scientific data

OPEN

DATA DESCRIPTOR

Two haplotype-resolved telomere-to-telomere genome assemblies of *Xanthoceras sorbifolium*

Yu Liu[1,2,6], Yijun Chen[2,3,6], Zizheng Ren[2,6], Kui Li[2,6], Xu Wang[2], Kai Wu[2], Jinfeng Liu[4], Nir Sade[3], Hang He[2,5] ✉, Shouke Li[4] ✉, Haiyang Jiang[1] ✉ & Xue Han[2] ✉

Yellowhorn (*Xanthoceras sorbifolium*) is widely used in northern China for landscaping, desertification control, and oil production. However, the lack of high-quality genomes has hindered breeding and evolutionary studies. Here, we present the first haplotype-resolved, telomere-to-telomere (T2T) yellowhorn genomes of PBN-43 (white single-flowered) and PBN-126 (white double-flowered) using PacBio HiFi and Hi-C data. These assemblies range from 464.34 Mb to 468.97 Mb and include all centromeres and telomeres. Genome annotation revealed that an average of 67.99% (317.09 Mb) of yellowhorn genomic regions consist of repetitive elements across all haplotypes. The number of protein-coding genes ranges from 35,039 to 35,174 among assemblies, representing an average 50.16% increase over the first published yellowhorn genome. Additionally, 93.90% of the annotated genes have functional annotations. We found yellowhorn experienced an LTR-RT burst during the last 0.45–0.48 Mya. These data provide a resource for investigating genomic variations, phylogenetic relationships, duplication modes, and the distribution of nucleotide-binding leucine-rich repeat (NLR) genes, and support further research into yellowhorn breeding.

Background & Summary

Xanthoceras sorbifolium, commonly known as yellowhorn, is a monotypic species of the Sapindaceae family and a deciduous shrub or small tree native to northern China[1]. Its spring blossoms display notable petal color changes and diverse shapes, making it a popular ornamental in landscaping[2]. With robust tolerance to drought, cold, and saline-alkali soils, yellowhorn thrives in arid and semi-arid regions, aiding desertification control in northwestern China[3]. Yellowhorn seed kernels contain 55%–65% oil, with 93% unsaturated fatty acids, supporting both edible oil and biodiesel production[4]. This combination of ornamental appeal, resistance, and high oil content makes yellowhorn a valuable species in horticultural and agricultural contexts.

A notable ornamental variant is the double-flowered form, characterized by multiple petal whorls but rendered sterile due to a transposon insertion in the intron region of the *AGAMOUS* (*AG*) gene *XsAG1*[5]. However, the complete structure of *XsAG1* remains unclear, and a comprehensive genome for the double-flowered variety is still lacking. Beyond ornamentation, yellowhorn displays significant resistance to plant pathogens and abiotic stress[6,7]. NLR genes are essential for pathogen detection and immune signaling[8–10], underscoring the need for high-quality genomes to identify NLR gene loci and further development of disease-resistant varieties[11].

Although several yellowhorn genomes have been assembled since 2019[7,12–16], limitations including incomplete telomeres and centromeres, unclosed gaps, and limited annotation quality constrain our genetic understanding. A haplotype-resolved T2T genome can serve as a blueprint for precise breeding. In this study, we constructed the first haplotype-resolved T2T gapless genomes for two ornamental yellowhorn varieties: "PBN-43", a single-flowered, white-petaled cultivar from Zhangjiakou (Hebei, China), and "PBN-126", a double-flowered, white-petaled cultivar from Chifeng (Nei Mongol, China) (Supplementary Fig. S1).

[1]National Engineering Laboratory of Crop Stress Resistance Breeding, School of Life Sciences, Anhui Agricultural University, Hefei, 230036, China. [2]Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences in Weifang, Weifang, 261325, China. [3]School of Plant Sciences and Food Security, Tel Aviv University, Tel Aviv, 69978, Israel. [4]Shandong Woqi Agriculture Development Co., Ltd, Weifang, 262100, China. [5]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, 100871, China. [6]These authors contributed equally: Yu Liu, Yijun Chen, Zizheng Ren, Kui Li. ✉e-mail: hang.he@pku-iaas.edu.cn; lishouke@163.com; hyjiang@ahau.edu.cn; xue.han@pku-iaas.edu.cn

Varieties	Library Type	Total bases	Number of sequences	N50 of read length (bp)	Depth
PBN-43	PacBio HiFi	98,321,815,446	5,680,974	17,616	210.27
	Hi-C	102,979,984,500	686,533,230	150	220.23
	Illumina NGS	28,193,334,300	187,955,562	150	60.29
	RNA-seq (NGS)	125,772,655,500	838,484,370	150	—
PBN-126	PacBio HiFi	94,074,430,759	5,632,693	16,802	200.26
	Hi-C	100,863,055,950	672,420,373	150	214.72
	Illumina NGS	28,662,192,450	191,081,283	150	61.02
	RNA-seq (NGS)	140,798,174,400	938,654,496	150	—

Table 1. Sequencing data statistics for PBN-43 and PBN-126.

We employed PacBio HiFi reads, high-throughput chromatin conformation capture (Hi-C) reads, Illumina paired-end (NGS) reads, and RNA-seq reads for genome assembly and annotation. The final haplotype genomes range in length from 464.34 Mb to 468.93 Mb and contain an average of 35,084 high-confidence protein-coding genes (PCGs). Additionally, we systematically revealed profiles of genetic variations, transposable elements (TEs), and NLR genes based on the assembled genomes, collectively providing valuable genomic resources for future studies and breeding.

Methods

Plant materials and sequencing. The single-flowered variety (PBN-43) and double-flowered variety (PBN-126) of *Xanthoceras sorbifolium* were cultivated at the Woqi Agriculture Development Co., Ltd yellowhorn orchard, Weifang, Shandong Province, China (Supplementary Fig. S1). For each variety, all tissue samples were collected from the same individual plant. Specifically, 10 g of young fresh leaves were collected and flash-frozen in liquid nitrogen for DNA and RNA extraction and sequencing. For flower tissues and mature seeds (77 days after flowering, only for PBN-43), 2 g per sample were collected for RNA extraction and sequencing. Genomic DNA was isolated using cetyltrimethylammonium bromide (CTAB) method. Frozen young leaves were ground into powders in liquid nitrogen for DNA extraction, and the quality of the isolated genomic DNA was assessed by monitoring DNA degradation and contamination on 1% agarose gels and measuring DNA concentration using the Qubit® DNA Assay Kit with a Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). Total RNA was isolated by utilizing RNAprep Pure Plant Plus Kit (Polysaccharides & Polyphenolics-rich, TIANGEN, Beijing, China) according to the manufacturer's instructions. RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). Sequencing was performed on the Illumina NovaSeq. 6000 platform.

To assemble and annotate haplotype-resolved T2T gapless genomes of yellowhorn varieties PBN-43 and PBN-126, we generated high-depth Illumina paired-end (NGS) reads (mean depth ~60.66×), PacBio HiFi reads (mean depth ~205.27×), high-throughput chromatin conformation capture (Hi-C) sequencing reads (mean depth ~217.48×), and RNA-seq reads for both accessions (Table 1). For Illumina paired-end (NGS) reads, a library was constructed following the manufacturer's protocol and sequenced on the Illumina NovaSeq. 6000 platform. Library preparation involved DNA fragmentation, end-polishing, A-tailing, adapter ligation, PCR amplification, and purification. The libraries were analyzed for size distribution using an Agilent 2100 Bioanalyzer and quantified by real-time PCR. For PacBio HiFi reads, a 15 kb library was constructed using a SMRTbell Express Template Prep Kit 3.0 (Pacific Biosciences, CA, USA). Library preparation included DNA shearing, damage repair, end repair, hairpin adapter ligation, size selection, and purification of the library. After quality control testing, the SMRTbell library was sequenced using a single 25 M SMRT Cell on the PacBio Revio platform (Pacific Biosciences, CA, USA). For Hi-C reads, the library was prepared from cross-linked chromatin isolated from young leaves using a standard Hi-C protocol. Chromatin was cross-linked with formaldehyde, digested with restriction enzyme *DpnII*, labeled with biotin-14-dCTP, and ligated to form proximity ligation products. DNA was then purified, fragmented, and enriched with streptavidin beads before adapter ligation. The resulting library was amplified and sequenced on the DNBSEQ platform to generate 2 × 150 bp paired-end reads.

Genome survey and assembly. The estimation of genome size and heterozygosity was conducted by Jellyfish (v2.2.10)[17] and GCE (v1.0.2)[18] based on k-mer size = 23. The genome assembly process involved several steps. Hifiasm (v0.19.9-r616)[19] was utilized to conduct *de novo* genome assembly by integrating HiFi reads and Hi-C reads with the parameters "-s 0.45 -l 2 -x 0.95 -y 0.6 --telo-m CCCTAAA" to accommodate heterozygosity and assemble more contigs with telomere motifs. Purge_haplotigs (v1.1.3)[20] was then used to improve haplotype phasing by further removing redundant contigs based on hifiasm preliminary assemblies and HiFi reads. Then, Hi-C data were used to anchor the contigs of each haplotype via the Juicer (v1.6)[21] and the 3D-DNA (v201013)[22] pipelines. After this step, almost half of chromosomes in each haplotype reached T2T level. Manual corrections for misassemblies were made using Juicebox (v2.17.00)[23], and short contigs were discarded during this process. Verkko (v2.1)[24] was additionally employed for assembly by integrating PacBio HiFi and Hi-C data to generate continuous long contigs with default parameters. Verkko assembly, hifiasm assembly, and HiFi reads were utilized for closing gaps with the help of quarTeT (v1.1.7)[25] GapFiller module (with default parameter) and Integrative Genomics Viewer (IGV)[26]. These assemblies and HiFi reads were also used for telomere patching using Teloclip pipeline (v0.0.4) (https://github.com/Adamtaranto/teloclip). Finally, PGA (v.0.2) (https://github.com/likui345/PGA) was

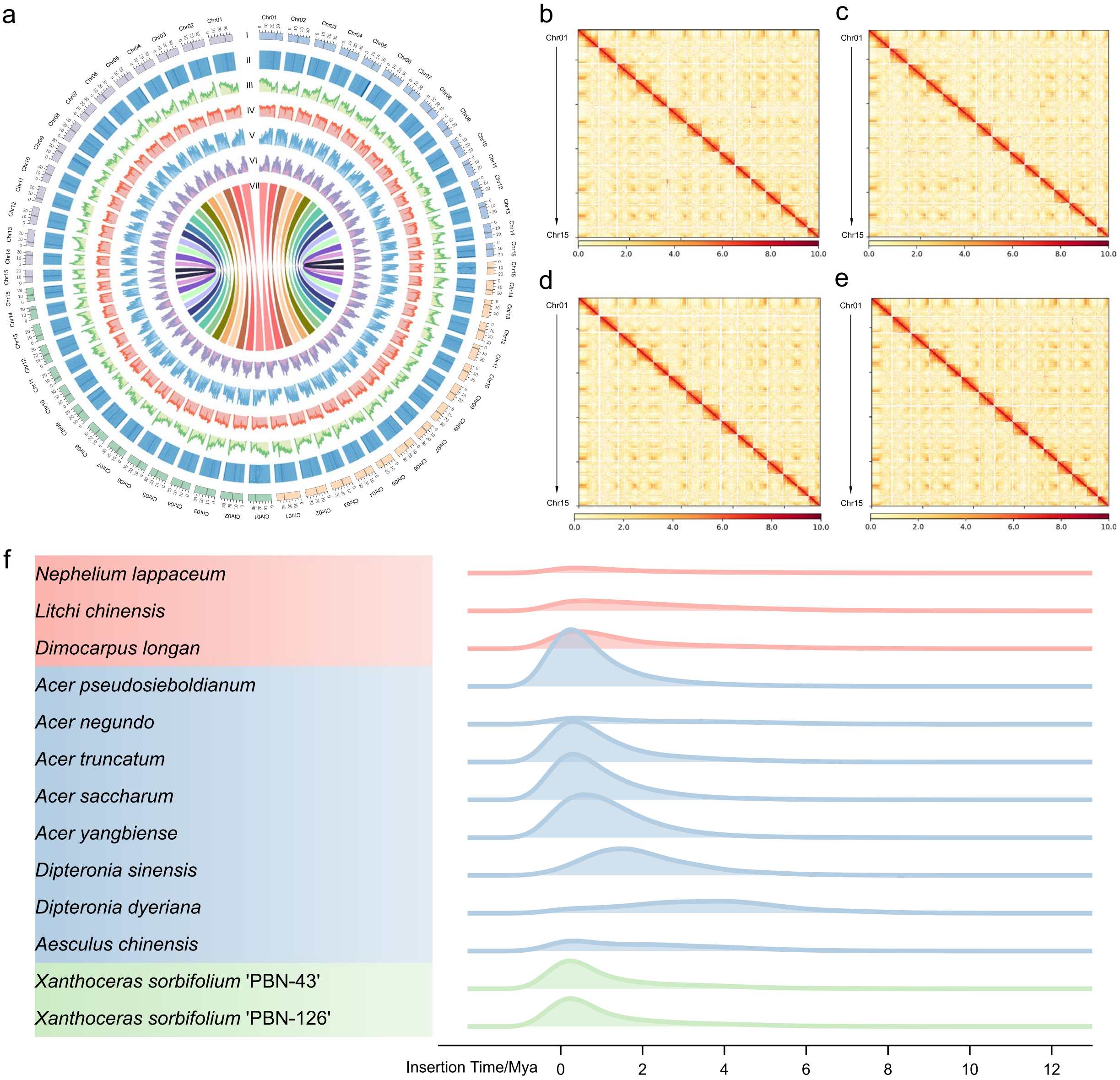

Fig. 1 Comprehensive overview of haplotype-resolved yellowhorn genomes. (**a**) Circos map of each haplotype's genomic features. Including: I. chromosomes (blue bar for PBN-43_hap1, orange bar for PBN-43_hap2, green bar for PBN-126_hap1, and purple bar for PBN-126_hap2; grey segments within the chromosome bars indicate centromere locations), II. GC content, III. gene density, IV. TE density, V. LTR/Copia density, VI. LTR/Gypsy density, VII. syntenic blocks. (**b–e**) Hi-C chromatin interaction heatmaps for PBN-43_hap1 (**b**), PBN-43_hap2 (**c**), PBN-126_hap1 (**d**), and PBN-126_hap2 (**e**), displaying chromosomal interaction frequencies for each haplotype. (**f**) LTR-RT insertion times (in Mya) across selected Sapindaceae species. Clade colors: red for Sapindoideae, blue for Hippocastanoideae, and green for Xanthoceroideae.

used to rearrange the final assemblies according to ZS4 with default parameters. For telomere identification, the quarTeT TeloExplorer module was used with the "-c plant" parameter. For centromere identification, centromere candidate regions were initially generated using the quarTeT CentroMiner module for each chromosome with repetitive elements and gene annotation under default parameters. These candidate regions were further manually determined by integrating information from repeat density, gene density, and Hi-C chromatin interaction heatmaps. After these steps, we obtained four haplotype genomes ranging from 464.34 Mb to 468.93 Mb, all of which reached gapless telomere-to-telomere (T2T) level with high-quality assessment metrics, representing the most comprehensive and complete yellowhorn genomes to date (Fig. 1a–e, Table 2, Supplementary Table S1).

Repetitive elements annotation. We conducted repetitive elements annotation using a combination of *de novo* prediction and homology-based searches. For both varieties, we first utilized RepeatModeler (v2.0.5)[27] to construct a *de novo* repeat library with default parameters. Next, to further refine the library, we classified all sequences labeled as "Unknown" in the *de novo* library using TEsorter (v1.4.6)[28] with default parameters. Additionally, we

		PBN-43_hap1	PBN-43_hap2	PBN-126_hap1	PBN-126_hap2
Assembly	Assembly Length (bp)	467,233,670	464,337,492	468,926,305	465,040,044
	Contig N50 (bp)	29,907,139	27,116,286	22,604,038	21,398,156
	Scaffold N50 (bp)	31,760,580	31,013,014	31,237,015	31,222,858
	Number of telomeres	30	30	30	30
	Number of centromeres	15	15	15	15
	Number of gaps	0	0	0	0
Quality Assessment	QV	67.01	66.79	67.02	66.79
	Completeness (%)	85.97	86.11	86.63	86.68
	Assembly BUSCOs (%)	99.01	98.82	98.95	98.89
	Annotation BUSCOs (%)	98.45	98.64	97.77	98.20
	LAI	18.04	17.64	16.73	16.53
	NGS reads mapping rates (%)	99.25	99.25	99.24	99.33
	HiFi reads mapping rates (%)	100.00	100.00	100.00	100.00
Genes and Repetitive Elements Annotation	Number of genes	35,069	35,054	35,174	35,039
	Average gene length (bp)	3,177	3,162	3,160	3,163
	GC content (%)	35.30	35.21	35.19	35.23
	Repeat elements content (%)	67.92	67.63	67.95	67.79
	Retroelements (bp)	196,168,915	193,606,048	194,035,806	194,267,292
	SINEs (bp)	1,320,181	1,328,923	1,467,265	1,458,220
	LINEs (bp)	16,522,509	16,549,464	12,767,495	12,710,649
	LTR elements (bp)	178,326,225	175,727,661	179,801,046	180,098,423
	LTR/Copia (bp)	76,940,797	76,208,096	82,985,455	82,752,361
	LTR/Gypsy (bp)	85,498,030	83,807,409	80,757,425	81,552,819
	DNA transposons (bp)	108,746,338	108,494,395	113,461,077	109,522,539
	Rolling-circles (bp)	980,283	1,016,332	768,188	878,192
	Small RNA (bp)	1,324,369	1,330,604	1,492,107	1,483,634
	Simple repeats (bp)	6,497,623	5,920,875	5,890,576	6,083,144
	Low complexity (bp)	1,017,948	1,007,166	1,083,071	1,075,734
	Unclassified (bp)	5,907,953	5,975,165	5,646,859	5,627,196
	Total bases masked	318,072,017	314,764,019	319,469,165	316,046,662
Functional Annotation	eggNOG-mapper	29,299	29,304	29,451	29,512
	InterProScan	31,202	31,117	31,289	31,292
	KEGG	10,874	10,882	10,822	10,826
	NCBI NR	31,147	31,269	31,406	31,432
	Uniprot_sprot	20,702	20,713	20,762	20,787
	Uniprot_trembl	30,178	30,231	30,341	30,417
	Total annotated genes	32,944	32,863	33,002	32,961
	Percentage of annotated genes	93.94%	93.75%	93.82%	94.07%

Table 2. Statistics for genome assembly and annotation of PBN-43 and PBN-126.

incorporated Sapindales-specific repeat families from the Dfam database[29] using the famdb.py tool, resulting in an updated species-specific repeat library. Finally, RepeatMasker (v4.1.6) (https://www.repeatmasker.org/) was used to annotate repeat regions, employing both the species-specific repeat library and the RepBase[30] library. We performed genome-wide LTR-RT identification and classification using LTR_retriever (v2.9.5)[31], with outputs from the LTRharvest (v1.6.5)[32] and LTR_FINDER_parallel (v1.1)[33] pipelines. The neutral mutation rate parameter was set to "-u 7e-9" (according to *Arabidopsis*[34]) in LTR_retriever to estimate LTR-RT insertion times.

For yellowhorn, an average of 67.99% (~317.09 Mb) of genomic regions across all haplotype genomes consisted of repetitive elements, particularly long terminal repeat retrotransposons (LTR-RT). The *Gypsy* superfamily was slightly more abundant than the *Copia* superfamily in PBN-43, whereas the opposite pattern was observed in PBN-126. Additionally, unclassified elements were reduced to an all-time low, averaging just 1.24% across all haplotype genomes (Table 2). We also found that most Sapindaceae species experienced LTR-RT insertion events within the last 6 million years (Mya), and several species (e.g., *Dipteronia dyeriana* and *Litchi chinensis*) exhibited complex and extended insertion patterns. Yellowhorn underwent a more recent LTR-RT insertion during the last 0.45–0.48 Mya (Fig. 1f).

Genes annotation. Multiple prediction methods contributed to the final gene annotation. For transcriptome-based prediction, we first used HISAT2 (v2.2.1)[35] to align transcriptome reads to the genomes to generate BAM files with default parameters, which were then sorted using SAMtools (v1.2.0)[36]. Then, two transcriptome-based prediction pipelines were employed: (1) a *de novo* transcript assembly pipeline consisting of

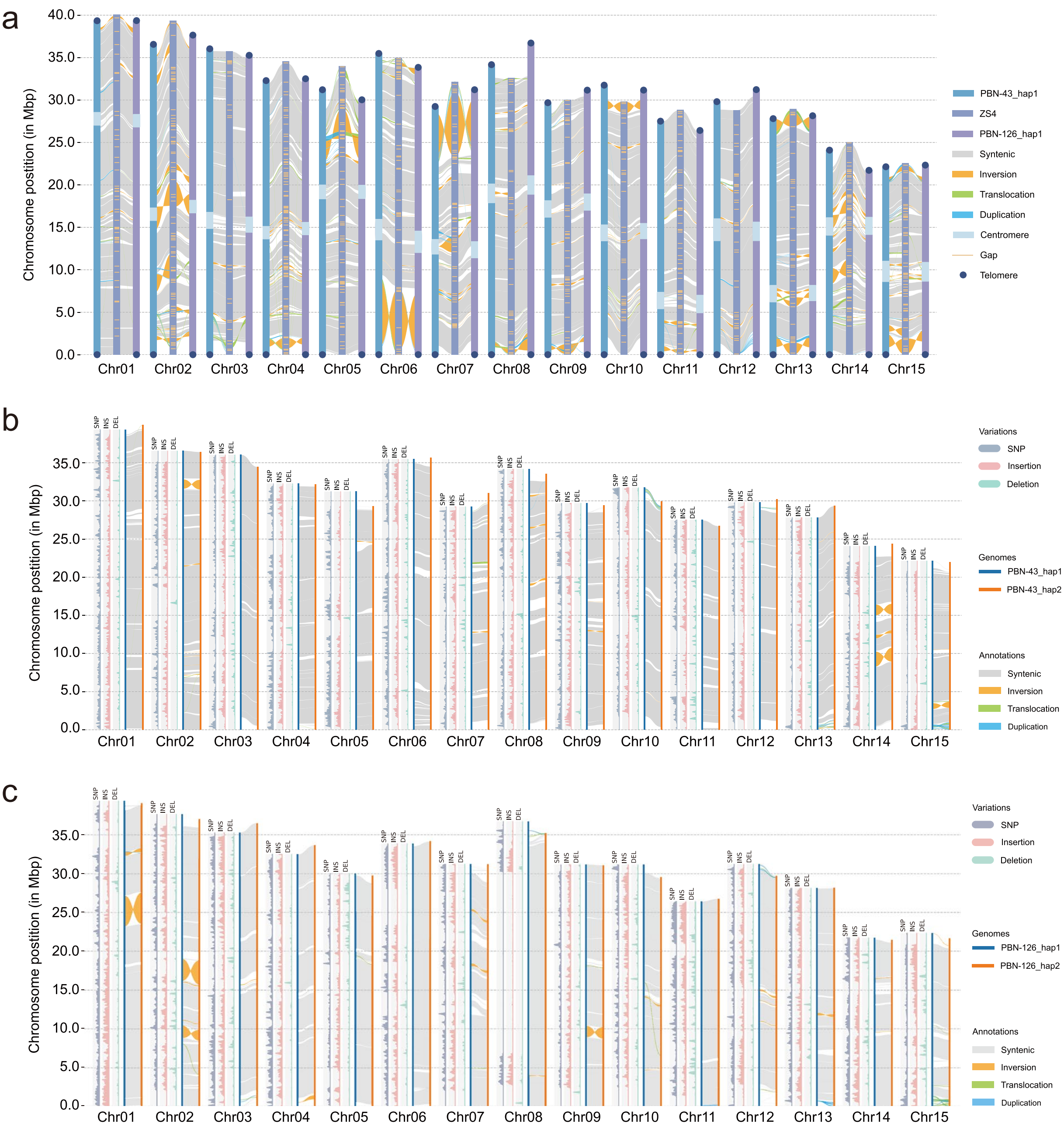

Fig. 2 Genomic variations among different yellowhorn varieties and within the haplotypes. (**a**) Genomic variations among PBN-43_hap1, PBN-126_hap1, and the previously published yellowhorn genome ZS4. (**b**-**c**) Haplotype-resolved variations in PBN-43 (**b**) and PBN-126 (**c**), with SNP and InDel densities calculated per 100 Kb using hap1 as the reference genome.

Trinity (v2.15.1)[37], PASA (v2.5.3)[38], and TransDecoder (v5.7.1) (https://github.com/TransDecoder/TransDecoder); and (2) a reference-guided transcript assembly pipeline using StringTie (v2.2.1)[39] and TransDecoder (v5.7.1). For homology-based prediction, we used a homologous protein dataset containing sequences from *Acer truncatum*, *Aesculus chinensis*, *Arabidopsis thaliana*, *Dimocarpus longan*, *Litchi chinensis*, and *Xanthoceras sorbifolium* "ZS4". Gene models were predicted using miniprot (v0.12-r237)[40] with the parameters "-I –gff –outc 0.8 –outn 1000" and GeMoMa (v1.9)[41] with default parameters. For *ab initio* prediction, we trained and generated gene models using AUGUSTUS (v3.5.0)[42] and GeneMark-ETP (v1.02)[43] via the BRAKER3 (v3.0.3)[44] pipeline, based on hard-masked genomes, aligned RNA-seq reads (BAM files), and the homologous protein dataset described above. After assigning weights to these three prediction methods, all predicted gene models were integrated into consensus gene models using EvidenceModeler (v2.0.0)[45]. For functional annotation, DIAMOND (v2.1.9)[46] was used to search protein sequences against the NCBI non-redundant (NR) protein database[47] and the UniProt

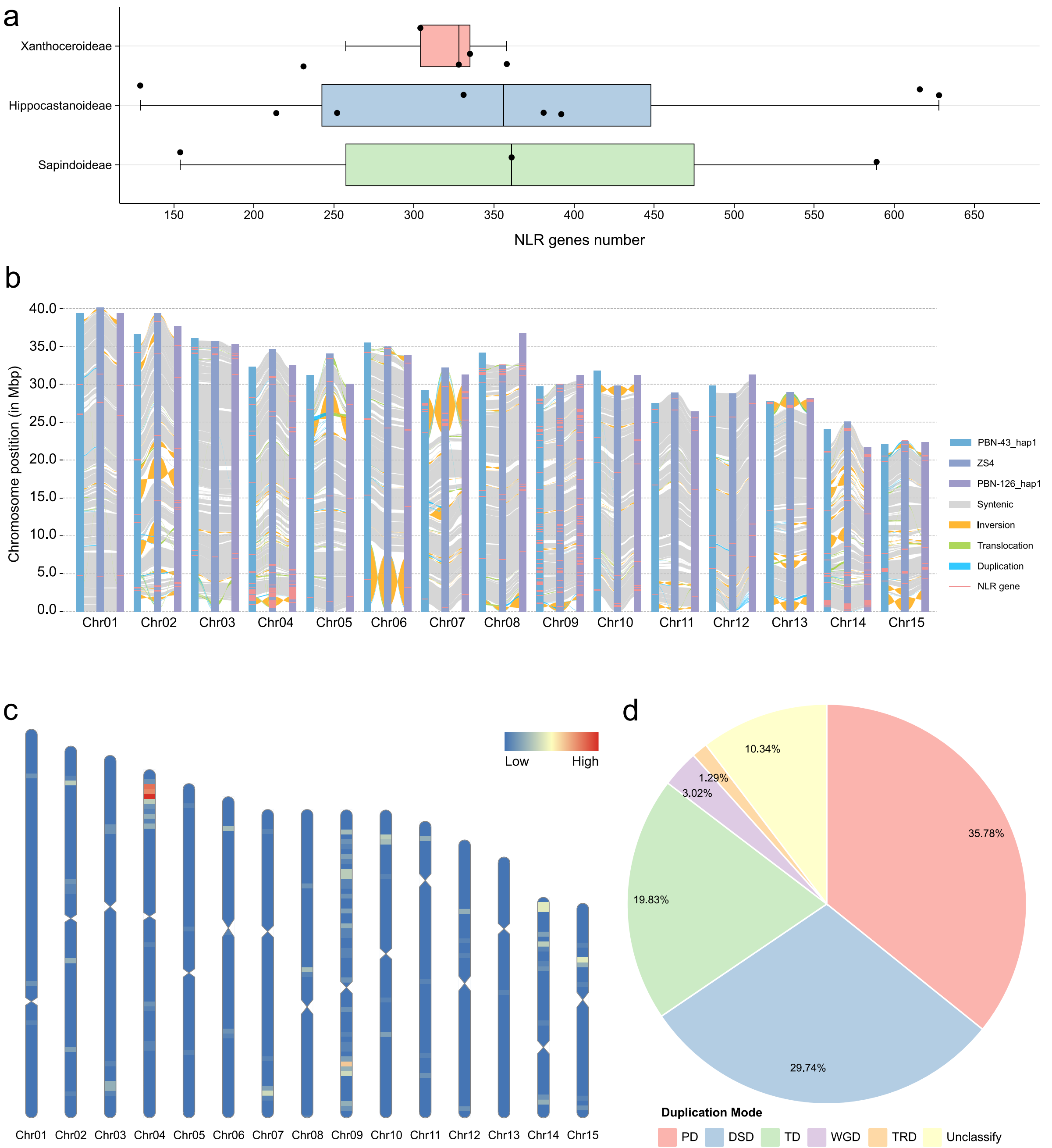

Fig. 3 NLR gene landscape in Sapindaceae species and yellowhorn. (**a**) NLR gene numbers across species from the three Sapindaceae subfamilies species. (**b**) Chromosomal localization of NLR genes in PBN-43, PBN-126, and ZS4 genomes. (**c**) Density of NLR genes along chromosomes in PBN-126_hap1. (**d**) Duplication modes of all NLR genes in both PBN-43 and PBN-126 genomes.

database[48] with the parameters "--evalue 1e-5 --max-target-seqs 1". Additionally, InterProScan (v5.47–82.0)[49], eggNOG-mapper (v2.1.12)[50], and kofam_scan (v1.3.0)[51] were utilized for further functional annotation.

Collectively, an average of 35,084 protein-coding genes was predicted with an average gene length of 3,166 bp. Across all haplotype genomes, at least 93.75% of all genes were functionally annotated (Table 2). Compared with the first published yellowhorn genome ZS4 (23,365 genes), a mean of 11,719 more genes have been predicted in haplotype genomes, and more than 99.8% of ZS4 genes were found in our genomes (Supplementary Table S2).

Genomes comparison and synteny analysis. To capture and catalog variations among haplotypes and previously published genomes, we first used Minimap2 (v2.26-r1175)[52] with the parameters "asm5 -ax –eqx" to align the genomes. Synteny and Rearrangement Identifier (SyRI, v1.6.3)[53] was then applied to identify variations based on the genome alignment results, followed by visualization using plotsr (v1.1.1)[54]. For variation annotation, SnpEff (v5.2-1)[55] databases were constructed for both varieties and used with default parameters on the VCF files generated by SyRI. Variations classified as "High" and "Moderate" impact by SnpEff were regarded as deleterious variations. With the haplotype-resolved yellowhorn genomes assembled, alleles were identified directly from the genome sequences using the JCVI (v0.0.0)[56] pipeline and manual curation. Gene annotation files for hap1 and hap2 were converted into BED format using the jcvi.formats.gff module with default parameters, and orthologous gene pairs were detected using the jcvi.compara.catalog module with the parameter --cscore=0.99, serving as preliminary allele pairs. Redundant pairs were manually removed based on coding protein sequence similarity and chromosomal location, particularly removing interchromosomal pairs, to produce the final allele set.

To investigate genomic variation among different yellowhorn varieties, we chose PBN-43_hap1 and PBN-126_hap1 as reference genomes and aligned them with each other, as well as individually compared them to the previously published yellowhorn genome ZS4. In comparing PBN-43_hap1 vs. PBN-126_hap1, we detected more syntenic regions (343.90 Mb and 399.72 Mb) but fewer inversions (43.89 Mb and 10.66 Mb) and translocations (10.29 Mb and 4.60 Mb) than in comparisons involving ZS4, indicating greater genomic similarity between these two varieties (Fig. 2a, Supplementary Table S3). Although the SNP density (5.31 SNPs/Kb and 5.23 SNPs/Kb) was slightly higher for PBN-43_hap1 vs. PBN-126_hap1, fewer of these SNPs resulted in deleterious variations such as frameshifts or stop-gained mutations, and more occurred in intergenic regions. Likewise, the numbers of insertions (105,700 and 129,675) and deletions (105,901 and 120,709) were significantly lower than in alignments with ZS4, and structural variants (SVs) showed a similar trend (Supplementary Tables S3, S4). We also identified several large-scale SVs, especially on Chr02, Chr06, and Chr07, which were supported by HiFi reads spanning the junction, confirming the accuracy of the assemblies (Fig. 2a, Supplementary Fig. S3).

At the haplotype level, variations between haplotypes for PBN-43 and PBN-126 were identified and annotated, and a total of 27,388 (PBN-43) and 27,888 (PBN-126) allelic genes were identified (Fig. 2b,c, Supplementary Tables S3, S4). Several large-scale SVs were detected among haplotypes and confirmed by HiFi read coverage on junction loci (Supplementary Fig. S3). Additionally, both the number of variations and deleterious variations in PBN-126 haplotypes is lower than those in PBN-43, which corresponds to the heterozygosity rate differences (Supplementary Table S3, S4). Notably, the previously discovered *AG* gene variation[5], which leads to the double-flowered phenotype, is present in both alleles in PBN-126 and we found that it is actually a 342 bp *LINE-RH* insertion instead of a *LINE1* (Supplementary Fig. S4).

Identification of NLR genes.

NLR-annotator (v2.1b)[57] was used to identify NLR domains across the genomes of 12 Sapindaceae species with default parameters. Genes were classified as NLR genes if they contained at least an NB-ARC domain, a C-terminal LRR domain, and an N-terminal domain of either Toll/interleukin-1 receptor (TIR), coiled-coil (CC), or RPW8, as determined by InterProScan functional annotation. Classification of NLR genes was based on the type of N-terminal domain.

In this study, we conducted genome-wide identification of NLR genes in both haplotypes of PBN-43 and PBN-126, as well as in the genomes of 11 other Sapindaceae species. Across the Sapindaceae family, the number of NLR genes varies tremendously, ranging from 90 in *Aesculus chinensis* to 589 in *Dimocarpus longan* (Fig. 3a, Supplementary Tables S5, S6). This variation aligns with the species-specific mechanisms of NLR gene expansion and contraction characteristic of flowering plants[58]. In yellowhorn, we identified 211/252 NLR genes in PBN-43 hap1/hap2 and 232/260 in PBN-126 hap1/hap2. Comparison of NLR gene locations among PBN-43, PBN-126, and ZS4 revealed that the main differences lie in some dispersed NLR genes, such as those on Chr09 (Fig. 3b). Regarding the NLR gene densities along chromosomes, we found that they tend to form gene clusters in yellowhorn, especially on the short arms of Chr04 and Chr09. Other chromosomes show a sparse distribution of NLR genes, reflecting that "hotspots" of NLR genes remain consistent among haplotypes and varieties (Fig. 3c, Supplementary Fig. S5). This phenomenon aligns with the duplication mode of yellowhorn NLR genes, with proximal and tandem duplications constituting the majority (Fig. 3d). NLR gene clusters may provide functional redundancy, allowing multiple genes in the same region to perform similar immune functions. In summary, this study presents for the first time complete haplotype-resolved genomes of yellowhorn, laying a foundation for future research into its ornamental traits, breeding potential, and broader genomic studies.

Data Records

The raw sequence data (PacBio HiFi, Illumina paired-end, Hi-C, and RNA-seq) reported in this paper have been deposited in the Genome Sequence Archive (GSA) of the National Genomics Data Center (NGDC, https://ngdc.cncb.ac.cn) under the accessions CRA021629[59], CRA021628[60], CRA021631[61], and CRA021624[62], respectively, with the BioProject accession PRJCA033489670[63]. These data have also been deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov) under the accession SRP573106[64], with the BioProject accession PRJNA1200135[65]. The assembled genomes have been deposited in the NCBI GenBank under the GenBank accessions JBMIRF000000000[66] (PBN-43_hap1), JBMIRG000000000[67] (PBN-43_hap2), JBMRGO000000000[68] (PBN-126_hap1), and JBMRGN000000000[69] (PBN-126_hap2). For broader accessibility, we have also deposited the corresponding files including genome assemblies, genome annotation, functional annotation, and VCF files at Figshare[70].

Technical Validation

Genome assembly and annotation quality assessment.

We assessed the quality and completeness of the genome assemblies in multiple ways. First, HiC-Pro (v3.1.0)[71] was used to align Hi-C reads to the final assemblies and generate matrix files with default parameters. These were then visualized using HiCPlotter (v0.6.02)[72] to create Hi-C chromatin interaction maps. We checked Hi-C chromatin interaction maps and found no significant contig misassemblies in any of the genome assemblies (Fig. 1b–e). For mapping rates, PacBio HiFi reads and NGS reads were aligned to each haplotype genome using Minimap2 and BWA-MEM (v0.7.18)[73], respectively. Mapping rates and genome coverages were calculated using SAMtools, resulting in an overall mapping rate of over 99%. Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.5.0)[74] was employed to assess the completeness of gene regions with the embryophyta_odb10 database (2024-01-08), and 98.82% to 99.01% of 1,614 plant core conserved genes were complete in all assemblies. Merqury (v1.3)[75] was used to calculate QV based on a 23-mer database generated from HiFi reads using the meryl tool, and the QV values of all assemblies ranges from 66.79 to 67.02, indicating the accuracies were >99.99%. Additionally, LTR_retriever utilized outputs from LTRharvest and LTR_FINDER_parallel to generate LTR Assembly Index (LAI) for evaluating the assembly continuity. LAI scores of the four haplotype genomes range from 16.53 to 18.04, which can be categorized as "Reference" level. Collectively, these assessments suggest that we have produced high-quality haplotype-resolved yellowhorn genome assemblies in terms of contiguity, accuracy, and completeness (Table 2). The annotated genes achieved an average BUSCO completeness of 98.27% based on the embryophyta_odb10 database (2024-01-08), and an average of 93.90% of all genes were functionally annotated across all assemblies, indicating high-quality annotations (Table 2).

Code availability

All analyses were conducted according to the manuals or protocols of the corresponding softwares and pipelines, as described in the Methods section, along with their versions. No custom code was generated for these analyses.

Received: 17 January 2025; Accepted: 23 April 2025;
Published online: 14 May 2025

References

1. Zong, J. *et al*. Growth, physiological, and photosynthetic responses of *Xanthoceras sorbifolium* bunge seedlings under various degrees of salinity. *Front. Plant Sci.* **12**, 730737 (2021).
2. Zhou, C., Wu, H., Sheng, Q., Cao, F. & Zhu, Z. Study on the phenotypic diversity of 33 ornamental *Xanthoceras sorbifolium* cultivars. *Plants* **12**, 2448 (2023).
3. Ruan, C. *et al*. The importance of yellow horn (*Xanthoceras sorbifolia*) for restoration of arid habitats and production of bioactive seed oils. *Ecological Engineering* **99**, 504–512 (2017).
4. Ma, Y. *et al*. Provenance variations in kernel oil content, fatty acid profile and biodiesel properties of *Xanthoceras sorbifolium* Bunge in northern China. *Industrial Crops and Products* **151**, 112487 (2020).
5. Wang, H. *et al*. The double flower variant of yellowhorn is due to a *LINE1* transposon-mediated insertion. *Plant Physiol* **191**, 1122–1137 (2022).
6. Liu, Z. *et al*. Identification of yellowhorn (*Xanthoceras sorbifolium*) WRKY transcription factor family and analysis of abiotic stress response model. *J. For. Res.* **32**, 987–1004 (2021).
7. Wang, J. *et al*. 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 (2023).
8. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444**, 323–329 (2006).
9. Huang, H. *et al*. Telomere-to-telomere haplotype-resolved reference genome reveals subgenome divergence and disease resistance in triploid Cavendish banana. *Horticulture Research* **10**, uhad153 (2023).
10. Yang, T. *et al*. A telomere-to-telomere gap-free reference genome assembly of avocado provides useful resources for identifying genes related to fatty acid biosynthesis and disease resistance. *Horticulture Research* **11**, uhae119 (2024).
11. Kourelis, J. & van der Hoorn, R. A. L. Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *The Plant Cell* **30**, 285–299 (2018).
12. Bi, Q. *et al*. Pseudomolecule-level assembly of the Chinese oil tree yellowhorn (*Xanthoceras sorbifolium*) genome. *Gigascience* **8**, giz070 (2019).
13. Liang, Q. *et al*. The genome assembly and annotation of yellowhorn (*Xanthoceras sorbifolium* Bunge). *Gigascience* **8**, giz071 (2019).
14. Liu, H. *et al*. 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 (2021).
15. Liang, Q. *et al*. Genomic and transcriptomic analyses provide insights into valuable fatty acid biosynthesis and environmental adaptation of yellowhorn. *Front Plant Sci* **13**, 991197 (2022).
16. Bi, Q. *et al*. The phased chromosome-scale genome of yellowhorn sheds light on the mechanism of petal color change. *Horticultural Plant Journal* **9**, 1193–1206 (2023).
17. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
18. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. Preprint at https://doi.org/10.48550/arXiv.1308.2012 (2020).
19. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
20. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
21. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98 (2016).
22. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
23. Rao, S. S. P. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
24. Rautiainen, M. *et al*. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474–1482 (2023).
25. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* **10**, uhad127 (2023).
26. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192 (2013).
27. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
28. Zhang, R. *et al*. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* **9**, uhac017 (2022).

29. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Research* **44**, D81–D89 (2016).
30. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
31. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422 (2018).
32. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
33. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48 (2019).
34. Ossowski, S. *et al*. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
35. Kim, D. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, (2019).
36. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* **29**, 644–652 (2011).
38. Haas, B. J. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
39. Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLOS Computational Biology* **18**, e1009730 (2022).
40. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
41. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* **44**, e89–e89 (2016).
42. Stanke, M. *et al*. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
43. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768 (2024).
44. Gabriel, L. *et al*. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
45. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
46. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
47. Wheeler, D. L. *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Research* **36**, D13–D21 (2008).
48. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515 (2019).
49. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
50. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).
51. Aramaki, T. *et al*. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
53. Goel, M., Sun, H., Jiao, W. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277 (2019).
54. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926 (2022).
55. Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
56. Tang, H. *et al*. JCVI: a versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
57. Steuernagel, B. *et al*. The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
58. Jacob, F., Vernaldi, S. & Maekawa, T. Evolution and Conservation of Plant NLR Functions. *Front. Immunol.* **4**, (2013).
59. *NGDC Genome Sequence Archive.* https://ngdc.cncb.ac.cn/gsa/browse/CRA021629 (2025).
60. *NGDC Genome Sequence Archive.* https://ngdc.cncb.ac.cn/gsa/browse/CRA021628 (2025).
61. *NGDC Genome Sequence Archive.* https://ngdc.cncb.ac.cn/gsa/browse/CRA021631 (2025).
62. *NGDC Genome Sequence Archive.* https://ngdc.cncb.ac.cn/gsa/browse/CRA021624 (2025).
63. *NGDC BioProject.* https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA033489 (2025).
64. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRP573106 (2025).
65. *NCBI BioProject.* https://identifiers.org/ncbi/bioproject:PRJNA1200135 (2025).
66. Chen, Y. Xanthoceras sorbifolium isolate PBN-43, whole genome shotgun sequencing project. Haplotype 1. *NCBI GenBank.* https://identifiers.org/ncbi/insdc:JBMIRF000000000 (2025).
67. Chen, Y. Xanthoceras sorbifolium isolate PBN-43, whole genome shotgun sequencing project. Haplotype 2. *NCBI GenBank.* https://identifiers.org/ncbi/insdc:JBMIRG000000000 (2025).
68. Chen, Y. Xanthoceras sorbifolium isolate PBN-126, whole genome shotgun sequencing project. Haplotype 1. *NCBI GenBank.* https://identifiers.org/ncbi/insdc:JBMRGO000000000 (2025).
69. Chen, Y. Xanthoceras sorbifolium isolate PBN-126, whole genome shotgun sequencing project. Haplotype 2. *NCBI GenBank.* https://identifiers.org/ncbi/insdc:JBMRGN000000000 (2025).
70. Chen, Y. Two haplotype-resolved telomere-to-telomere genome assemblies of *Xanthoceras sorbifolium*. *figshare* https://doi.org/10.6084/m9.figshare.28188326.v2 (2025).
71. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).
72. Akdemir, K. C. & Chin, L. HiCPlotter integrates genomic data with interaction matrices. *Genome Biol* **16**, 198 (2015).
73. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
74. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
75. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).

Acknowledgements

This study was supported by funding from the Key R&D Program of Shandong Province, China (2024CXPT034), and the project SYS202206 supported by Shandong Provincial Natural Science Foundation. This study was also supported by the Taishan Scholars Program. We would like to thank the Bioinformatics Platform at Peking University Institute of Advanced Agricultural Sciences for providing the high-performance computing resources.

Author contributions

X.H., H.J., S.L. and H.H. conceived and designed the study. Y.L., Y.C., Z.R., K.L. and K.W. performed the experiments, conducted the bioinformatic analyses, and prepared the data visualizations. Y.L., Y.C., Z.R. and X.W. carried out the plant tissue sampling. S.L. provided plant materials and, together with J.L., guided the sampling process. N.S. offered suggestions for this research. All authors contributed to the manuscript, read, and approved the final version for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05057-x.

Correspondence and requests for materials should be addressed to H.H., S.L., H.J. or X.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.