frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 21 March 2023
DOI 10.3389/fpls.2023.1147946

OPEN ACCESS

EDITED BY
Yunpeng Cao,
Wuhan Botanical Garden (CAS), China

REVIEWED BY
Lin Zhang,
Hubei University of Chinese Medicine, China
Sajid Fiaz,
The University of Haripur, Pakistan

*CORRESPONDENCE
Jinping Guo
jinpguo@126.com
Baopeng Ding
dingbaopeng2006@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 19 January 2023
ACCEPTED 06 March 2023
PUBLISHED 21 March 2023

CITATION
Wang J, Hu H, Liang X, Tahir ul Qamar M, Zhang Y, Zhao J, Ren H, Yan X, Ding B and Guo J (2023) High-quality genome assembly and comparative genomic profiling of yellowhorn (*Xanthoceras sorbifolia*) revealed environmental adaptation footprints and seed oil contents variations.
Front. Plant Sci. 14:1147946.
doi: 10.3389/fpls.2023.1147946

High-quality genome assembly and comparative genomic profiling of yellowhorn (*Xanthoceras sorbifolia*) revealed environmental adaptation footprints and seed oil contents variations

Juan Wang[1,2†], Haifei Hu[3,4†], Xizhen Liang[1,2], Muhammad Tahir ul Qamar[5], Yunxiang Zhang[1,2], Jianguo Zhao[6], Hongqian Ren[1,2], Xingrong Yan[1,2], Baopeng Ding[1,6*] and Jinping Guo[1,2*]

[1]College of Forestry, Shanxi Agricultural University, Taigu, Shanxi, China, [2]Shanxi Key Laboratory of Functional Oil Tree Cultivation and Research, Shanxi Agricultural University, Taigu, Shanxi, China, [3]Rice Research Institute, Guangdong Key Laboratory of New Technology in Rice Breeding, Guangzhou, China, [4]Guangdong Rice Engineering Laboratory, Guangdong Academy of Agricultural Sciences, Guangzhou, China, [5]Integrative Omics and Molecular Modeling Laboratory, Department of Bioinformatics and Biotechnology, Government College University Faisalabad (GCUF), Faisalabad, Pakistan, [6]Engineering Research Center of Coalbased Ecological Carbon Sequestration Technology of the Ministry of Education, Datong University, Taigu, Shanxi, China

Yellowhorn (*Xanthoceras sorbifolia*) is a species of deciduous tree that is native to Northern and Central China, including Loess Plateau. The yellowhorn tree is a hardy plant, tolerating a wide range of growing conditions, and is often grown for ornamental purposes in parks, gardens, and other landscaped areas. The seeds of yellowhorn are edible and contain rich oil and fatty acid contents, making it an ideal plant for oil production. However, the mechanism of its ability to adapt to extreme environments and the genetic basis of oil synthesis remains to be elucidated. In this study, we reported a high-quality and near gap-less yellowhorn genome assembly, containing the highest genome continuity with a contig N50 of 32.5 Mb. Comparative genomics analysis showed that 1,237 and 231 gene families under expansion and the yellowhorn-specific gene family NB-ARC were enriched in photosynthesis and root cap development, which may contribute to the environmental adaption and abiotic stress resistance of yellowhorn. A 3-ketoacyl-CoA thiolase (*KAT*) gene (*Xso_LG02_00600*) was identified under positive selection, which may be associated with variations of seed oil content among different yellowhorn cultivars. This study provided insights into environmental adaptation and seed oil content variations of yellowhorn to accelerate its genetic improvement.

KEYWORDS

yellowhorn, pan-genomics, genomic profiling, adaptation, oil contents

Introduction

Yellowhorn (*Xanthoceras sorbifolia*), belonging to the *Xanthoceras* genus (*Sapindaceae* family), is a unique woody tree plant species widely growing in Northern and Central China (Bi et al., 2019; Liu et al., 2021a). Yellowhorn shows strong abiotic stress resistance ability and can grow under extreme environmental conditions, including extreme temperature, drought conditions, saline, and alkaline land (Ruan et al., 2017). Furthermore, yellowhorn is easy to reproduce, sowing, root cutting, and grafting and is now considered a promising afforestation species for many arid areas. This oil-rich tree produces capsular fruits from hermaphrodites, with about 60% of its seed kernel containing edible seed oil for the human diet and around 4% nervonic acid essential for nerve and brain development with high medicinal and ornamental value (Liang et al., 2019; Liang et al., 2022). However, yellowhorn also contains moderate erucic acid (about 9% of the total fatty acid) that can damage the heart at high doses (Liu et al., 2021b). Therefore, to make yellowhorn a more desirable species for oil production, it is essential to underly the genetic basis of its oil synthesis pathway and design and cultivate new species with a lower level of erucic acid and a higher level of nervonic acid.

The 3-ketoacyl-CoA thiolase (KAT) is a member of thiolase and can catalyze the final step of fatty acid β-oxidation and the claisen condensation reaction between two Acetyl-CoAs and lead to carbon chain elongation, which is a key step in the fatty acid biosynthetic pathways (Footitt et al., 2007). So far, KAT has been reported to play an important role in producing various energy-storage molecules, such as fatty acids and affecting seed oil content and synthesis in *Arabidopsis thaliana* (Germain et al., 2001) and *Jatropha curcas* (Gomes et al., 2010). However, the mechanism of KAT regulation in yellowhorn and how it underly the fatty acid synthesis remains to be elucidated.

The rapid development of sequencing technologies has facilitated the development of yellowhorn genomes, with two good-quality yellowhorn genome assemblies being published recently (Liang et al., 2019; Liang et al., 2022). These published yellowhorn genomes were sampled and collected from a valley terrain environment with mountains and rivers in Shandong Province. However, yellowhorn also grew and adapted to the loess plateau with a more extreme climate. Therefore, in this study, using long-read sequencing, we sequenced and assembled a gapless *Xanthoceras sorbifolia* genome of the superior line "G11" (Data named *XsoG11*), which was collected from the loess plateau located in Shanxi province. By performing the comparative genomic analysis among representative angiosperms, we revealed that gene families with functions of photosynthesis and root cap development were expanded and existed in yellowhorn, which may associate with adaptation to extreme environmental conditions. With the availability of high-quality yellowhorn reference genomes, we performed the pangenome-wide analysis among three yellowhorn genomes and identified gene content variations that may associate with environment adaptation and oil content variations of different yellowhorn cultivars. All the above results will provide new insights into genetic diversity study of yellowhorn and helps in its genetic improvement.

Materials and methods

Plant materials

Xanthoceras sorbifolia superior cultivar "XsoG11" (*Xanthoceras sorbifolia* superior G11) is a strain with highly comprehensive evaluation selected by *Xanthoceras sorbifolia* research group of Shanxi Agricultural University collected from Lvliang Mountain (Shanxi province; 111°47′17″East, 37°15′57″North) (Supplementary Figure 1), located in semi-arid area, which is extremely cold in winter. The DNA sequencing libraries of PacBio HiFi long reads, Illumina short reads, and Hi-C reads were prepared according to the standard Illumina and PacBio library construction protocol for the generation of genome assembly "XsoG11" (Liang et al., 2021).

Genome assembly

The clean PacBio HiFi reads were assembled using Hifiasm (v.0.15) (Cheng et al., 2021) with default parameters. Then, the original assembly result is polled using pilon (v1.23) (Walker et al., 2014) to get the final genome assembly result. Chromosome-length scaffolds were generated by aligning the raw HiC-reads to the draft assembly using Juicer (v.1.6) (Durand et al., 2016) with the resulting alignments processing by the 3D-DNA pipeline (v.19) (Dudchenko et al., 2017) to generate the candidate chromosome-length assemblies. This candidate assembly was reviewed and curated using Juicebox Assembly Tools (v.1.11.08) (Robinson et al., 2018). BUSCO V3 (Simao et al., 2015) with eukaryota_odb9 was used to assess the completeness of the assembly.

Repeat sequence annotation

For the repeat sequence annotation, trf (v4.09) (Benson, 1999) was used to predict tandem repeats; Microsatellite sequence uses misa Pl program prediction; LTR First use LTR separately_Finder and LTR_Harvest Identify, then use LTR_ Retriever (v2.7) (Ou and Jiang, 2018) integrates the results of the above two software to obtain the final LTR identification results; LINE, SINE, and DNA transposons were identified by RepeatMasker (v4.0.9) (Tarailo-Graovac and Chen, 2009). The two methods are combined to identify the repeat contents in our genome, homology-based and *de novo* prediction. Homology-based analysis: We identified the known TEs within the XsoG11 genome using RepeatMasker (v4.0.9) (Tarailo-Graovac and Chen, 2009) with the Repbase TE library. *De novo* prediction: We constructed a *de novo* repeat library of the XsoG11 genome using RepeatModeler, which can automatically execute two core *de novo* repeat-finding programs, namely, RECON (v1.08) (Bao and Eddy, 2002) and RepeatScout (v1.0.5) (Benson, 1999), to comprehensively conduct, refine and classify consensus models of putative interspersed repeats for the XsoG11 genome. Furthermore, we performed a *de novo* search for long terminal repeat (LTR) retrotransposons against the XsoG11 genome sequences using LTR_Finder (v1.0.7) (Xu and Wang, 2007), LTR_harvest (v1.5.11) and LTR_retriever (v2.7) (Ou and Jiang, 2018). We also identified

tandem repeats using the Tandem Repeat Finder (TRF) package and the SimpleSequence Repeats (SSR) using misa (v1.0) (Beier et al., 2017). Finally, we merged the library files of the two methods to identify and determine the repeat contents.

Gene annotation

We predicted protein-coding genes of the XsoG11 genome using three methods: ab initio gene prediction, homology-based gene prediction, and RNA-Seq-guided gene prediction. Before gene prediction, the assembled XsoG11 was hard and soft masked using RepeatMasker (v4.0.9) (Tarailo-Graovac and Chen, 2009). We adopted Augustus (v3.3.3) (Stanke et al., 2008) to perform *ab initio* gene prediction. Models used for each gene predictor were trained from a set of high-quality proteins generated from the RNA-Seq dataset. We used maker (v2.31.10) (Holt and Yandell, 2011) to conduct homology-based gene prediction. First, the protein sequences and transcripts sequences were aligned to our genome assembly and predicted coding gene using maker with the default parameters. To carry out RNA-Seq-guided gene prediction, we first aligned clean RNA-Seq reads to the genome using hisat2 (v2.0.0) (Kim et al., 2015), and the gene structure was formed using Trinity (v2.3.2) (Grabherr et al., 2011), Transdecoder (v2.01) (Haas et al., 2017) and maker (v2.31.10) (Holt and Yandell, 2011). Finally, EVidenceModeler (v1.1.1) (Haas et al., 2008) was used to integrate the prediction results of the three methods to predict gene models. Functional annotation was performed by comparing proteins with various functional databases including NR, swiss pro, KOG and TrEMBL, using BLASTP (e-value < 1e-5) (Camacho et al., 2009).

Comparative genomics analysis

Using the assembled yellowhorn genome (XsoG11) and nine other related angiosperm genomes, we performed a comparative genome analysis using OrthoFinder (v 2.4.0) (Emms and Kelly, 2019) to identify the orthologous gene families in the yellowhorn genome. The analysis process of the OrthoFinder was indicated as follows: 1) Use the diamond to input all protein sequences for all-*vs*-all comparison and detect homologous gene pairs (Evaluate < 1e-5 and the minimum coverage is > 40%). 2) Input the list of homologous gene pairs into MCL program for family clustering. A maximum likelihood phylogenetic tree of ten species was constructed based on shared single-copy genes using Mega V5 (Tamura et al., 2011). Expanded and contracted gene families were detected using CAFÉ (v4.2.1) (De Bie et al., 2006). The expanded gene families were functionally annotated on Pfam v32.0 (Mistry et al., 2021) and Swiss-Prot (UniProt, 2021) databases. The functional enrichment of each gene family was determined using a Fisher's exact test (false discovery rate < 0.05).

Pan-genomics analysis

The genome sequences and protein sequences of two published yellowhorn (WF18 and Xsv2) were downloaded from Liang and Liu study (Liang et al., 2019, Liu et al., 2021a). Orthologous genes among the yellowhorn genomes were identified by Orthovenn2 (Xu et al., 2019), a web tool used to identify orthologous and paralogous genes, with a pairwise sequence similarity cut-off of 10- 5 and inflation of 1.5 to define orthologous cluster structure. KaKs_Calculator 2.0 (Wang et al., 2010) was used to calculate orthologous gene clusters' non-synonymous/synonymous substitution ratio. Orthologous clusters and gene pairs under positive selection (Ka/ks > 1) were analyzed by UniProt search and TopGO (Alexa et al., 2006) using Fisher's exact tests for functional annotation and enrichment analysis. Furthermore, the two published genomes were compared to our genome assembly using the mumandco_ V3 program (parameter default) (O'donnell and Fischer, 2020).

Results

Genome assembly and annotation

In this study, we used the long read sequencing to *de novo* assemble a near gapless genome assembly of the Shanxi yellowhorn cultivar XsoG11 (Figure 1A; Table 1). We generated approximately 30-fold coverage of PacBio CCS (HiFi) reads and assembled the CCS (HiFi) reads using Hifiasm: a haplotype-resolved assembler for accurate HiFi reads (Cheng et al., 2021). The assembly length of the XsoG11 genome is 489.18 Mb with a contig N50 of 32.5 Mb, showing the highest contiguity than the previously published yellowhorn genomes (Table 2). The contigs were further polished using pilon, then ordered, oriented and anchored to chromosomes using *in-situ* Hi-C sequencing. We found that around 96.2% (470.79 Mb) of sequences are anchored to the chromosome and seven chromosomes in our genome do not contain any gaps (Table 1; Figure 1B). XsoG11 assembly has 95.7% complete BUSCOs (Table 2; Figure 1C), comparable to the previously published Xsv2 (Liu et al., 2021b) and WF18 yellowhorn genomes (Liang et al., 2019). In addition, we identified approximately 68.71% repeat sequences in the assembled genome, in which the long terminal repeat (LTR) retrotransposon element represents the most abundant transposable elements (TEs) class, accounting for 35.8% TEs (Table 3). Using RNA-seq transcript mapping combined with ab initio prediction and homologous protein searches, we predicted 35,039 protein-coding genes with an average gene length of 2,662 bp in the XsoG11 genome (Supplementary Table 1), in which 27,082 (85%) genes have functional annotation from at least one functional protein database, including nr (84%), TrEMBL (84%), KOG (41%), Swiss-Prot (57%), Pfam (69%), Gene Ontology (GO) (42%), and KEGG (23%) (Supplementary Table 2). We also identified 1,250 tRNAs, 770 small nucleolar RNAs and 4,691 small nuclear RNAs (Supplementary Table 3).

Yellowhorn phylogenetics and gene family expansion analyses

The change in gene family size plays an important role in the evolution of angiosperms' environmental adaptation and trait

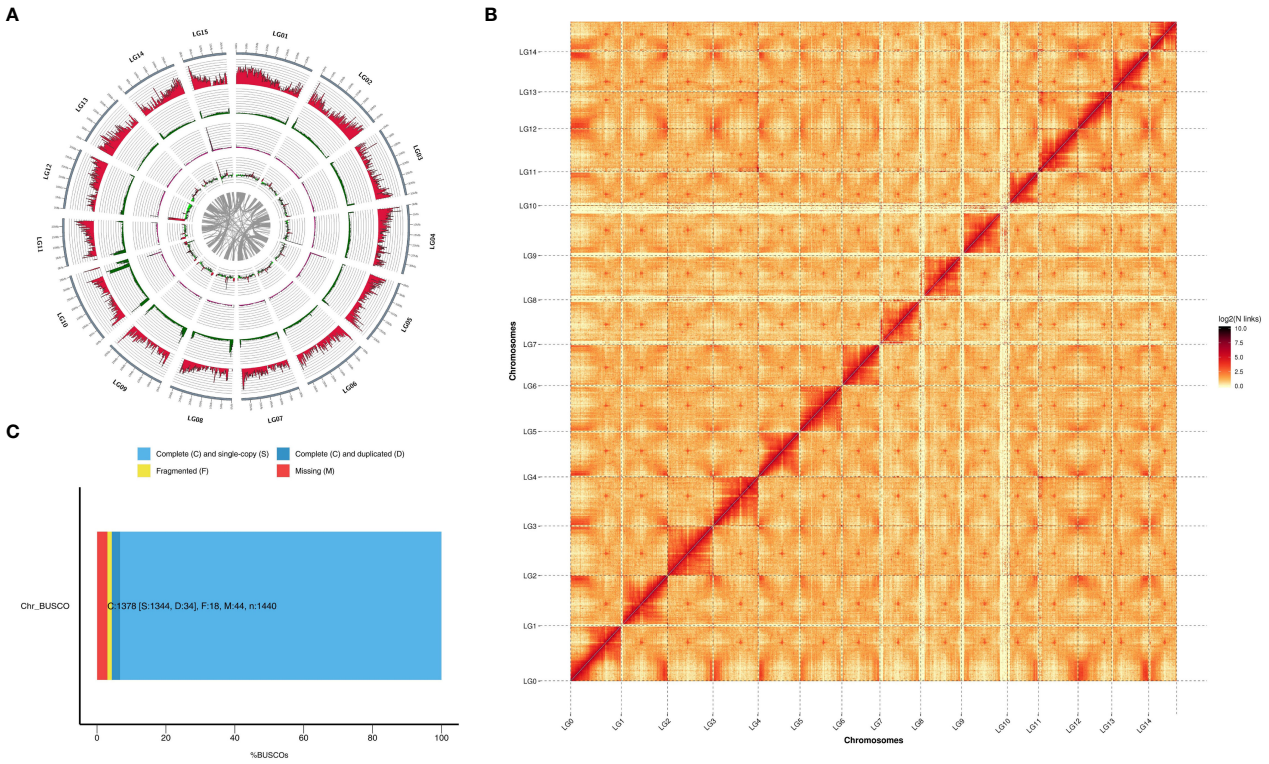

FIGURE 1
The characteristics of yellowhorn XsoG11 genome assembly. **(A)** The yellowhorn genome feature. From outer-most track to innermost track: gene density, transposable element density, repeat sequence density, GC content, and Intra-genome collinear blocks. **(B)** Buscos of complete genome. **(C)** Contact map of Hi-C links among 14 pseudochromosomes.

TABLE 1 Genome Statistic of *Xanthoceras sorbifolia* superior G11.

Chr	Chr_size(bp)	Contig_number	Contig_size(bp)	GC_content(%)
LG01	39494243	1	39494243	34.9
LG02	35873774	1	35873774	35.5
LG03	35257883	4	35257583	35.05
LG04	34944601	2	34944501	35.14
LG05	32593928	4	32593628	34.99
LG06	32495627	1	32495627	35.18
LG07	29437501	1	29437501	35.28
LG08	31632995	1	31632995	35.22
LG09	31780288	3	31780088	35.34
LG10	35671781	2	35671681	35.4
LG11	24152185	3	24151985	35.44
LG12	30754806	7	30754206	36.83
LG13	26198562	1	26198562	35.16
LG14	28601909	1	28601909	35.09
LG15	21899742	2	21899642	35.21
ChrAll	470789825	34	470787925	35.31

TABLE 2 Statistic of different yellowhorn genome assemblies.

TypeParameter	WF18v1	Xsv2	ZS4	WF18	XsoG11
Assembly Genome size(Mb)	490.44	470	504.2	440	489.18
Chromosome-scale scaffolds(Mp)	490.24 (99.96%)	446.2 (94.9%)	489.29 (97.04%)	420 (95.4%)	470.79 (96.24%)
Total num. of scaffolds	22	988	297	267	417
Total num. of chromosomes	15	15	15	15	15
ScaffoldN50(Mb)	34	30.8	32.17	29.4	31.6
Total num. of Contigs	2,428	3,302	3,035	2,002	417
Contig N50(Mb)	0.42	0.42	1.04	0.64	31.6
Complete BUSCOs	98.70%	97.50%	98.70%	84.60%	95.70%
GC content of the genome(%)	34.71	34.94	36.95	32.75	35.70%
Protein-coding genes	29,888	22,049	24,672	21,059	35,039
Reference	(Liang et al., 2022)	(Liu et al., 2021a)	(Bi et al., 2019)	(Liang et al., 2019)	This study

formation during evolution (Van de Peer et al., 2009). To further dissect the genetic basis of high seed oil content and the ability to adapt to extreme environments of yellowhorn, we performed the comparative genomics and gene family expansion analysis in yellowhorn and eight other representative angiosperms and the outgroup species (*Taxus chinensis*). We first determined the phylogeny position of yellowhorn by constructing a phylogenetic tree using 17 single-copy orthologous genes conserved in 10 representative angiosperms (Figure 2A; Supplementary Table 4). Our result inferred that *Acer yangbiense* was the most recent common ancestor of yellowhorn, which diverged around 76.5 million years ago (Figure 2B).

We further performed the collinear analysis by comparing the yellowhorn (XsoG11) genome with *Acer yangbiense* and *Sapindus mukorossi* genomes (Figure 2C). Although these species belong to the *Sapindaceae* family, we identified significant structural variations in yellowhorn and *Acer yangbiense* and *Sapindus mukorossi*, suggesting that significant chromosomal differentiation occurred since they derived from the last common ancestor. A total of 4,622 gene families were shared by ten studied species

TABLE 3 Repeat elements of the XsoG11 yellowhorn genome.

repeat_type	total_size	Percentage of genome (%)
DNA/CMC-EnSpm	1456919	0.30%
DNA/hAT-Ac	5043893	1.03%
DNA/hAT-Tag1	1039259	0.21%
DNA/hAT-Tip100	1516665	0.31%
DNA/MuLE-MuDR	4508774	0.92%
DNA/PIF-Harbinger	887491	0.18%
DNA/TcMar-Pogo	209047	0.04%
LINE/L1	19514034	3.99%
LINE/RTE-X	67436	0.01%
Low_complexity	1285729	0.26%
LTR	1516161	0.31%
LTR/Caulimovirus	1471510	0.30%
LTR/Copia	46538428	9.51%
LTR/Gypsy	62413913	12.76%
Simple_repeat	8520634	1.74%
Unknown	180122927	36.82%
Total	336112820	68.71%

10.3389/fpls.2023.1147946

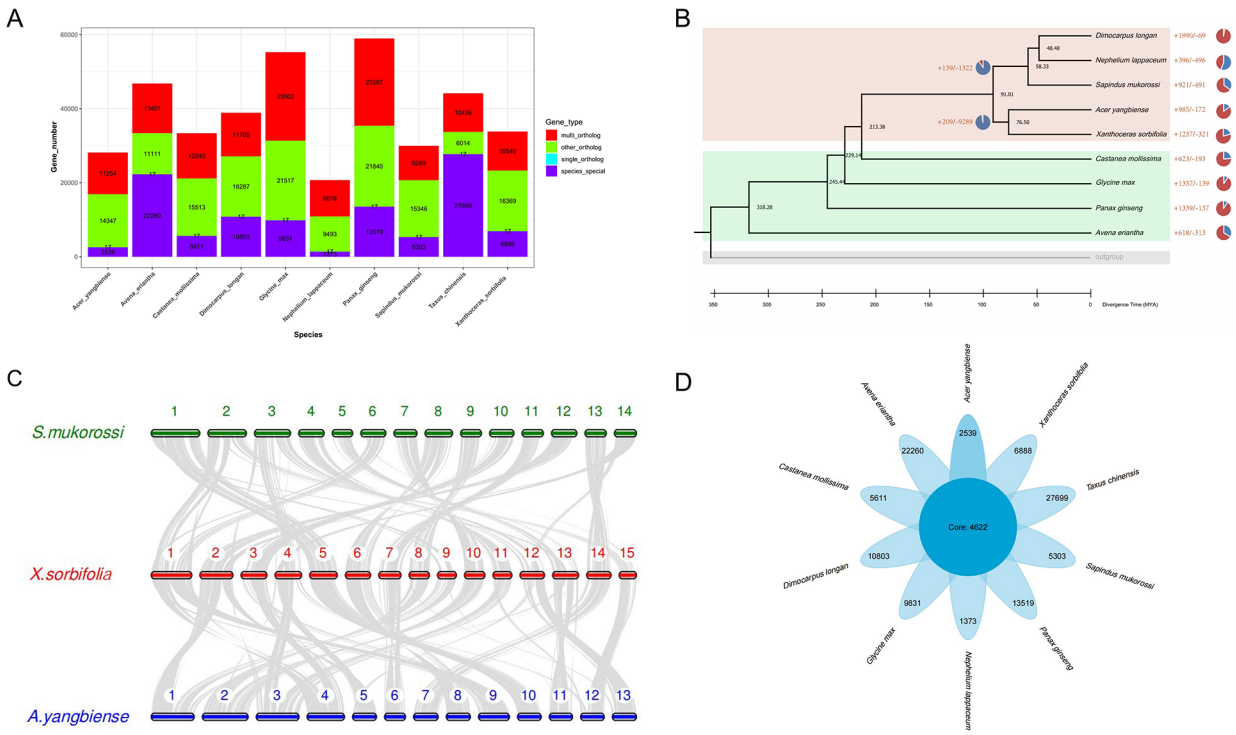

FIGURE 2
Comparative analysis yellowhorn and nine other representative angiosperms. **(A)** Gene family clusters. Multi_ortholog: This family exists in all species, and the number of members in at least one species is greater than or equal to two; single_ortholog: a single copy orthologous gene family, which exists in all species and has 1 member in all species; other_ortholog: The family exists in at least two species and does not exist in at least one species; species_special: the unique gene family of each species. **(B)** Maximum-likelihood phylogeny of yellowhorn and nine other representative angiosperms. **(C)** Collinear analysis by comparing the yellowhorn (*X.sorbifolia*, XsoG11) genome with *Acer yangbiense* (*A. yangbiense*) and *Sapindus mukorossi* genomes (*S. mukorossi*). **(D)** The comparative genomic analysis shows the number of the core and species specific gene families.

(Figure 2D), with yellowhorn having 6,888 species-specific gene families. GO enrichment analysis indicates that these species-specific gene families are significantly enriched in functions associated with photosynthesis (Supplementary Table 5). Additionally, the gene family size analysis showed that 1,237 and 231 gene families were found to be expanding and contracting (Supplementary Table 6). Enrichment analysis showed that the expanding gene families were enriched for functions associated with disease resistance (Pfam: NB-ARC domain) (Supplementary Table 7) and photosynthesis (GO: "photosynthetic electron transport in photosynthesis", "photosynthetic electron transport chain", "photosynthesis, light reaction" and "photosynthesis"; KEGG: ko00195: Photosynthesis) (Supplementary Tables 8, 9). These results may suggest that gene family expansion was associated with photosynthesis and biotic stress resistance in yellowhorn.

Pan-genome analysis of yellowhorn genomes

High-quality genome assemblies enable the accurate discovery of structural variations and genetic variations among genomes. Using the high-quality genome assembly (XsoG11) as the reference genome, we further discovered abundant structural variations (SVs), including inversions, translocations, insertions and deletions between the XsoG11 genome and the other two published yellowhorn genomes (Xsv2 and WF18) (Figure 3A; Supplementary Tables 10, 11). We identified 3,515 and 2,262 sequences uniquely present in XsoG11 by comparing this assembly with Xsv2 and WF18 genomes, respectively, in which 1,005 sequences are present in the XsoG11 genome but missing in both Xsv2 and WF18 genomes (Figure 3B). In addition, we further performed the pan-genome wide analysis of gene families (Zia et al., 2022). The pangenome-wide gene family clustering analysis revealed that these three genomes shared 11,750 core orthologous clusters, whereas at least one genome (but less than three) shared 8,844 dispensable orthologous clusters, with 677 XsoG11-specific dispensable orthologous clusters (Figure 3C). The evolutionary analysis of the three yellowhorn showed that they are clustered together (Supplementary Figure 2) and the results of collinearity is consistent (Supplementary Figure 3). Functional annotation of genes located in these 1,005 sequences found oxidative phosphorylation (ko00190, $P < 2.19E-83$), ribosome (ko03010, $P < 1.06E-61$), RNA polymerase genes (ko03020, $P < 1.44E-13$) were the first three most significantly enriched in KEGG pathyway. In addition, photosynthetic pathway genes were also significantly enriched (ko00195, $P < 0.003$) (Figure 4). Gene ontology (GO) enrichment analysis shows that genes with essential biological functions, including RNA-DNA hybrid ribonuclease activity,

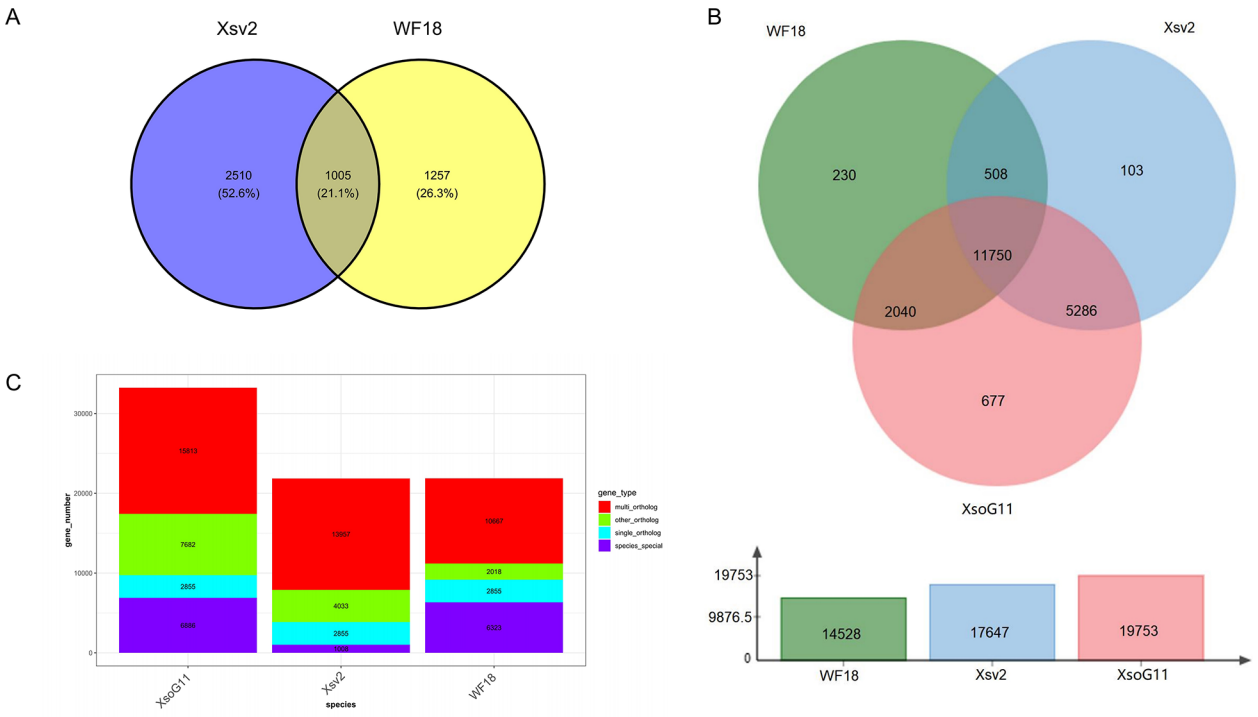

FIGURE 3
Pan-genome analysis of yellowhorn genomes. **(A)** The Venn diagram shows the number of SVs that are uniquely and commonly absent in the Xsv2 and WF18 genome, by comparing with the XsoG11 genome. **(B)** Pangenome-wide analysis of the core and dispensable gene families among three different yellowhorn genomes. **(C)** Gene family clustering,.

DNA recombination, oxidation-reduction process and DNA integration, were enriched in core orthologous clusters (Supplementary Table 12). By contrast, genes with functions potentially associated with fatty acid synthesis and abiotic stress responses, such as photosynthesis and root cap development, are enriched in XsoG11-specific dispensable orthologous clusters (Supplementary Table 13). We further examined the non-synonymous/synonymous substitution ratio (Ka/Ks) of homologous gene pairs of XsoG11 and two other published yellowhorn genomes (Xsv2 and WF18). The result showed that a total of 364 genes in the XsoG11 are under positive selection (Ka/Ks > 1) (Supplementary Table 14), including a gene (*Xso_LG02_00600*) encoding 3-ketoacyl-CoA thiolase associated with the formation of fatty acid (ko01040: Biosynthesis of unsaturated fatty acids; ko00592, alpha-Linolenic acid metabolism).

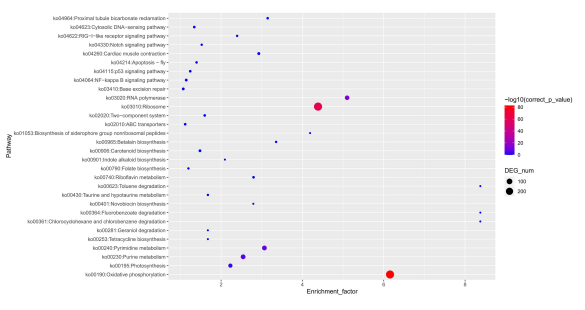

FIGURE 4
XsoG11 special gene enrichment in KEGG.

Discussion

A high-quality yellowhorn genome assembly is key to underlying the genetic basis of its ability to adapt to extreme environments and produce high oil seed content. In this study, using Circular consensus sequencing (CCS) long-read sequencing, we present a high-quality chromosome-scale genome assembly of Shanxi's yellowhorn cultivar (XsoG11). Compared with the previously published yellowhorn Shandong's cultivar "Shanyou 1" genome (WF18) (Liang et al., 2021), we assembled a higher contiguity (Contig N50: XsoG11: 31.6 Mb *vs* WF18v1: 0.42 Mb) and a near-gapless yellowhorn genome, with seven out of 15 chromosomes having no gaps. This additional high-quality genome can provide novel genomic resources for future yellowhorn improvement.

Photosynthesis is an important physiological process that converts light energy into chemical energy, affecting plant growth and development, respiration and transpiration (Goudriaan et al., 1985). Recent studies suggest that genes involved in the photosynthesis pathway are essential for environmental adaptation to different light regimes and coping with climate change by regulating the circadian clock and light perception (Kreps and Kay, 1997; Quint et al., 2016). For example, a pan-genome study of mung beans indicates that the presence/absence

variation (PAV) of genes regulating the photosynthesis pathway enables mung beans to adapt to different environments (Liu et al., 2022). Gene expansion through tandem duplication is important for stress response (Hanada et al., 2008). Photosynthesis-related genes, early light-induced proteins (ELIPs), were found to be expanded in plants showing drought resistance (Liu et al., 2020). Combined with the public genome assemblies of representative angiosperms and yellowhorn, this high-quality assembly enabled us to identify gene family differences and the expansion/contraction of yellowhorn during speciation and divergence and among yellowhorn cultivars growing in a different climate region. Compared with other angiosperms and yellowhorn cultivars collected in Shandong Province, the yellowhorn assembled in this study grows in the Loess Plateau of the middle of Shanxi Province with poor and dry soil (Yao et al., 2013). Both GO enrichment analyses of yellowhorn's expanding gene families and pangenome-wide XsoG11-specific dispensable gene families showed that gene families are significantly enriched in functions associated with photosynthesis. The enrichment of photosynthesis can lead to the accumulation of plant carbohydrates and other carbon sources, resulting in the generation of more energy to cope with the adverse stress of environmental factors in the Loess Plateau (Farrar and Williams, 1991). In addition, we also found that genes associated with root cap development were also enriched in the XsoG11-specific dispensable gene families that were missing in other yellowhorn cultivars. This may reflect that the yellowhorn cultivar growing in the drought Loess Plateau requires a more robust and extended root cap system (Du et al., 2011).

We also revealed that a gene (*Xso_LG02_00600*) encoding 3-ketoacyl-CoA thiolase (KAT) was under positive selection by comparing the value of Ka/Ks of homologous gene pairs of our yellowhorn cultivar with the Shandong's yellowhorn cultivars. KAT is an important catalyst for the process of fatty acid beta-oxidation. In *Arabidopsis thaliana*, the *KAT* gene was demonstrated to be activated in the early germination and seedling stage and led to fatty acyl-CoAs accumulation and the form of triacylglycerol, facilitating lipid storage in the oil seed (Germain et al., 2001; Footitt et al., 2007). A similar finding was reported in *Ophiocordyceps sinensis* that KAT was regulated and participated in the fatty acid pathway and provided sufficient energy for organisms by catalyzing the tricarboxylic acid cycle and electronic respiratory chain (Wang et al., 2022). The positive selection of the *KAT* gene in our Shanxi's yellowhorn cultivar may associate with its seed oil content variations among the comparative yellowhorn cultivars.

Conclusion

In a nutshell, we assembled a high-quality and near gap-less yellowhorn genome collected from Loess Plateau, providing valuable genomic resources for future yellowhorn genetic improvement. The functional analysis shows that the gene family under expansion and the yellowhorn-specific gene family are enriched in functions associated with photosynthesis and root cap development, which may relate to the environmental adaption of yellowhorn. A *KAT* gene under positive selection was identified, reflecting variations of seed oil content among different yellowhorn cultivars. This study provide a foundation for further genetic improvement of yellowhorn.

Data availability statement

All sequencing data and genome assembly generated in this study are available on SRA at the NCBI with the accession numbers PRJNA924504.

Author contributions

JG and BD designed the experiment. JW and HH performed research and did the methodology, finish writing and editing. XL, YZ and JZ helped with the part of data and results. MTQ, HR and XY critically revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by Shanxi Agricultural University Doctoral Research Initiation Project (2021BQ17); Basic Research Program of Shanxi Province (202103021224168); The Shanxi Postdoctoral Research Fund Project (K462102907), Shanxi Province doctoral graduates and postdoctoral researchers come to work in Shanxi Province to reward the fund scientific research project (SXBYKY2021061).

Acknowledgments

We would like to acknowledge The Engineering Research Center of Coal-based Ecological Carbon Sequestration Technology of the Ministry of Education for suggestion.We would also like to acknowledge Beijing Biomarker Technology Co. Ltd. and Wuhan Carboncode Biotechnologies Co.,Ltd. for performing DNA sequencing and genome assembly.

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.2023.1147946/full#supplementary-material

SUPPLEMENTARY FIGURE 1
XsoG11 yellowhorn tree used in this study.

SUPPLEMENTARY FIGURE 2
Evolution tree of three yellowhorn ("WF18", "XsoG11" and "Xsv2").

SUPPLEMENTARY FIGURE 3
The collinearity of three yellowhorn ("WF18","XsoG11" and "Xsv2").

References

Alexa, A., Rahnenfuhrer, J., and Lengauer, T. (2016). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22, 1600–1607. doi: 10.1093/bioinformatics/btl140

Bao, Z., and Eddy, S. R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* 12, 1269–1276. doi: 10.1101/gr.88502

Beier, S., Thiel, T., Munch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

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

Bi, Q., Zhao, Y., Du, W., Lu, Y., Gui, L., Zheng, Z., et al. (2019). Pseudomolecule-level assembly of the Chinese oil tree yellowhorn (*Xanthoceras sorbifolium*) genome. *GigaScience* 8 (6), giz070. doi: 10.1093/gigascience/giz070

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10, 1–9. doi: 10.1186/1471-2105-10-421

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* 18, 170–175. doi: 10.1038/s41592-020-01056-5

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Du, S., Wang, Y. L., Kume, T., Zhang, J. G., Otsuki, K., Yamanaka, N., et al. (2011). Sapflow characteristics and climatic responses in three forest species in the semiarid loess plateau region of China. *Agric. For. Meteorol.* 151, 1–10. doi: 10.1016/j.agrformet.2010.08.011

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the *Aedes aegypti* genome using Hi-c yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-c experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Farrar, J., and Williams, M. (1991). The effects of increased atmospheric carbon dioxide and temperature on carbon partitioning, source-sink relations and respiration. *Plant Cell Environ.* 14, 819–830. doi: 10.1111/j.1365-3040.1991.tb01445.x

Footitt, S., Cornah, J. E., Pracharoenwattana, I., Bryce, J. H., and Smith, S. M. (2007). The *Arabidopsis 3-ketoacyl-CoA thiolase-2* (*kat2-1*) mutant exhibits increased flowering but reduced reproductive success. *J. Exp. Bot.* 58, 2959–2968. doi: 10.1093/jxb/erm146

Germain, V., Rylott, E. L., Larson, T. R., Sherson, S. M., Bechtold, N., Carde, J. P., et al. (2001). Requirement for 3-ketoacyl-CoA thiolase-2 in peroxisome development, fatty acid β-oxidation and breakdown of triacylglycerol in lipid bodies of *Arabidopsis* seedlings. *Plant J.* 28, 1–12. doi: 10.1046/j.1365-313X.2001.01095.x

Gomes, K. A., Almeida, T. C., Gesteira, A. S., Lôbo, I. P., Guimarães, A. C. R., De Miranda, A. B., et al. (2010). ESTs from seeds to assist the selective breeding of jatropha curcas l. for oil and active compounds. *Genomics Insights* 3, GEI. S4340. doi: 10.4137/GEI.S4340

Goudriaan, J., Van Laar, H., and H.v. and Louwerse, W. (1985). Photosynthesis, CO_2 and plant production. *Wheat Growth Model.* 86, 107–122. doi: 10.1007/978-1-4899-3665-3_10

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. *Nat. Biotechnol.* 29, 644. doi: 10.1038/nbt.1883

Haas, B., Papanicolaou, A., and Yassour, M. (2017). TransDecoder. Available at: https://github.com/TransDecoder/TransDecoder.

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

Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K., and Shiu, S. H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* 148, 993–1003. doi: 10.1104/pp.108.122457

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinf.* 12, 491. doi: 10.1186/1471-2105-12-491

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

Kreps, J. A., and Kay, S. A. (1997). Coordination of plant metabolism and development by the circadian clock. *Plant Cell* 9, 1235-1244. doi: 10.1105/tpc.9.7.1235

Liang, Q., Fang, H., Liu, J., Zhang, B., Bao, Y., Hou, W., et al. (2021). Analysis of the nutritional components in the kernels of yellowhorn (*Xanthoceras sorbifolium* bunge) accessions. *J. Food Compos. Anal.* 100, 103925. doi: 10.1016/j.jfca.2021.103925

Liang, Q., Li, H., Li, S., Yuan, F., Sun, J., Duan, Q., et al. (2019). The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium* bunge). *GigaScience* 8, giz071. doi: 10.1093/gigascience/giz071

Liang, Q., Liu, J., Fang, H., Dong, Y., Wang, C., Bao, Y., et al. (2022). Genomic and transcriptomic analyses provide insights into valuable fatty acid biosynthesis and environmental adaptation of yellowhorn. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.991197

Liu, C., Wang, Y., Peng, J., Fan, B., Xu, D., Wu, J., et al. (2022). High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Commun.* 3, 100352. doi: 10.1016/j.xplc.2022.100352

Liu, F., Wang, P., Xiong, X., Zeng, X., Zhang, X., and Wu, G. (2021b). A review of nervonic acid production in plants: Prospects for the genetic engineering of high nervonic acid cultivars plants. *Front. Plant Sci.* 12, 626625. doi: 10.3389/fpls.2021.626625

Liu, H., Yan, X., Wang, X., Zhang, D., Zhou, Q., Shi, T., et al. (2021a). Centromere-specific retrotransposons and very-long-chain fatty acid biosynthesis in the genome of yellowhorn (*Xanthoceras sorbifolium*, sapindaceae), an oil-producing tree with significant drought resistance. *Front. Plant Sci.* 12, 766389. doi: 10.3389/fpls.2021.766389

Liu, X., Zhang, Y., Yang, H., Liang, Y., Li, X., Oliver, M. J., et al. (2020). Functional aspects of early light-induced protein (ELIP) genes from the desiccation-tolerant moss *Syntrichia caninervis*. *Int. J. Mol. Sci.* 21, 1411-1429. doi: 10.3390/ijms21041411

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

O'donnell, S., and Fischer, G. (2020). MUM&Co: accurate detection of all SV types through whole-genome alignment. *Bioinformatics* 36, 3242–3243. doi: 10.1093/bioinformatics/btaa115

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

Quint, M., Delker, C., Franklin, K. A., and Wigge, P. A. (2016). Molecular and genetic control of plant thermomorphogenesis. *Nat. Plants* 2, 15190. doi: 10.1038/nplants.2015.190

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdottir, H., Mesirov, J. P., and Aiden, E. L. (2018). Juicebox.js provides a cloud-based visualization system for Hi-c data. *Cell Syst.* 6, 256–258 e251. doi: 10.1016/j.cels.2018.01.001

Ruan, C. J., Yan, R., Wang, B. X., Mopper, S., Guan, W. K., and Zhang, J. (2017). The importance of yellow horn (*Xanthoceras sorbifolia*) for restoration of arid habitats and production of bioactive seed oils. *Ecol. Eng.* 99, 504–512. doi: 10.1016/j.ecoleng.2016.11.073

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24, 637–644. doi: 10.1093/bioinformatics/btn013

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood,

evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf. Chapter 4 Unit* 4, 10. doi: 10.1002/0471250953.bi0410s25

UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489. doi: 10.1093/nar/gkaa1100

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* 10, 725–732. doi: 10.1038/nrg2600

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* 9, e112963. doi: 10.1371/journal.pone.0112963

Wang, Y., Li, C., Zhang, J., Xu, X., Fu, L., Xu, J., et al. (2022). Polyunsaturated fatty acids promote the rapid fusion of lipid droplets in caenorhabditis elegans. *J. Biol. Chem.* 298, 102179. doi: 10.1016/j.jbc.2022.102179

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genom. Proteom Bioinf.* 8, 77–80. doi: 10.1016/S1672-0229(10)60008-3

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., et al. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* 47, W52–W58. doi: 10.1093/nar/gkz333

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

Yao, Z.-Y., Qi, J.-H., and Yin, L.-M. (2013). Biodiesel production from *Xanthoceras sorbifolia* in China: Opportunities and challenges. *Renew. Sust. Energ. Rev.* 24, 57–65. doi: 10.1016/j.rser.2013.03.047

Zia, K., Rao, M., Sadaqat, M., Azeem, F., Fatima, K., Tahir Ul Qamar, M., et al. (2022). Pangenome-wide analysis of cyclic nucleotide-gated channel (CNGC) gene family in citrus spp. revealed their intraspecies diversity and potential roles in abiotic stress tolerance. *Front. Genet.* 13, 1034921. doi: 10.3389/fgene.2022.1034921