Feng, L. Y., Gao, L. Z., et al. (2021). Gapless indica rice genome reveals synergistic contributions of active transposable elements

and segmental duplications to rice genome evolution. *Mol. Plant* 14, 1745–1756. doi:10.1016/j.molp.2021.06.017

Li, L., Stoeckert, C. J., Jr., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503

Liang, G., He, H., Nai, G., Feng, L., Li, Y., Zhou, Q., et al. (2021). Genome-wide identification of BAM genes in grapevine (*Vitis vinifera* L.) and ectopic expression of *VvBAM1* modulating soluble sugar levels to improve low-temperature tolerance in tomato. *BMC Plant Biol.* 21, 156. doi:10.1186/s12870-021-02916-8

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656

Liu, H., Wu, S., Li, A., and Ruan, J. (2020a). *SMARTdenovo: A de novo Assembler Using Long Noisy Reads*.

Liu, X., Chen, L., Shi, W., Xu, X., Li, Z., Liu, T., et al. (2021). Comparative transcriptome reveals distinct starch-sugar interconversion patterns in potato genotypes contrasting for cold-induced sweetening capacity. *Food Chem.* 334, 127550. doi:10.1016/j.foodchem.2020.127550

Liu, X. Y., Li, B., Yang, J. H., Sui, N., Yang, X. M., and Meng, Q. W. (2008). Overexpression of tomato chloroplast omega-3 fatty acid desaturase gene alleviates the photoinhibition of photosystems 2 and 1 under chilling stress. *Photosynthetica* 46, 185. doi:10.1007/s11099-008-0030-z

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

Mao, D., Yu, L., Chen, D., Li, L., Zhu, Y., Xiao, Y., et al. (2015). Multiple cold resistance loci confer the high cold tolerance adaptation of Dongxiang wild rice (*Oryza rufipogon*) to its high-latitude habitat. *Theor. Appl. Genet.* 128, 1359–1371. doi:10.1007/s00122-015-2511-3

Mathan, J., Singh, A., and Ranjan, A. (2021). Sucrose transport and metabolism control carbon partitioning between stem and grain in rice. *J. Exp. Bot.* 72, 4355–4372. doi:10.1093/jxb/erab066

Matteucci, M., D'angeli, S., Errico, S., Lamanna, R., Perrotta, G., and Altamura, M. M. (2011). Cold affects the transcription of fatty acid desaturases and oil quality in the fruit of Olea europaea L. genotypes with different cold hardiness. *J. Exp. Bot.* 62, 3403–3420. doi:10.1093/jxb/err013

Meyer, R. S., and Purugganan, M. D. (2013). Evolution of crop species: Genetics of domestication and diversification. *Nat. Rev. Genet.* 14, 840–852. doi:10.1038/nrg3605

Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi:10.1104/pp.17.01310

Panibe, J. P., Wang, L., Li, J., Li, M. Y., Lee, Y. C., Wang, C. S., et al. (2021). Chromosomal-level genome assembly of the semi-dwarf rice Taichung Native 1, an initiator of Green Revolution. *Genomics* 113, 2656–2674. doi:10.1016/j.ygeno.2021.06.006

Parra, G., Bradnam, K., and Korf, I. (2007). Cegma: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi:10.1093/bioinformatics/btm071

Patel, R. K., and Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS One* 7, e30619. doi:10.1371/journal.pone.0030619

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, 3542–3558.e16. doi:10.1016/j.cell.2021.04.046

Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant genomes. *Brief. Funct. Genomics* 13, 296–307. doi:10.1093/bfgp/elu016

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962, 227–245. doi:10.1007/978-1-4939-9173-0_14

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, 878–896. doi:10.1038/s41422-022-00685-z

Shen, C., Li, D., He, R., Fang, Z., Xia, Y., Gao, J., et al. (2014). Comparative transcriptome analysis of RNA-seq data for cold-tolerant and cold-sensitive rice genotypes under cold stress. *J. Plant Biol.* 57, 337–348. doi:10.1007/s12374-014-0183-1

Soria-Garci, A. I., Rubio, M. a. C., Lagunas, B., Li Pez-Gomolli, N. S., Luji, N. M., Di Az-Guerra, R. L., et al. (2019). Tissue distribution and specific contribution of *Arabidopsis* FAD7 and FAD8 plastid desaturases to the JA- and ABA-mediated cold stress or defense responses. *Plant Cell Physiol.* 60, 1025–1040. doi:10.1093/pcp/pcz017

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi:10.1093/bioinformatics/btl446

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). Augustus: Ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi:10.1093/nar/gkl200

Sun, C. Q., Wang, X. K., Li, Z. C., Yoshimura, A., and Iwata, N. (2001a). Comparison of the genetic diversity of common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (O. *sativa* L.) using RFLP markers. *Theor. Appl. Genet.* 102, 157–162. doi:10.1007/s001220051631

Sun, C. Q., Wang, X. K., Li, Z. C., Yoshimura, A., and Iwata, N. (2001b). Comparison of the genetic diversity of common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (O. *sativa* L.) using RFLP markers. *Theor. Appl. Genet.* 102, 157–162. doi:10.1007/s001220051631

Sun, H., Ding, J., Piednoel, M., and Schneeberger, K. (2018a). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* 34, 550–557. doi:10.1093/bioinformatics/btx637

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., et al. (2018b). Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* 50, 1289–1295. doi:10.1038/s41588-018-0182-0

Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 27, 737–746. doi:10.1101/gr.214270.116

Vaughan, D. A., Ge, S., Kaga, A., and Tomooka, N. (2008). "Phylogeny and biogeography of the genus Oryza," in *Rice biology in the genomics era*. Editors H.-Y. Hirano, Y. Sano, A. Hirai, and T. Sasaki (Berlin, Heidelberg: Springer Berlin Heidelberg), 219–234.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi:10.1093/bioinformatics/btx153

Wang, M., Yu, Y., Haberer, G., Marri, P. R., Fan, C., Goicoechea, J. L., et al. (2014). The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat. Genet.* 46, 982–988. doi:10.1038/ng.3044

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, 43–49. doi:10.1038/s41586-018-0063-9

Wang, X., and Wang, L. (2016). Gmata: An integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* 7, 1350. doi:10.3389/fpls.2016.01350

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi:10.1093/nar/gkr1293

Wei, Y., Xu, F., and Shao, X. (2017). Changes in soluble sugar metabolism in loquat fruit during different cold storage. *J. Food Sci. Technol.* 54, 1043–1051. doi:10.1007/s13197-017-2536-5

Wellenreuther, M., Merot, C., Berdan, E., and Bernatchez, L. (2019). Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.* 28, 1203–1209. doi:10.1111/mec.15066

Xie, J., Agrama, H. A., Kong, D., Zhuang, J., Hu, B., Wan, Y., et al. (2009a). Genetic diversity associated with conservation of endangered Dongxiang wild rice (*Oryza rufipogon*). *Genet. Resour. Crop Evol.* 57, 597–609. doi:10.1007/s10722-009-9498-z

Xie, J., Agrama, H. A., Kong, D., Zhuang, J., Hu, B., Wan, Y., et al. (2009b). Genetic diversity associated with conservation of endangered Dongxiang wild rice (*Oryza rufipogon*). *Genet. Resour. Crop Evol.* 57, 597–609. doi:10.1007/s10722-009-9498-z

Xu, P., and Cai, W. (2014). RAN1 is involved in plant cold resistance and development in rice (*Oryza sativa*). *J. Exp. Bot.* 65, 3277–3287. doi:10.1093/jxb/eru178

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., et al. (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30, 105–111. doi:10.1038/nbt.2050

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. doi:10.1093/nar/gkm286

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., et al. (2019a). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat. Genet.* 51, 1052–1059. doi:10.1038/s41588-019-0427-6

Yang, N., Wu, S., and Yan, J. (2019b). Structural variation in complex genome: Detection, integration and function. *Sci. China. Life Sci.* 62, 1098–1100. doi:10.1007/s11427-019-9664-4

Yang, Z. (2007). Paml 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* 296, 79–92. doi:10.1126/science.1068037

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, 853–863. doi:10.1101/gr.276015.121

Zhang, X., Zhou, S., Fu, Y., Su, Z., Wang, X., and Sun, C. (2006). Identification of a drought tolerant introgression line derived from Dongxiang common wild rice (*O. rufipogon* Griff.). *Plant Mol. Biol.* 62, 247–259. doi:10.1007/s11103-006-9018-x

Zhao, J., Qin, J. J., Song, Q., Sun, C. Q., and Liu, F. X. (2016). Combining QTL mapping and expression profile analysis to identify candidate genes of cold tolerance from Dongxiang common wild rice (*Oryza rufipogon* Griff.). *J. Integr. Agric.* 15, 1933–1943. doi:10.1016/s2095-3119(15)61214-x

Zhao, L., Yang, T., Xing, C., Dong, H., Qi, K., Gao, J., et al. (2019). The beta-amylase PbrBAM3 from pear (*Pyrus betulaefolia*) regulates soluble sugar accumulation and ROS homeostasis in response to cold stress. *Plant Sci.* 287, 110184. doi:10.1016/j.plantsci.2019.110184

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Publisher Correction: Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50, 1196. doi:10.1038/s41588-018-0136-6