BMC Genomics

RESEARCH Open Access

High-quality chromosome-level genomic insights into molecular adaptation to low-temperature stress in *Madhuca longifolia* in southern subtropical China

Shuyu Wang[1], Haoyou Lin[1], Shuiyun Ye[1], Zhengli Jiao[2], Zhipeng Chen[1], Yifei Ma[1] and Lu Zhang[1*]

Abstract

Background *Madhuca longifolia*, the energy-producing and medicinal tropical tree originally from southern India, faces difficulties in adapting to the low temperatures of late autumn and early winter in subtropical southern China, impacting its usability. Therefore, understanding the molecular mechanisms controlling the ability of this species to adapt to environmental challenges is essential for optimising horticulture efforts. Accordingly, this study aimed to elucidate the molecular responses of *M. longifolia* to low-temperature stress through genomic and transcriptomic analyses to inform strategies for its effective cultivation and utilisation in colder climates.

Results Herein, the high-quality reference genome and genomic assembly for *M. longifolia* are presented for the first time. Using Illumina sequencing, Hi-C technology, and PacBio HiFi sequencing, we assembled a chromosome-level genome approximately 737.92 Mb in size, investigated its genomic features, and conducted an evolutionary analysis of the genus *Madhuca*. Additionally, using transcriptome sequencing, we identified 17,941 differentially expressed genes related to low-temperature response. Through bioinformatics analysis of the *WRKY* gene family, 15 genes crucial for *M. longifolia* low-temperature resistance were identified.

Conclusions This research not only lays the groundwork for the successful ecological adaptation and cultivation of *M. longifolia* in China's southern subtropical regions but also offers valuable insights for the genetic enhancement of cold tolerance in tropical species, contributing to their sustainable horticulture and broader industrial, medicinal, and agricultural use.

Keywords Whole genome, Transcriptome, *WRKY*, Tropical tree species

*Correspondence:
Lu Zhang
zhanglu@scau.edu.cn
[1]College of Forestry and Landscape Architecture, South China Agricultural University, Guangzhou, Guangdong 510642, China
[2]School of Life Sciences, Guangzhou University, Guangzhou 510006, China

Background

Ecological adaptation is a key mechanism for species to cope with environmental changes. Understanding the molecular mechanisms involved in adaptation is crucial for cultivating and utilising individual species. Temperature is an abiotic factor that considerably affects plant growth, development, and physiological activities. To mitigate the damage resulting from low temperatures, plants have developed cold acclimation mechanisms that involve a complex set of physiological and biochemical responses to environmental factors [1].

Madhuca longifolia (J. Koenig ex L.) J. F. Macbr, called 'Mahua' in India, is a broadleaf evergreen tree species originating from southern India and Burma and belonging to the family Sapotaceae. The species produces edible fruit and is a source of hardwood. *M. longifolia* is considered a panacea in Indian traditional medicine, with its leaves, flowers, seeds, and bark utilised as medicines [2–5]. Moreover, studies of *M. longifolia* suggest promising roles in food processing, renewable energy, urban greening and other fields. The flowers of the tree are universally utilised for food, feed, and fuel [6]. Its seed oil is a promising non-edible oil that can be employed to improve the quality of bitumen [7]. *M. longifolia* provides a solution for the three major "Fs", namely food, forage and fuel [8]. This plant can be considered an effective agent against oral diseases like dental caries [9]. A nano-composite material prepared from the seed extract of *M. longifolia* has a remarkable insecticidal effect on vector larvae and can be used to control mosquitos and other major vectors as an eco-friendly substitute for modern chemical synthetic insecticides [10]. Furthermore, the bark fibre of *M. longifolia* has robust tensile strength, light weight, and good thermal steadiness and is appropriate for manufacturing biodegradable materials for sporting goods, automotive body panels, wallboards, partitions, and non-structural lightweight components in the construction industry [11]. The tree has a strong tolerance to air pollution and can help in its alleviation through its high capacity for absorbing greenhouse gases and other air pollutants. It is one of the most promising street trees for urban greening in the humid tropics [12].

Valued for its industrial and medicinal uses, *M. longifolia* was introduced to China in 1964, where it has shown potential for anti-typhoon forest development in Hainan province [13]. However, climatic differences can lead to unforeseeable outcomes when attempting to acclimatise woody species to locations beyond their native habitat [14]. In China, *M. longifolia* faces challenges in temperature acclimatisation, with low temperatures causing growth stagnation and mortality in seedlings. Our team's previous research found that after the introduction of *M. longifolia* to the southern subtropics in China, there is a risk of the air temperature plummeting to around 10 ℃ during the cold snap in late autumn and early winter when the buds emerge from the soil. Once this happens, even if the air temperature rises after a few days, it will cause harm to the young buds, showing that the young buds and young leaves lose water and wilt, leading to their growth stagnation, and even causing their death in serious cases. To date, there have been no reports on the temperature adaptation of *M. longifolia* after its introduction to South China, and there are very few studies on its genomics, and information on ploidy, chromosome number and genome size is unknown. This absence of whole-genome data has limited research on its phylogenetic origin and evolutionary history and hindered attempts to improve breeding, ecological adaptation, and related biological aspects.

Plant genome sequencing serves as a highly useful tool for studying the molecular mechanisms underlying plant adaptation and determining the impacts of environmental stressors on the evolution, growth, and development of plants as well as allowing for the domestication of useful traits [15]. For example, by sequencing and screening possible genes linked with flowering, growth, and responses to osmotic pressure and temperature stress, Sork et al. [16] revealed important insights into the spatial selection of climate-related genes within natural populations of *Quercus lobata*, highlighting potential environmental adaptation mechanisms of the species. Similarly, the construction of a high-quality reference genome for *Corylus heterophylla* elucidated the molecular mechanisms underlying its response to environmental stress and informed genetic guidelines for optimised breeding [17].

In addition to genome sequencing, RNA-sequencing (RNA-seq) is an important tool for analysing differential gene expression with the transcriptome and understanding genome function, precisely determining gene expression and supporting precise bioinformatic analyses [18, 19]. Transcriptome analyses have been conducted to screen plant genomes for functional genes regarding low-temperature stress resistance and to obtain high-throughput expression data for these key genes. For instance, a transcriptome analysis of cold-tolerant *Zea mays* under cold stress yielded 43 million high-quality sequences, from which a weighted gene co-expression network analysis recognised *Zm00001d037590* and *Zm00001d012321* as the most likely key genes concerning cold hardiness at the seedling stage [20]. Similarly, transcriptome sequencing analysis of *Eremochloa ophiuroides* showed that the expression of genes encoding AUX_IAA as well as *WRKY* and heat shock factor (HSF) transcription factors (TFs) increased with different low-temperature stress treatments [21]. A transcriptome and weighted gene co-expression network analysis of *Ilex dabieshanensis* after cryogenic treatments of different

durations revealed 5,750 differentially expressed genes (DEGs), among which the hub genes for stress response to low temperature were *evm.TU.CHR1.1507* and *1821* and *evm.TU.CHR2.210*, *244*, and *89* [22].

TFs assume an essential role in governing the transduction of signals and the management of gene expression in response to environmental stress. Identification of TFs is infuenced by the annotation quality of genome [23]. The developments of whole genome sequencing in walnut and woody species have revealed evidence of cold and chilling stress and the genome-wide identification of gene families related to stress studies [24]. The *WRKY* gene family is specific to algae and higher plants and greatly affects many plant life processes, particularly in response to biological stress [25, 26]. In *Oryza sativa*, *OsWRKY71* acts as a beneficial controller of responses to low-temperature stress, enhancing the photosynthesis and survival of plants [27]. In contrast, in *Arabidopsis thaliana*, *AtWRKY34* negatively influences plant growth in cold temperatures [28]. In *Musa acuminata* fruits, four *MaWRKY*s enhance low-temperature resistance through an abscisic acid (ABA)-mediated signalling pathway [29]. Based on these examples, studying the *WRKY* gene family through transcriptome analysis is a particularly effective method for studying plant stress resistance at the molecular level. Despite their value, genomic and transcriptomic analyses have not yet been used to elucidate the molecular mechanisms underlying low-temperature stress responses in *M. longifolia.*

The present study aimed to identify the key genes controlling the low-temperature response of *M. longifolia* through transcriptome and whole-genome sequencing. The results could provide a theoretical basis for the comprehensive cultivation and utilisation of this species in subtropical China and a reference for future ecological studies of *Madhuca*-related species.

Methods

A comprehensive methodology was employed to explore the genomic and transcriptomic responses of *M. longifolia* to low-temperature stress. The approach included the cultivation of *M. longifolia* from seeds, isolation of genomic DNA for sequencing and assembly, and genome annotation to identify coding and non-coding regions. A phylogenetic analysis was conducted to place *M. longifolia* in its evolutionary context. Transcriptome sequencing of seedlings under low-temperature conditions was performed to analyse gene expression changes, focusing particularly on the *WRKY* gene family owing to its role in stress response. Quantitative real-time PCR (qPCR) was utilised to verify key findings obtained from RNA-seq. This methodology aimed to elucidate the molecular mechanisms underpinning *M. longifolia*'s adaptation to low temperatures.

Sample collection, genomic DNA extraction and chromosome counts

The mature seeds of *M. longifolia* that had fallen to the ground were collected at South China National Botanical Garden, Tianhe District, Guangzhou, Guangdong, China (113°21′50″E, 23°11′7.3″N). *Madhuca longifolia* seeds were germinated on wet paper towels and then transplanted to peat soil for cultivation after the embryonic roots emerged. The cultivation site was outside the College of Forestry and Landscape Architecture, South China Agricultural University, Tianhe District, Guangzhou City, Guangdong Province, China (113°21′20″E, 23°9′44″N). From among the artificially grown seedlings, we randomly selected a well-grown, healthy, and pest-free specimen and collected its healthy mature leaves for DNA collection. Genomic DNA extraction was conducted utilising cetyltrimethylammonium bromide (CTAB) method. The leaves were grinded into fine powder by liquid nitrogen and then transfer approximately 100 mg to a pre-cooled 2 mL centrifuge tube. After mixing with 1 mL pre-warmed CTAB extract and 20 μL β-Mercaptoethanol, the samples were incubated in a 65 ℃ constant temperature water bath for 1 h, and the samples were mixed upside down several times during the water bath. Then the samples were cooled to room temperature and centrifuged at room temperature, 12,000 rpm for 10 min. The supernatant was mixed with an equal volume of chloroform: isoamyl alcohol (24:1) and centrifuged at room temperature for 10 min at 12,000 rpm. Then the supernatant was mixed with an equal volume of phenol: chloroform: isoamyl alcohol (25:24:1) and centrifuged at room temperature for 10 min at 12,000 rpm. The supernatant was again aspirated and mixed with an equal volume of chloroform: isoamyl alcohol (24:1) and centrifuged at room temperature and 12,000 rpm for 10 min. The nucleic acid was precipitated with isopropanol, washed with 75% ethanol and dissolved in 50 μL ddH_2O. DNA quality was confirmed using NanoDrop and Qubit spectrophotometers (Thermo Fisher Scientific, Waltham, MA, USA). Lastly, 1% agarose gel electrophoresis was employed for testing the sample DNA integrity. The seedlings of 10-month-old *M. longifolia* were selected for chromosome counts. At approximately 9:00 am when the meristems of the plant root tips were flourishing, a 1–2 cm section was excised from the root tips using a blade. The apical materials were treated with 0.002 mol/L quinolin-8-ol solution for 1.5 h. The plant material was rinsed twice with distilled water, transferred to Carnot's fixative for 24 h, and then washed twice with water for 20 min each. Then 2.5% cellulase and 2.0% pectinase solution were added and treated for 2.5 h at 37 ℃. After removing the enzyme solution, the sample was rinsed with distilled water for 2–3 times and let stand in the water for more than 40 min. The treated

root tip was placed on a glass slide and the root cap and elongation area were excised, leaving only the meristems. Finally, the prepared specimen was stained in basic fuchsin dye solution for 10–15 min, and then dried for chromosome counting in root tip cells. The micrographs were taken with an CX33 microscope camera system (Olympus).

Genome sequencing and assembly

Quality-checked genomic DNA was interrupted with Yeasen enzyme digestion kit (Hieff NGS® OnePot TM II DNA Library Prep Kit for MGI®). A total of 1 μg of genomic DNA, 10 μL of Smearase® Mix, and ddH_2O was added to the PCR system to a total volume of 60 μL. The PCR programme was 4 ℃ for 1 min, 30 ℃ for 15 min, 72 ℃ for 20 min and 4 ℃ hold. Then, the NEBNext Ultra DNA Library Prep Kit library (New England Biolabs, Ipswich, MA, USA) was adopted for repairing ends and adding A-tails and Illumina sequencing connectors. DNA fragments of 300–400 bp in length were concentrated by PCR; an AMPure XP system (Beckman Coulter, Brea, CA, USA) was used for PCR product purification. The library was assayed using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA), and qPCR was used for quantification, followed by sequencing using a NovaSeq 6000 sequencer (Illumina, San Diego, CA, USA) according to the PE 150 sequencing strategy. Initial data were filtered using the Illumina platform with fastp (version 0.18.0) [30] for the eradication of reads with ≥10% unknown nucleotides (N), 50% reads with a 20 base-Phred quality score, and those containing connectors. Next, we performed k-mer analysis using high-quality reads. Jellyfish (version 2.2.6) [31] was employed to predict the genomic features (size, heterozygosity rate, and repeat content) according to k-mer (k=21) distributions. Jellyfish used default parameters.

Genomic DNA was processed into fragments of 8–10 kb in length utilising the G-tube method (Covaris, Woburn, MA, USA), followed by DNA fragment end repair. Exonuclease III and IV digestions were performed to remove DNA with gaps or not joined to the ring junction and to remove the junction dimer. To improve sequencing quality and obtain longer average insert fragments, the library was fragmented using the BluePippin Nucleic Acid Fragment Recovery System (Sage Science, Beverly, MA, USA) to remove short fragments of library molecules. After library construction, the Qubit instrument was used for quality assurance. An Agilent 2100 system was employed to evaluate the size of the insert, followed by sequencing using the PacBio platform. To guarantee the reliability and accuracy of the follow-up analysis, the subreads obtained from the raw sequencing reads after removing sequence junctions were considered clean data, and the length distribution of the subreads was used as the main content to evaluate the sequencing effect. Because of the long length and minimal error rate of the PacBio HiFi data, Hifiasm (version 0.15.1-r334) [32] was utilised to splice and assemble the triple sequencing data. The initial genomic assembly integrity assessment was performed using the core eukaryotic gene mapping approach (CEGMA) and benchmarking universal single-copy orthologs (BUSCO) approaches.

Chromosome-level assembly with Hi-C data and evaluation

Filtered reads were compared with the preliminary sequencing results obtained from the assembly of HiFi utilising the MEM algorithm with the Burrows–Wheeler Aligner (version 0.7.12) [33], and a scaffold was established according to interactions between sequences. Scaffolds were sorted and oriented to acquire the ultimate quasi-chromosome-level genome. Hi-C data were analysed using LACHSIS (version 2014-09-12.12) [34], ALLHIC (version 0.9.8) [35] and 3D-DNA (version 180114) [36]. The construction of interaction matrices, calibration of chromosome-constructed genomes, and evaluation of results were performed via ICE software [37]. BUSCO (version 4) was applied for sequence integrity evaluations [38].

Genome annotation

Repeated sequence annotation

In order to use homology to identify genomic repetitive sequences, we matched the genomic sequences with existing databases of repetitive sequences. Homology alignment of the genome sequences and repeats from the Repbase library (version 19.05, http://www.girinst.org/repbase) was performed through ProteinMask and RepeatMasker (version open-4.05) [39], and detected repeat sequences were annotated. Multiple copies of genomic repeated sequences were detected by cross-matching sequences. Using the amino acid sequences of plants in the uniprot library/closely related species as homologous proteins, the genomes aligned by spaln were obtained as a result of homology prediction (spaln: -XQ90 -Q7 -O0 -LS -ya0 -yX2 -d spaln). The Augustus was trained using RNAseq+homology prediction results as a training set (etraining: default parameter). Evidence-based *de novo* prediction was performed using Augustus (augustus: --UTR=off --alternatives-from-evidence=true --allow_hinted_splicesites=atac, gcag --softmasking=1 --gff3=on).

First, we used PILER (version 1.0) [40], RepeatModeler (version 2.0.1) [41], and RepeatScout (version 1.05) [42] to retrieve multiple copies of sequences in the genome through internal alignment of the genome and established a repeat library *de novo*. We removed redundant repeats established from scratch and filtered out

misidentifications, thus establishing a repeat sequence library for *M. longifolia*. We used the obtained repeat sequence library as a reference and again used RepeatMasker to recognise the genomic repeat regions through homologous alignment. LTR_FINDER (version 1.0.7) [43] was utilised to scan and extract the sequences of long terminal repeat (LTR) transposons in the genome, and RepeatMasker was used to annotate their position information. TRF software (version 4.04) [44] was employed to locate simple sequences in tandem.

Coding gene annotation

Protein-coding gene prediction in *M. longifolia* was conducted using a joint strategy including homology, RNA-seq, and methods based on *de novo* prediction. Employing hidden Markov models, the entire genome's coding genes were predicted using Augustus (version 2.7) [45]. Augustus was trained with default parameters using RNAseq and homologous prediction results as a training set. We compared the known homologous species' coding protein sequences with the novel species' genome sequence. Afterward, the new species' related gene region was determined via clustering algorithms like solar and GeneWise for the purpose of homology prediction (MAKER, version 2.2.1) [46]. The EST/cDNA sequence and the genome were compared. Using EVM (version r2012-06-25) [47] and MAKER, gene sets estimated through diverse approaches were combined into a non-redundant and more comprehensive gene set. The parameters used for EVM are: EVidenceModeler: --search_long_introns 25,000 -w weight.txt (weight.txt: stringtie2: 4, scallop2: 1, psiclass: 1, bloomT: 2, bloomP: 2, hom: 1, augustus: 6). By means of manual integration, the final dependable gene set was derived. Transcription set data assembled by Tophat (version 2.0.8b) alignment and Cufflinks (version 2.2.1) were also employed to supplement and complete the final gene set (including supplementing variable shear and UTR information).

BLAST (version 2.2.29+) [48] was adopted to annotate the predicted gene protein sequences based on the Gene Ontology (GO), RefSeq Non-Redundant Protein (NR), Clusters of Orthologous Groups of Proteins (COG), SwissProt, and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. An alignment array was used as a function of the target array. Because of the large number of possible matching results per sequence during the alignment process, matches were filtered using a threshold value of $\leq 1\times10^{-5}$ to ensure that further analyses were biologically meaningful. The 20 highest-scoring sequences were chosen from each sequence's comparison results and used as alignment results.

Non-coding RNA annotation

tRNAscan-SE software (version 1.3.1) [49] was utilised to search for genomic transfer RNA (tRNA) sequences according to tRNA structural features. The reference sequence was selected from the ribosomal RNA (rRNA) sequences of closely related species using BLASTN (version 2.6.0+) [50] alignment because rRNA is highly conserved. Using the Rfam 11.0 covariance model, we predicted the sequence information for small nuclear RNA (snRNA) and microRNA (miRNA).

Species evolution and phylogenetic analysis

In total, 14 species were employed to construct a phylogenetic tree: *Amborella trichopoda*, *Rhododendron simsii*, *Camellia sinensis*, *Vitis vinifera*, *Populus trichocarpa*, *Theobroma cacao*, *Juglans regia*, *Gossypium hirsutum*, *Actinidia chinensis*, *Z. mays*, *O. sativa*, *A. thaliana*, *M. pasquieri* (unpublished), and *M. longifolia* (Table S1). Divergence time was estimated according to the following methods: orthogroups were detected via OrthoMCL [51] with the DIAMOND [52] aligner. For each single-copy orthogroup, protein sequence alignment was conducted via MUSCLE (http://www.drive5.com/muscle/) [53], with all alignments combined into a supergene. This was used to establish a maximum-likelihood phylogenetic tree through RAxML under the GTR+F+R4 model containing 1000 bootstrap replicates, and maximum-likelihood evolutionary trees were constructed via IQ-tree [54]. Subsequently, the mcmctree functionality [55] in the PAML package (version 4.9) [56] was applied to predict the inter-species differentiation time, referring to other species' known differentiation times in the TimeTree database (http://www.timetree.org).

Based on evolutionary trees with varying times and gene family clustering results, we utilised birth rate and mortality models to estimate the number of ancestral gene family members per branch using CAFÉ software (version 4.0) [57]. We predicted expanded and contracted gene families in *M. longifolia*, as compared to their ancestral state. *P*-values <0.05 across the family were considered to determine significance. Expanded/contracted gene families in *M. longifolia* underwent GO and KEGG enrichment analyses. Genome-wide replication event analysis was performed for *M. longifolia*, *M. pasquieri*, *R. simsii*, *C. sinensis*, and *V. vinifera* genomes using the synonymous mutation rate (Ks) method.

Low-temperature experiment, RNA-seq, and transcriptome sequencing

Young leaves of *M. longifolia* are susceptible to chilling injury in winter during sudden temperature drops, which may cause wilting of the leaves and hinder seedling growth. We collected naturally dropped *M. longifolia* seeds from the same plant that provided the leaves

for DNA isolation. Seeds were germinated on wet paper towels; seedlings were then planted in peat soil for cultivation after the embryonic roots grew, using the same cultivation site noted in Sect. 2.1. A batch of uniformly grown specimens (age: 9 months) was selected from among these seedlings and transferred to an intelligent artificial climate chamber (5 °C, 65% humidity, 12 h photoperiod, 17,600 lx), and young leaves were collected after 0 (control check, CK), 1 (D1), 3 (D3), 5 (D5), and 7 (D7) days of exposure for RNA extraction (Figure S1). Three biological replicates per group were used. A TRIzol kit (Invitrogen, Carlsbad, CA, USA) was utilised to isolate total RNA. The Agilent 2100 Bioanalyzer was employed to determine the RNA quality, which was then confirmed via RNAse-free agarose gel electrophoresis. Subsequently, oligo (dT) beads were used to enrich mRNA. Short fragments were generated from the enriched mRNA using a fragmentation buffer. Reverse transcription was carried out utilising the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB #7530; New England Biolabs) to obtained cDNA. An Illumina Novaseq6000 by Gene Denovo Biotechnology Co. (Guangzhou, China) was utilised to sequence the obtained cDNA library.

Comparative transcriptome analysis of leaves after varying low-temperature exposure durations

Reads were fast filtered using fastp (version 0.18.0) to acquire extremely high-quality reads [30]. Reads from raw RNA were filtered and truncated to produce desired reads. Fastp (v 0.18.0) has parameters -a AGATCGGAAG AGC -q 20 -u 50 -n 15 -l 50. The purpose of this step is: (1) Remove reads containing adapter; (2) Remove reads with N ratio greater than 10%; (3) Remove reads with all A bases; (4) Remove low-quality reads (the number of bases with quality value Q≤20 accounts for more than 50% of the whole read). After establishing the reference indicator of the *M. longifolia* genome, the clean paired-end reads were aligned to this reference genome using HISAT2 (version 2.4) [58]. The parameter is set to "-rna-strandness RF" and the rest of the parameters are default. For each sample, read mapping and assembly was conducted via StringTie v1.3.1 following published protocols [59, 60]. To quantify the expression and variation of each transcription area, reads were normalised via fragments per kilobase of transcript per million mapped reads calculation using RSEM software [61]. The input data for the gene differential expression analysis were the read counts data obtained from the gene expression level analysis, which were analysed using the edgeR [62] software. The analysis was divided into three parts: (1) normalisation of the read counts; (2) calculation of the probability of hypothesis *P*-value according to the model.

DESeq2 [63] software was adopted to perform inter-group analysis of differentially expressed RNA. All DEGs were aligned with GO terms and KEGG pathways, and each term's gene count was computed. Hypergeometric testing was performed to detect GO terms remarkably enriched in the DEG in comparison with the genomic context, and KOBAS software [64] was employed to quantify significant KEGG pathway enrichment of the DEGs. *P*-values were quantified using hypothesis testing and corrected by FDR. Pathways with a *Q*-value≤0.05 were defined as significantly enriched among DEGs.

Identification, alignment, and phylogenetic analysis of *WRKY* gene family

Using blastp (parameter settings: e-value 10^{-5} and identity 50%), the protein sequences of *M. longifolia* were contrasted with the *WRKY* protein family of *A. thaliana* to identify genomic *WRKY* gene family members. Using the hmmsearch programme in hmmer 3.3.1 [65] under default parameters, the corresponding gene family members among *M. longifolia* protein sequences were determined according to the *WRKY* gene family structural domains. The members obtained from these two steps are merged as the result of the final *WRKY* gene family, where if the IDs of the genes obtained from the two steps are the same, they are retained only once. Using genome sequencing and chromosomal information of the *WRKY* genes of *M. longifolia*, chromosomal localisation analysis was performed utilising MG2C (http://mg2c.iask.in/mg2c_v2.0/) to precisely localise each *WRKY* member and facilitate gene homology analysis over evolutionary history. Based on the *A. thaliana* and *M. longifolia* *WRKY* gene family members, neighbour-joining tree construction with MUSCLE multiple sequence alignment (default parameters) was performed using MEGA11.0.8 [66] (parameters: Poisson model, partial deletion 80%, and 1000 bootstraps). Conserved motif analysis was carried out utilising the Multiple Em for Motif Elicitation (MEME) suite (version 5.3.0) [67] (parameter settings: repetition count: any; maximum motif count: 20; and optimal width per motif: 6–100 residues), and motif functional analysis was conducted as per the NCBI Conserved Domains Database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

Validation of key gene expression patterns via real-time fluorescence reverse transcription qPCR (RT-qPCR)

RT-qPCR was performed to verify RNA-seq findings. Fifteen genes were chosen for expression analysis at 1D, 3D, 5D and 7D using mikado Chr 09G46 as the internal reference gene, which encoding ubiquitin-coupled enzyme. Each set of experiments consisted of 3 biological replicates. Primer Premier 5.0 (Premier Biosoft International, Palo Alto, CA, USA) was utilised to design primers (Table S2). Total RNA was isolated from each sample using TRIzol reagent [68], and reverse transcription was

performed on a T100 Thermal Cycler (Bio-Rad, Hercules, CA, USA) to acquire template cDNA. qPCR was carried out on a TianLong 988 Real-Time PCR System (Tianlong Technologies, Xi'an, China) with ChamQ SYBR qPCR Master Mix (Vazyme, Nanjing, China) using the following programme: 90 s denaturation at 95 °C, and 40 amplification cycles were conducted of 95 °C for 5 s, 60 °C for 15 s, and 72 °C for 20 s. Each sample underwent three rounds of testing. Relative expression was computed using the $2^{-\Delta\Delta Ct}$ method [69].

Results

Madhuca longifolia genome size prediction and assembly

Altogether, 31.23 Gb (42.27×) PacBio long reads and 26.79 Gb (36.26×) Illumina short paired-end reads were obtained (Table S3). Using the valid *M. longifolia* genome data obtained from Illumina sequencing, a total of 22,037,014,232 k-mers with a main peak k-mer depth of 31.82 were identified (Table S4). A clear heterozygosity peak appeared approximately halfway to the major k-mer distribution curve peak (Figure S2), indicating a high heterozygosity rate. Through further calculation and correction, the genome size was calculated as 646.92 Mb, with a 1.48% heterozygosity ratio and a 38.75% repeat sequence ratio.

Based on preliminary genome assembly and correction, the genome of *M. longifolia* contained 340 contigs, and the genome scale was 739.00 Mb with the GC content of 33.76%. Contig N50 was 56.71 Mb long, with the longest assembled contig size being 76.57 Mb (Table 1). Chromosome counts showed that *M. longifolia* was diploid (2n = 24, Figure S3). Contigs obtained from preliminary Hi-C assembly were clustered into 12 pseudochromosomes (Fig. 1A, Table S5). The Hi-C interaction heat map showed a well-organised diagonal pattern of intrachromosomal interactions, indicating a satisfactory genome assembly (Fig. 1B). The ultimate assembled genome dimension was 737.92 Mb in size, and contig N50 and scaffold N50 had lengths of 56.71 Mb and 60.05 Mb, respectively. The longest assembled contig and scaffold sizes were 76.63 Mb and 84.26 Mb, respectively, with a GC content of 33.74% (Table 1).

Illumina data were re-matched to the initially assembled *M. longifolia* genome to assess assembly integrity and accuracy, and the Illumina reads showed 98.62% alignment and 99.81% coverage (Tables S6 and S7). CEGMA assessment showed a detection rate of 234/248 (94.35%) for core eukaryotic genes (Table S8). The BUSCO evaluation results indicated that 1,571 of 1,614 direct homologous single-copy genes (97.34%) were observed in the *M. longifolia* genome (Table S9). The final assembly outcomes were also evaluated using the BUSCO software and yielded 1,535 single-copy genes, representing 95.11% of the gene counts (Table S10). This indicated that the final assembled genome of *M. longifolia* had good integrity and high quality.

Table 1 Statistics of *Madhuca longifolia* genome assembly

PacBio assembly		
Statistics	Contig length (bp)	Number
Max	76,565,288	1
N10	76,565,288	1
N20	72,498,706	2
N30	60,044,954	4
N40	57,664,288	5
N50	56,707,685	6
N60	51,577,671	8
N70	51,515,675	9
N80	41,195,910	11
N90	30,563,916	13
Total length	738,965,282	340
GC rate (%)	33.7631	-
Hi-C assembly		
Statistics	Scaffold	Contig
Total number	329	331
Total length (bp)	737,923,470	737,921,470
Gap (N) (bp)	2,000	0
Average length (bp)	2,242,928.48	2,229,370
N50 length (bp)	60,054,753	56,707,731
N90 length (bp)	41,669,588	30,985,487
Maximum length (bp)	84,256,480	76,625,452
Minimum length (bp)	13,157	13,157
GC content (%)	33.74	33.74

Genome annotation

A total of 1,294 rRNA, 2,043 tRNA, 264 miRNA, and 287 snRNA sequences were obtained by annotating the non-coding RNAs of the *M. longifolia* genome. The average length for these four RNA types was 2,581.70 bp, 74.54 bp, 133.95 bp, and 123.97 bp, respectively (Table S11). The genome contained 486.30 Mb of repeated sequences, occupying 65.90% of the total genome. Among the detected transposable elements, which are important components of repeated sequences, the LTR transposon class was predominant at 69.16 Mb in size and occupying 9.36% of the genome. The next most common class was the long interspersed nuclear element class, with a dimension of 18.03 Mb, occupying 2.44% of the genome. Among the DNA transposons, the highest proportion of genomes was in the Helitron class (62.31%), followed by miniature inverted-repeat transposable elements (1.63%) (Table S12). Structural annotation of genomic coding genes yielded 46,610 coding genes. The mean gene length of the *M. longifolia* genome was 5,561.26 bp, with the N50 gene being 9,993 bp in length. The average mRNA length was 1,625.14 bp. The mean coding sequence was 1,154.17 bp long. The mean exon length was 272.37 bp, and the mean count of exons in each gene was 5.78 (Table

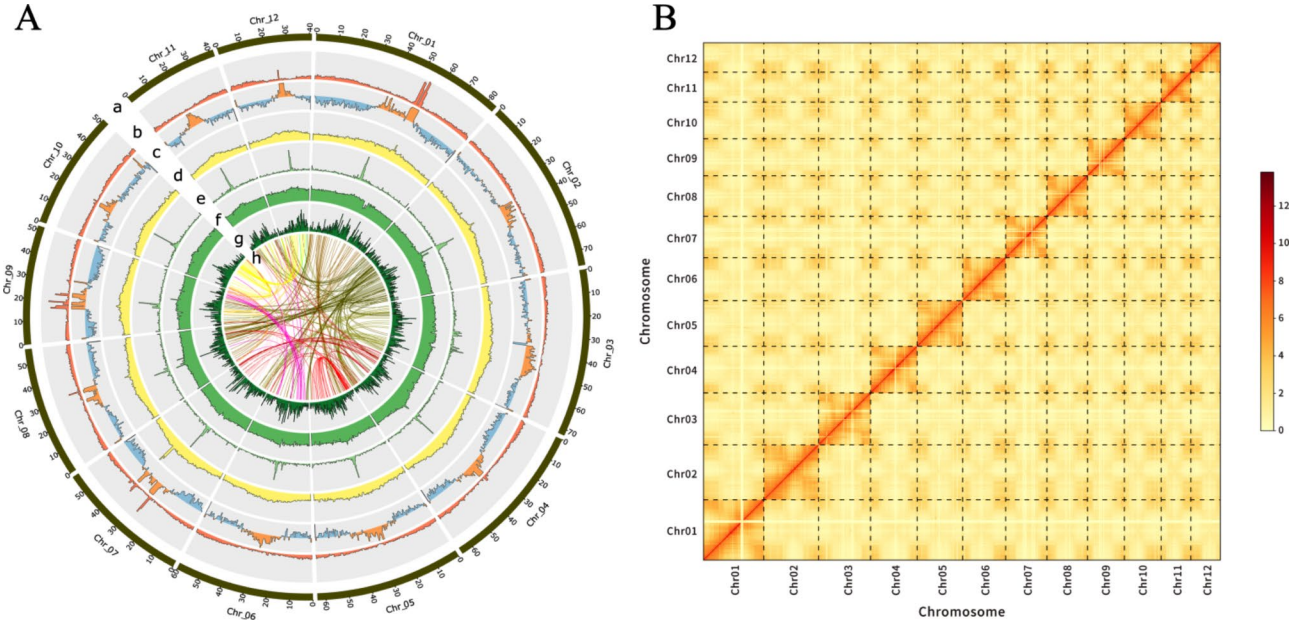

Fig. 1 Chromosomal features and Hi-C map of the *Madhuca longifolia* genome (**A**) Landscape of the *M. longifolia* genome. Inwards from the outside: Chromosome (a), Gene density (b), GC distribution density (c), Transposon distribution density (d), LINE distribution density (e), Illumina short read distribution density (f), LTR copy distribution density (g), Schematic of major inter-chromosomal relationships in the *M. longifolia* genome (h). (**B**) Heat map of the Hi-C interaction density between 12 pseudochromosomes in *M. longifolia*
LTR, long terminal repeat; LINE, long interspersed nuclear element

S13 and S14). BUSCO assessment (Table S15) showed 96.47% completeness, indicating excellent annotation of the encoded genes. Among the genes with coded proteins, 34,927 were annotated against the NR database (75% coverage), 24,100 for SwissProt (52%), 26,665 for GO (57%), 20,243 for COG (43%), and 34,162 for KEGG (73%; Table S16).

Evolution of gene families

The *M. longifolia* genome contained 15,545 gene families consisting of 26,724 genes. Table S17 shows the clustered gene families in the remaining 13 analysed species. A total of 10,517 genes were identified and shared by *M. longifolia*, *M. pasquieri*, *R. simsii*, *V. vinifera*, and *A. chinensis* (Fig. 2A). Forty-one single-copy gene families were shared among all species. Homologous single-copy gene sequence comparison and maximum-likelihood tree construction showed that *M. longifolia* and *M. pasquieri* diverged at 15–75 mya, slightly later than *O. sativa* and *Z. mays* (Fig. 2B). The gene family expansion/contraction analysis for *M. longifolia* revealed that 632 gene families underwent expansion and 161 experienced contraction (Fig. 2B). Among the expanded gene families, 2,669 genes were annotated with GO terms, with 1,540 genes enriched in biological process terms, 313 in cellular component terms, and 815 in molecular function terms (Figure S4). Altogether, 749 genes were enriched in 91 KEGG pathways (Figure S5). Genes that underwent expansion were significantly enriched in metabolic, glutathione metabolism, RNA polymerase, and oxidative phosphorylation pathways.

Whole-genome duplication

The results of the covariance analysis showed a 2:1 syntenic depth rate between *M. longifolia* and *V. vinifera* and a 2:2 syntenic depth rate between *M. pasquieri* and *R. simsii* (Fig. 2C). The *V. vinifera* genome did not exhibit the whole-genome doubling (WGD) followed by whole-genome tripling common to core dicotyledons [70], whereas this was the case in *R. simsii* [71]. Ks distribution mapping for *M. longifolia*, *M. pasquieri*, *V. vinifera*, *C. sinensis*, and *R. simsii* highlighted a shared peak at ~1.5 Ks representing a genome-wide triploidisation event (γ event) common to core dicots, after which *M. longifolia* experienced another, more recent WGD event (Fig. 2D).

Comparative transcriptome analysis of young leaves under different low-temperature treatment durations

Across all pairwise comparisons of seedlings from the CK, 1D, 3D, 5D, and 7D groups, 17,941 DEGs were recognised, with 3,382 overlapping DEGs (Fig. 3A). The D7 group showed the highest number of DEGs (14,291) against the CK group, of which 4,765 were upregulated and 9,526 were downregulated. The fewest DEGs (5,945) were found between CK and D1, of which 3,237 were upregulated and 2,708 were downregulated in the D1 group. Altogether, 13,116 DEGs were detected between the CK and D3 groups, where 4,512 were upregulated

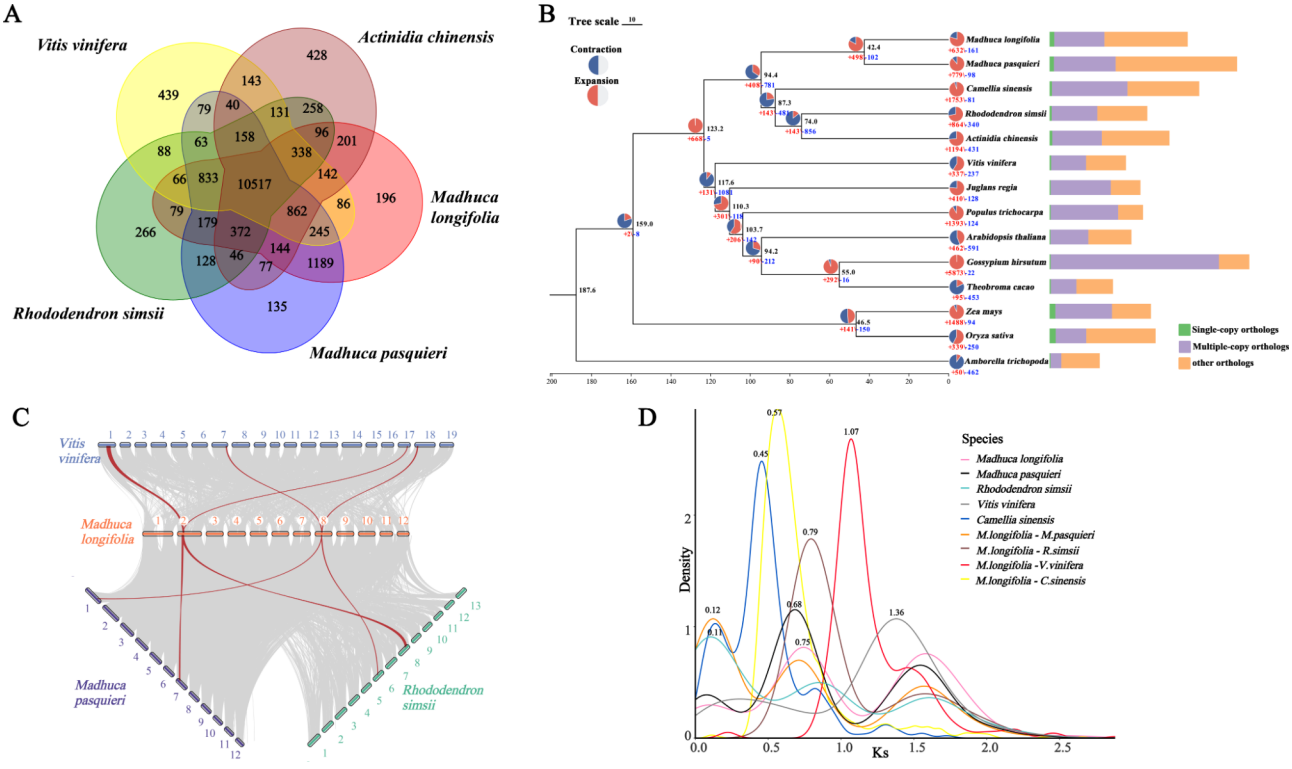

Fig. 2 Comparative genomic analysis of *Madhuca longifolia* with other plants (**A**) Venn diagram of unique and common gene families in *M. longifolia*, *M. pasquieri*, *Rhododendron simsii*, *Vitis vinifera*, and *Actinidia chinensis*. (**B**) Phylogenetic analysis of the *M. longifolia* genome based on phylogenetic relationships with 14 species. Node labels denote node ages. Gene family expansion or contraction is presented in the pie chart. Numbers of gene family cluster classes in each species are presented in the histogram. (**C**) Phylogenetic relationships of *M. longifolia* and other species (*V. vinifera*, *M. pasquieri*, and *R. simsii*). (**D**) Distribution of synonymous substitution rates (Ks) for homologous genomes used for intrachromosomal comparisons. The Ks value peaks (Ks = 0.7 and 1.5) indicate the occurrence of a recent WGD and an ancient WGD, respectively, in the *M. longifolia* genome
WGD, whole-genome doubling

and 8,604 were downregulated in the D3 group; 13,988 DEGs were found between the CK and D5 groups, of which 4,629 were upregulated and 9,359 were downregulated in the D5 group (Fig. 3B). Based on GO annotation with hypergeometric tests, DEGs were predominantly enriched in the organonitrogen compound metabolic process (GO:1901564), ion binding (GO:0043167), and catalytic complex (GO:1902494) GO terms (Figure S6). KEGG analysis showed that DEGs between CK and D1, CK and D3, CK and D5, and CK and D7 were enriched in 129, 135, 138, and 137 pathways, respectively (Figure S7). Further trend analysis showed that DEGs were clustered into 20 profiles. Among these, 15,993 DEGs were clustered into five profiles at $P<0.05$ (downregulation mode, profile 0; upregulation mode, profile 19; upregulation then stable then downregulation, profile 18; upregulation then downregulation then stable, profile 12; and stable then downregulation then upregulation, profile 8) (Fig. 3C). Profile 0 contained 9,391 DEGs that were downregulated in the 1D, 3D, 5D, and 7D groups compared to levels in the CK group. Profile 19 contained 2,941 DEGs that were upregulated in these groups.

Identification of the *WRKY* gene family

Altogether, 94 *WRKY* putative genes were identified in the genome of *M. longifolia*. Chromosomal localisation analysis revealed that these genes were distributed on 12 chromosomes, with each chromosome containing 14, 15, 4, 6, 12, 7, 6, 8, 5, 6, 7, and 4 *MlWRKY*s, respectively (Fig. 4A and Table S18). Phylogenetic analyses of *WRKY* gene families for *M. longifolia* and *A. thaliana* revealed three principal families (groups I–III), among which group II was subdivided into five subfamilies: II-a to II-e (Fig. 4B). *WRKY* gene family groups I, II, and III respectively contained 15, 58, and 21 genes, and group II subfamilies IIa, IIb, IIc, IId, and IIe respectively contained 4, 11, 25, 8, and 10 genes (Table S19). Utilising the MEME Suite, 20 conserved amino acid sequences corresponding to members of the *M. longifolia*'s *WRKY* gene family were found. Four conserved sequences with characteristic *WRKY* structural domains were identified (Figure S8). Genes containing four *WRKY* structural domain sequences (motifs 1, 2, 3, and 5) were all located in group I. Of 94 *MlWRKY*s, only 14 (*MlWRKY*1, *MlWRKY*2, *MlWRKY*3, *MlWRKY*6, *MlWRKY*14, *MlWRKY*20, *MlWRKY*24, *MlWRKY*25, *MlWRKY*31, *MlWRKY*37,

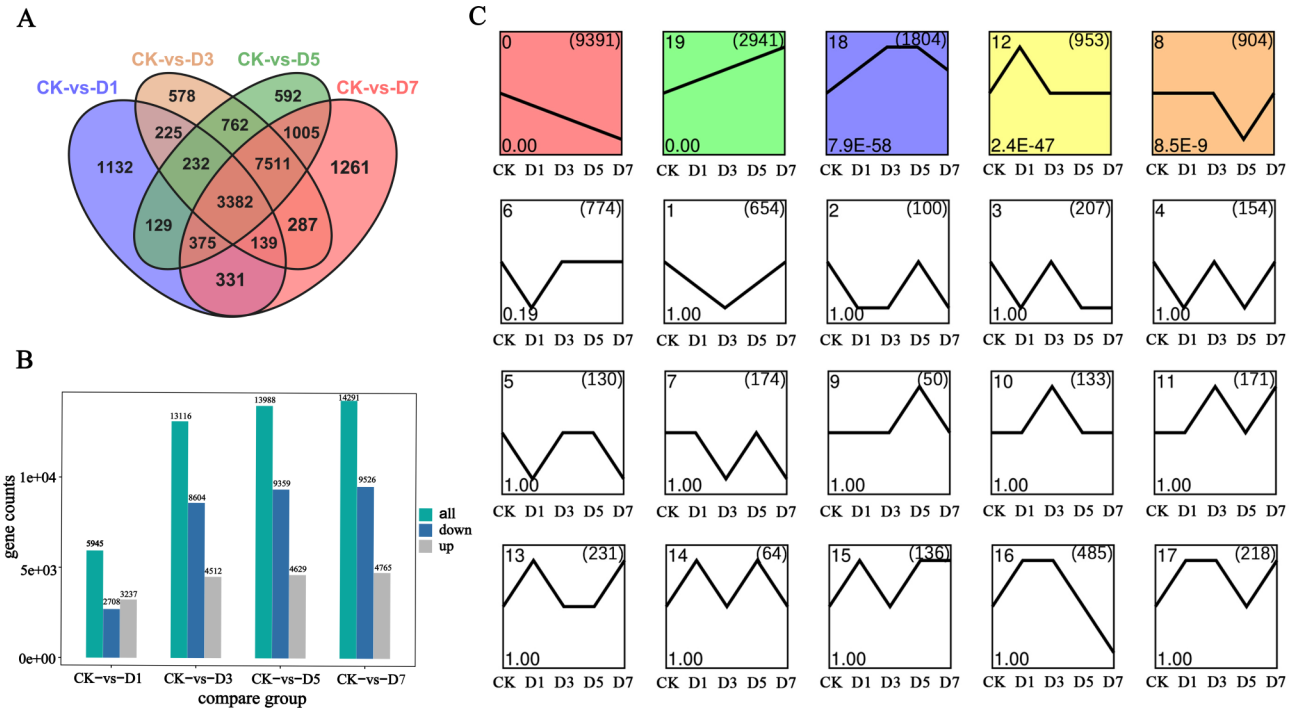

Fig. 3 Low-temperature transcriptome analysis of *Madhuca longifolia* (**A**) Venn diagram of the number of DEGs under different durations of low-temperature stress: 0 days versus 1 day (CK-vs-D1), 0 days versus 3 days (CK-vs-D3), 0 days versus 5 days (CK-vs-D5), and 0 days versus 7 days (CK-vs-D7). (**B**) Number of up- and downregulated DEGs in the four comparisons shown in A. (**C**) Total trend of all expression changes in DEGs under low-temperature stress in *M. longifolia*. For each profile, the number in the lower left corner denotes the *P*-value, the number in the upper left corner symbolises the profile ID, the number in parentheses in the upper right corner indicates the number of genes assigned to that profile, and the coloured profile represents instances of $P < 0.05$

DEG, differentially expressed gene; CK, blank control group with 0 days of low-temperature treatment; D1, D3, D5 and D7, control groups with 1, 3, 5, 7 days of low-temperature treatment

*MlWRKY*47, *MlWRKY*84, *MlWRKY*88, and *MlWRKY*89) contained either motif 1 or motif 2 but not both, whereas each of the other genes contained both motifs, which were closely related (Figure S9).

The transcriptome trend analysis revealed that of the 94 *WRKY* genes identified in the *M. longifolia* genome with differential expression under low-temperature stress, 17 upregulated *MlWRKY*s corresponded to profile 19 and 15 downregulated *MlWRKY*s represented profile 0. Based on this result, a gene expression heat map was plotted via the TBtools software [72] (Figure S10). The top eight differentially expressed *MlWRKY*s in profile 19 and the equivalent top seven in profile 0 were identified as genes likely to be tightly linked with the response to low-temperature stress in *M. longifolia* (Table S20). These 15 differentially expressed *MlWRKY*s (*MLWRKY81*, *MLWRKY21*, *MLWRKY52*, *MLWRKY74*, *MLWRKY99*, *MLWRKY46*, *MLWRKY54*, *MLWRKY5*, *MLWRKY59*, *MLWRKY90*, *MLWRKY13*, *MLWRKY48*, *MLWRKY28*, *MLWRKY97*, and *MLWRKY76*) were selected and validated via qRT-PCR, and their expression in the transcriptome was largely consistent with the fluorescence quantification of expression in different samples (Fig. 4C), further supporting the credibility of the RNA-seq data.

Discussion

In this study, Illumina sequencing, PacBio HiFi sequencing, and Hi-C technology were incorporated to sequence and assemble the complete *M. longifolia* genome to obtain a high-quality chromosome-level reference genome. This is the first complete chromosome-level reference genome for the genus *Madhuca* and provides considerable genomic data for investigations of other species in the genus. The *M. longifolia* genome also provides a basis for future research on molecular breeding, phylogeny, and resistance mechanisms. The size of the assembled genome is approximately 737.92 Mb, with contig N50 (56.71 Mb) and scaffold N50 (60.05 Mb) both notably larger than contig N50 (2.2 Mb) and scaffold N50 (36 Mb) of the closely related species *R. simsii* [71]. In this study, 65.90% of the total genome of *M. longifolia* was represented by duplicated sequences, a considerably larger percentage than in *R. simsii* (47.48%) and *R. delavayi* (51.77%) [71]. This indicates that *M. longifolia* could have undergone greater sequence differentiation and genome expansion than these species. Altogether, 46,610

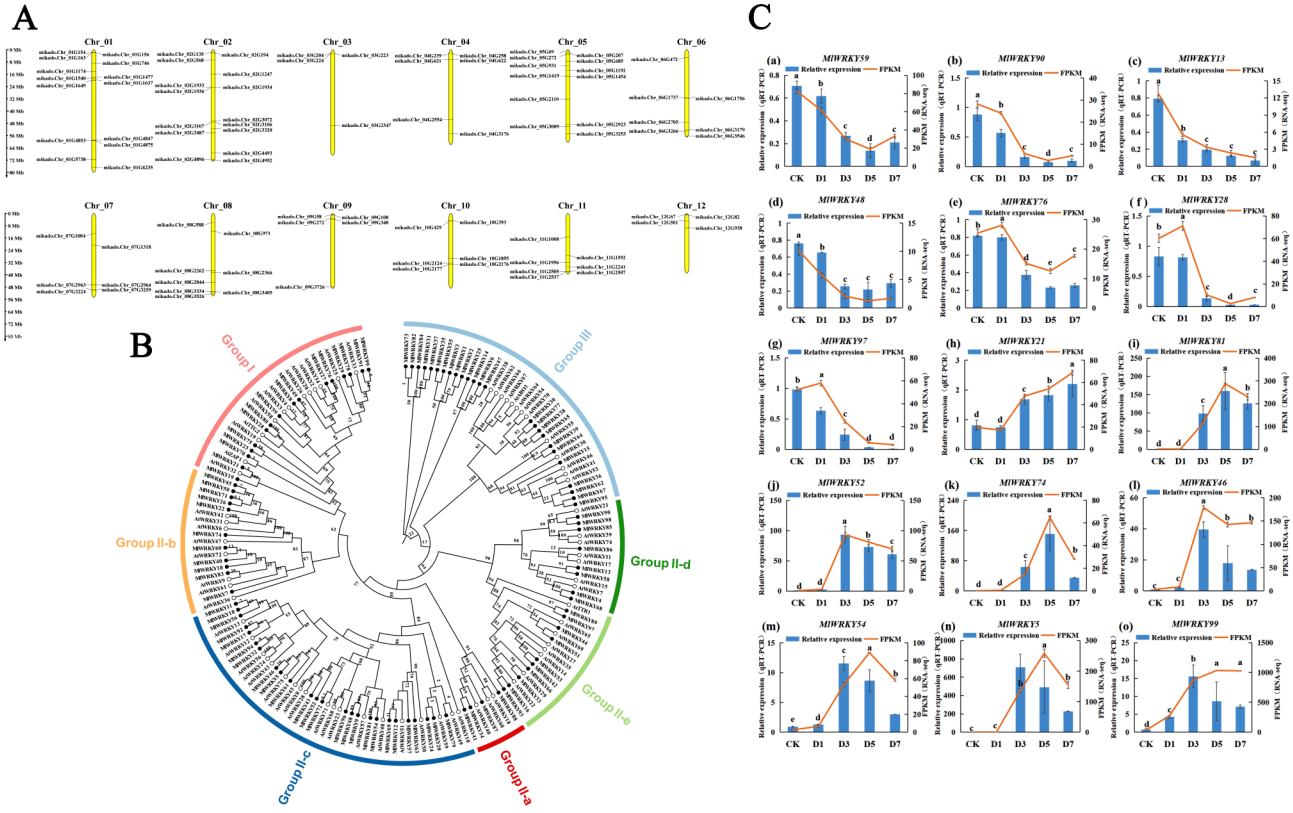

Fig. 4 Analysis of the *WRKY* gene family and RT-qPCR validation of key *WRKY* genes for low-temperature response in *Madhuca longifolia* (**A**) Chromosomal localisation of 94 *WRKY* genes on *M. longifolia* chromosomes. (**B**) Phylogeny of the *WRKY* gene family in *M. longifolia* and *Arabidopsis thaliana*. (**C**) RT-qPCR validation of 15 *MIWRKY* genes in response to low-temperature stress in *M. longifolia*. In (c), panels (a)–(g) show *WRKY* genes consistently downregulated in profile 0. Panels (h)–(o) show *WRKY* genes consistently upregulated in profile 19

coding genes corresponded to known functional annotations for *M. longifolia*, close to the count for the closely related species *A. chinensis* (40,464) [73] but markedly higher than that for *R. simsii* (32,999).

Adaptive evolution is a key strategy for the survival of all species. Understanding the molecular mechanisms controlling adaptive evolution helps us to comprehend the development of adaptive characteristics and the correlations that support species diversification, phenotype convergence, and interspecific interactions. It may also provide valuable knowledge of the formation and sustainability of biodiversity [74]. Gene family expansion significantly affects species differentiation. We found that 632 gene families in *M. longifolia* have expanded during its evolutionary history and exhibit enrichment in oxidative phosphorylation, glutathione metabolism, and RNA polymerase pathways. These gene families may have been associated with resistance to environmental stress during the species' evolutionary history. The natural habitat of *M. longifolia* is southern India and Myanmar, where it is warm throughout the year and hot in the summer. It can therefore be assumed that *M. longifolia* has gradually evolved to become more tolerant to hot environments and less adaptable to low temperatures. This may be one of the reasons for the poor overwintering adaptability of the species during cold waves in winter after its introduction to southern China.

The divergence time of *A. chinensis* and *R. simsii* in the present study was approximately 74.0 Ma, which is almost identical to the previously reported divergence time of 74.78 Ma. In contrast, the differentiation time of 87.3 Ma for *A. chinensis* and *C. sinensis* was greater than that of 60.95–76.84 Ma reported in other studies [73]. This may be due to the use of genomes from more species in this investigation, particularly the inclusion of genomic data from the Sapotaceae family, which is closely related to the Actinidiaceae and Theaceae families. Genome-wide replication events are important drivers of species evolution and can lead to changes in plant genome size and gene number [75, 76]. *M. longifolia* exhibited covariate relationships of 2:1, 2:2, and 2:2 with *V. vinifera*, *M. pasquieri*, and *R. simsii*, respectively. Both *M. longifolia* and *M. pasquieri* have 12 pseudochromosomes, whereas the closely related species *Synsepalum dulcificum* [77] and *R. simsii* have 13 pseudochromosomes. Therefore, the evolution of the chromosomes of *M. longifolia* is of great importance to the Sapotaceae and Ericales. *M. longifolia* and *M. pasquieri* peaked at Ks=0.75 and

Ks=0.68, respectively, while *S. dulcificum* of the same family peaked at Ks=0.56. Given *M. longifolia*'s evolutionary position in the phylogenetic tree, we speculate that this WGD event was not unique to the Sapotaceae family. It has been reported that *C. sinensis* underwent only one WGD event following a whole-genome tripling event, and that this was the same WGD event shared by *A. chinensis*, *C. sinensis*, *R. simsii*, and *Diospyros kaki* [78], which may also be shared with other species in the family. However, because of the lack of genomic studies on *M. longifolia*, the neutral mutation rate of the tightly related species *A. chinensis* was utilised to calculate the timing of the WGD event in this analysis. This may have led to less accurate results and could be corrected in conjunction with more relevant subsequent research results.

Low temperatures are a key factor limiting the large-scale cultivation of high-quality tropical trees in China. Low-temperature stress often results in the severe dehydration of plant cells, leading to tissue injury, stunted growth, and wilting. Various physiological, molecular, and metabolic responses driven by multiple pathways occur when plants resist the adverse effects of low temperatures [79, 80]. We found that the total number of DEGs between the low-temperature groups and the CK group gradually rose with an increasing duration of seedling exposure to stress. More genes were downregulated than upregulated on days 3, 5, and 7 of low-temperature treatment but not on day 1, when the count of upregulated genes surpassed that of downregulated genes. Enrichment analyses of several groups of DEGs using the GO and KEGG databases demonstrated that DEGs were remarkably enriched in membrane pathways (GO:0016020) on D1 as compared to CK. One of the key mechanisms for adapting to low-temperature stress involves modifying the plasma membrane's function and composition [81]. As the low-temperature treatment duration escalated, the DEGs in plant tissues became remarkably enriched in pathways related to stress response. Examples of other pathways associated with membranes included the integral component (GO:0005887) and the obsolete intrinsic component of the plasma membrane (GO:0031226). The enrichment analysis results in this study are highly similar to previously reported enrichment pathways of DEGs in *Kandelia obovata* during cold acclimation in coastal environments [82].

When exposed to abiotic stress factors, some *WRKY* TFs quickly promote signal transduction and lead to differential gene expression [83]. The *WRKY* expression modes and functional identification are identified through transcriptome analysis and RT-qPCR. The ongoing expansion of plant genome and transcriptome databases has led to the detection of a rising number of *WRKY* genes, e.g., 82 in *Solanum tuberosum* [84] and 95 in *Daucus carota* [85]. A previous study predicted 96 *WRKY* TFs in the genome of *M. pasquieri* [86]; herein, 94 *MlWRKYs* were recognised for the first time in the *M. longifolia* genome. The *WRKY* gene count in both species was very similar, which may be related to their close affinity. *WRKY* genes in *M. longifolia* were nonuniformly distributed across the species' 12 chromosomes. Phylogenetic analysis allowed the classification of these genes into three groups, of which group II was subdivided into five subgroups (II-a to II-e). This clustering result is accordant with that found in a prior study [87]. Subgroup II-c had the largest number of members (25), whereas II-a had only four *MlWRKY* members; this distribution is similar to the clustering results of *WRKY* subgroups in *S. lycopersicum* [88] and *Manihot esculenta* [89].

Various investigations have indicated that *WRKY* genes influence many plants' responses to low-temperature stress. The *WRKY71* protein is localised in the nucleus of *Fragaria×ananassa* seedlings and plays a role in responses to abiotic stressors such as cold, salt, and low phosphate levels [90]. *BcWRKY46* in *Brassica campestris* is triggered by low-temperature stress and ABA to improve plant resistance through the activation of relevant genes in ABA signalling pathways [91]. In the present study, 15 key *MlWRKYs* were found to respond to low temperatures, all of which were enriched in GO terms related to the control of transcription and transcription with a DNA template (GO:0006355), DNA binding specific to sequences (GO:0043565), and transcription factor activity binding to DNA (GO:0003700) (Table S20). These genes may be key regulators of related pathways that rapidly respond to low-temperature stress, regulating the expression of related genes and altering metabolite synthesis and secretion, among other responses.

Despite our valuable findings, the study has some limitations. First, the transcriptome samples we used were leaf sections that were not further disassembled for sequencing to analyse the differential expression among different tissues; therefore, it remains unclear whether there are differences in the expression modes of *MlWRKY*s in distinct tissues under low-temperature stress. Further, it is unknown how these *WRKY* gene family members respond to low-temperature signals and regulate gene expression. Further experiments are needed to validate the functions of these genes and investigate their specific roles in the complicated molecular mechanisms related to the low-temperature response of *M. longifolia*. Moreover, there is a lack of metabolomic data related to *M. longifolia* under low-temperature stress, limiting our analysis of differential metabolites and core metabolic pathways. In the future, further combinations of genomic, transcriptomic, metabolomic, proteomic and other multi-omics data can be used to establish a gene–metabolite regulatory network, mine additional

hub genes, and clarify the related regulatory relationships. Finally, although RT-qPCR verified recognised hub gene expression, insufficient direct molecular experiments were performed to functionally verify them. As a next step, a genetic transformation system of *M. longifolia* needs to be established to facilitate the verification of gene functions. Combined with multi-omics association analysis and molecular biology techniques, we plan to explore the principal target genes and pathways related to low-temperature stress responses and clarify the existence of any interactions between these, so as to analyse the molecular mechanism of these responses in depth.

Conclusions

In this study, high-quality chromosome-level genome assembly was performed for *M. longifolia*, and key genes controlling low-temperature responses were identified for the first time based on genomic and transcriptomic data. The genomic data and comparative genomic analyses provide valuable references for further studies on the adaptive evolution of *M. longifolia* and related species. The derived transcriptome information constitutes a basis for further elucidating the adaptive mechanisms of *M. longifolia* to unfavourable low-temperature environmental conditions and for optimising the molecular breeding and cultivation of other high-quality tropical woody plants. Future studies can use our genome assembly, annotation, and transcriptome data to enhance the ecological adaptability and exploitability of valuable tropical trees to different environments following their introduction.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10769-2.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors appreciate Gene Denovo Biotechnology Co. (China)'s support on the project and particularly Zhiyuan Yang for her technical guidance and assistance.

Author contributions

ZL secured funding and managed the project, which reviewed and edited this article. WSY prepared the experimental materials and completed the resource survey, data collation and visualization, and was a major contributor in writing the manuscript. LHY was involved in the experiments and the data analysis. YSY assisted in the preparation of experimental materials and software analysis. JZL participated in the revision and editing of the experiment and writing. CZP and MYF assisted in the software analysis and data visualization process. All the authors read and approved the final manuscript.

Funding

This work was funded by the National Natural Science Foundation of China [No. 32371742], the Wildlife Conservation and Management Projects of Guangdong Forestry Administration (2022 and 2023), and the Forestry Department of Guangdong Province, China, for non-commercial ecological forest research [No. 2020STGYL0019].

Data availability

The whole-genome sequence data used in this paper have been stored in the Genome Warehouse in the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number GWHDTZT00000000. The data are accessible to the public at https://ngdc.cncb.ac.cn/gwh.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 February 2024 / Accepted: 4 September 2024
Published online: 18 September 2024

References

1. Chinnusamy V, Zhu J, Zhu JK. Cold stress regulation of gene expression in plants. Trends Plant Sci. 2007;12:444–51.
2. Akshatha KN, Mahadeva Murthy S, Lakshmidevi N. Ethnomedical uses of *Madhuca longifolia*-A review. Int J Life Sci Pharm Res. 2013;3:44–53.
3. Sharma M, Yadav S, Ganesh N, Srivastava MM, Srivastava S. Biofabrication and characterization of flavonoid-loaded ag, au, Au-Ag bimetallic nanoparticles using seed extract of the plant *Madhuca longifolia* for the enhancement in wound healing bio-efficacy. Prog Biomater. 2019;8:51–63.
4. Pinakin DJ, Kumar V, Kumar S, Kaur S, Prasad R, Sharma BR. Influence of pre-drying treatments on physico-chemical and phytochemical potential of dried mahua flowers. Plant Foods Hum Nutr. 2020;75:576–82.
5. Simon JP, Evan Prince S. Ameliorative activity of aqueous leaf extract from *Madhuca longifolia* against diclofenac-administered toxicity on rat stomach and intestine. J Histotechnol. 2021;44:114–26.
6. Pinakin DJ, Kumar V, Suri S, Sharma R, Kaushal M. Nutraceutical potential of tree flowers: a comprehensive review on biochemical profile, health benefits, and utilization. Food Res Int. 2020;127:108724.
7. Pradhan SK, Sahoo UC. Evaluation of recycled asphalt mixtures rejuvenated with *Madhuca longifolia* (Mahua) oil. Int J Pavement Res Technol. 2021;14:43–53.
8. Asanthi H, Yasasvi J, Ashoka G, et al. Nutritional, functional properties and applications of Mee (*Madhuca longifolia*) seed fat. Agronomy. 2023;13:2445.
9. Fatma A, Ahuja V, Ahuja A, et al. Evaluation of antibacterial activity of *Madhuca longifolia* (Mahua) stem extract against *Streptococcus mutans*: an in vitro study. Cureus. 2024;16(1):e52210.
10. Vinotha V, Vaseeharan B. Bio-fabricated zinc oxide and cry protein nanocomposites: synthesis, characterization, potentiality against Zika, malaria and West Nile virus vector's larvae and their impact on non-target organisms. Int J Biol Macromol. 2023;224:699–712.
11. Gopinath R, Billigraham P, Sathishkumar TP. Characterization studies on novel cellulosic fiber obtained from the bark of *Madhuca longifolia* tree. J Nat Fibers. 2022;19:14880–97.
12. Bandara WART, Dissanayake CTM. Most tolerant roadside tree species for urban settings in humid tropics based on Air Pollution Tolerance Index. Urban Clim. 2021;37:100848.
13. Hou Q, Li Y, Kang W, Zhou T, Liu J, Luo J et al. Selection of tree species for anti-typhoon shelter forests of tropical coastal city in Hainan. J Cent S Univ Technol. 2011;31:184 – 91, 240.
14. Ebrahimi A, Lawson SS, McKenna JR, Jacobs DF. Morpho-physiological and genomic evaluation of Juglans species reveals regional maladaptation to cold stress. Front Plant Sci. 2020;11:229.
15. Shang J, Tian J, Cheng H, Yan Q, Li L, Jamal A, et al. The chromosome-level wintersweet (*Chimonanthus praecox*) genome provides insights into floral scent biosynthesis and flowering in winter. Genome Biol. 2020;21:200.

16. Sork VL, Squire K, Gugger PF, Steele SE, Levy ED, Eckert AJ. Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*. Am J Bot. 2016;103:33–46.
17. Zhao T, Ma W, Yang Z, Liang L, Chen X, Wang G, et al. A chromosome-level reference genome of the hazelnut. Corylus heterophylla Fisch GigaScience. 2021;10:giab027.
18. Simon SA, Zhai J, Nandety RS, McCormick KP, Zeng J, Mejia D, et al. Short-read sequencing technologies for transcriptional analyses. Annu Rev Plant Biol. 2009;60:305–33.
19. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
20. He RY, Yang T, Zheng JJ, Pan ZY, Chen Y, Zhou Y, et al. QTL mapping and a transcriptome integrative analysis uncover the candidate genes that control the cold tolerance of maize introgression lines at the seedling stage. Int J Mol Sci. 2023;24:2629.
21. Liu Y, Xiong Y, Zhao J, Bai S, Li D, Chen L, et al. Molecular mechanism of cold tolerance of centipedegrass based on the transcriptome. Int J Mol Sci. 2023;24:1265.
22. Li H, Zhou T, Chong X, Lu X, Li Y, Zheng B, et al. Transcriptome and expression analysis of genes related to regulatory mechanisms in Holly (*Ilex dabieshanensis*) under cold stress. Forests. 2022;13:2150.
23. Song S, Ma D, Xu C, et al. In silico analysis of NAC gene family in the mangrove plant *Avicennia marina* provides clues for adaptation to intertidal habitats. Plant Mol Biol. 2023;111:393–413.
24. Zhou H, M J, Liu H, et al. Genome-wide identification of the CBF gene family and ICE transcription factors in walnuts and expression profiles under cold conditions. Int J Mol Sci. 2024;25:25.
25. Zhang Y, Wang L. The *WRKY* transcription factor superfamily: its origin in eukaryotes and expansion in plants. BMC Evol Biol. 2005;5:1.
26. Pandey SP, Somssich IE. The role of *WRKY* transcription factors in plant immunity. Plant Physiol. 2009;150:1648–55.
27. Kim C, Vo KTX, Nguyen CD, Jeong D, Lee S, Kumar M, et al. Functional analysis of a cold-responsive rice *WRKY* gene, *OsWRKY71*. Plant Biotechnol Rep. 2016;10:13–23.
28. Zou C, Jiang W, Yu D. Male gametophyte-specific *WRKY34* transcription factor mediates cold sensitivity of mature pollen in *Arabidopsis*. J Exp Bot. 2010;61:3901–14.
29. Luo DL, Ba LJ, Shan W, Kuang JF, Lu WJ, Chen JY. Involvement of *WRKY* transcription factors in abscisic-acid-induced cold tolerance of banana fruit. J Agric Food Chem. 2017;65:3627–35.
30. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
31. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.
32. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.
33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
34. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25.
35. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 2019;5:833–45.
36. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.
37. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9:999–1003.
38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
39. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009, Chap. 4;Chapter:4.10.1–4.10.14.
40. Edgar RC, Myers EW. PILER: identification and classification of genomic repeats. Bioinformatics. 2005;21(Suppl 1):i152–8.
41. Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12:1269–76.
42. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl 1):i351–8.
43. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265–8.
44. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
45. Stanke M, Tzvetkova A, Morgenstern B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol. 2006;7(Suppl 1):S111–8.
46. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164:513–24.
47. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9:R7.
48. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6.
49. Kollmar M, TRNAscan, -SE. Searching for tRNA genes in genomic sequences. London: Springer; 2019. pp. 1–14.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
51. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
52. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
53. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
54. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
55. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997;13:555–6.
56. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
57. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22:1269–71.
58. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
59. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
60. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11:1650–67.
61. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
62. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
63. Love MI, Huber W, Anders S. Moderated estimation of Fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
64. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39Suppl2:W316–22.
65. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14:755–63.
66. Kumar S, Tamura K, Nei M. MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput Appl Biosci. 1994;10:189–91.
67. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.
68. Rio DC, Ares MJ, Hannon GJ, Nilsen TW. Purification of RNA using trizol (TRI reagent). Cold Spring Harb Protoc. 2010;2010:pdb.prot5439.
69. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method. Methods. 2001;25:402–8.
70. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–7.
71. Yang FS, Nie S, Liu H, Shi TL, Tian XC, Zhou SS, et al. Chromosome-level genome assembly of a parent species of widely cultivated azaleas. Nat Commun. 2020;11:5269.

72. Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, et al. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant. 2020;13:1194–202.
73. Wu H, Ma T, Kang M, Ai F, Zhang J, Dong G, et al. A high-quality *Actinidia chinensis* (kiwifruit) genome. Hortic Res. 2019;6:117.
74. Hu Y, Wang X, Xu Y, Yang H, Tong Z, Tian R, et al. Molecular mechanisms of adaptive evolution in wild animals and plants. Sci China Life Sci. 2023;66:453–95.
75. Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 2016;30:159–65.
76. Wu S, Han B, Jiao Y. Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. Mol Plant. 2020;13:59–71.
77. Yang Z, Liu Z, Xu H, Chen Y, Du P, Li P, et al. The chromosome-level genome of miracle fruit (*Synsepalum dulcificum*) provides new insights into the evolution and function of miraculin. Front Plant Sci. 2021;12:804662.
78. Wang Y, Chen F, Ma Y, Zhang T, Sun P, Lan M, et al. An ancient whole-genome duplication event and its contribution to flavor compounds in the tea plant (*Camellia sinensis*). Hortic Res. 2021;8:176.
79. Thomashow MF. Role of cold-responsive genes in plant freezing tolerance. Plant Physiol. 1998;118:1–8.
80. Thomashow MF. Molecular basis of plant cold acclimation: insights gained from studying the CBF cold response pathway. Plant Physiol. 2010;154:571–7.
81. Barrero-Sicilia C, Silvestre S, Haslam RP, Michaelson LV. Lipid remodelling: unravelling the response to cold stress in *Arabidopsis* and its extremophile relative *Eutrema salsugineum*. Plant Sci. 2017;263:194–200.
82. Su W, Ye C, Zhang Y, Hao S, Li QQ. Identification of putative key genes for coastal environments and cold adaptation in mangrove *Kandelia obovata* through transcriptome analysis. Sci Total Environ. 2019;681:191–201.
83. Jiang Y, Deyholos MK. Functional characterization of *Arabidopsis* NaCl-inducible *WRKY25* and *WRKY33* transcription factors in abiotic stresses. Plant Mol Biol. 2009;69:91–105.
84. Liu Q, Liu Y, Xin Z, Zhang D, Ge B, Yang R, et al. Genome-wide identification and characterization of the *WRKY* gene family in potato (*Solanum tuberosum*). Biochem Syst Ecol. 2017;71:212–8.
85. Li MY, Xu ZS, Tian C, Huang Y, Wang F, Xiong AS. Genomic identification of *WRKY* transcription factors in carrot (*Daucus carota*) and analysis of evolution and homologous groups for plants. Sci Rep. 2016;6:23101.
86. Kan L, Liao Q, Su Z, Tan Y, Wang S, Zhang L. Single-molecule real-time sequencing of the *Madhuca pasquieri* (Dubard) Lam. Transcriptome reveals the diversity of full-length transcripts. Forests. 2020;11:866.
87. Eulgem T, Rushton PJ, Robatzek S, Somssich IE. The *WRKY* superfamily of plant transcription factors. Trends Plant Sci. 2000;5:199–206.
88. Huang S, Gao Y, Liu J, Peng X, Niu X, Fei Z, et al. Genome-wide analysis of *WRKY* transcription factors in *Solanum lycopersicum*. Mol Genet Genomics. 2012;287:495–513.
89. Wei Y, Shi H, Xia Z, Tie W, Ding Z, Yan Y, et al. Genome-wide identification and expression analysis of the *WRKY* gene family in cassava. Front Plant Sci. 2016;7:25.
90. Yue M, Jiang L, Zhang N, Zhang L, Liu Y, Wang Y, et al. Importance of *FaWRKY71* in strawberry (*Fragaria* × *ananassa*) fruit ripening. Int J Mol Sci. 2022;23:12483.
91. Wang F, Hou X, Tang J, Wang Z, Wang S, Jiang F, et al. A novel cold-inducible gene from pak-choi (*Brassica campestris* ssp. *chinensis*), *BcWRKY46*, enhances the cold, salt and dehydration stress tolerance in transgenic tobacco. Mol Biol Rep. 2012;39:4553–64.

Publisher's note