GigaScience, 8, 2019, 1–9

doi: 10.1093/gigascience/giz136
Data Note

DATA NOTE

Trochodendron aralioides, the first chromosome-level draft genome in Trochodendrales and a valuable resource for basal eudicot research

Joeri S. Strijk [1,2,3,*,†], Damien D. Hinsinger [2,3,†], Fengping Zhang[4] and Kunfang Cao[1]

[1]State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, College of Forestry, Guangxi University, Daxuedonglu 100, Nanning, Guangxi 530005, China; [2]Biodiversity Genomics Team, Plant Ecophysiology & Evolution Group, Guangxi Key Laboratory of Forest Ecology and Conservation, College of Forestry, Daxuedonglu 100, Nanning, Guangxi 530005, China; [3]Alliance for Conservation Tree Genomics, Pha Tad Ke Botanical Garden, Khem Khong, PO Box 959, 06000 Luang Prabang, Laos and [4]Evolutionary Ecology of Plant Reproductive Systems Group, Key Laboratory of Biodiversity and Biogeography, 123 Lanhei Road, Kunming Institute of Botany, Kunming, China

[*]**Correspondence address.** Joeri S. Strijk, State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, College of Forestry, Guangxi University, Daxuedonlu 100, Nanning, Guangxi 530005, China E-mail: jsstrijk@hotmail.com http://orcid.org/0000-0003-1109-7015
[†]Contributed equally to this work.

Abstract

Background: The wheel tree (*Trochodendron aralioides*) is one of only 2 species in the basal eudicot order Trochodendrales. Together with *Tetracentron sinense*, the family is unique in having secondary xylem without vessel elements, long considered to be a primitive character also found in *Amborella* and Winteraceae. Recent studies however have shown that Trochodendraceae belong to basal eudicots and demonstrate that this represents an evolutionary reversal for the group. *Trochodendron aralioides* is widespread in cultivation and popular for use in gardens and parks. **Findings:** We assembled the *T. aralioides* genome using a total of 679.56 Gb of clean reads that were generated using both Pacific Biosciences and Illumina short-reads in combination with 10XGenomics and Hi-C data. Nineteen scaffolds corresponding to 19 chromosomes were assembled to a final size of 1.614 Gb with a scaffold N50 of 73.37 Mb in addition to 1,534 contigs. Repeat sequences accounted for 64.226% of the genome, and 35,328 protein-coding genes with an average of 5.09 exons per gene were annotated using *de novo*, RNA-sequencing, and homology-based approaches. According to a phylogenetic analysis of protein-coding genes, *T. aralioides* diverged in a basal position relative to core eudicots, ~121.8–125.8 million years ago. **Conclusions:** *Trochodendron aralioides* is the first chromosome-scale genome assembled in the order Trochodendrales. It represents the largest genome assembled to date in the basal eudicot grade, as well as the closest order relative to the core-eudicots, as the position of Buxales remains unresolved. This genome will support further studies of wood morphology and floral evolution, and will be an essential resource for understanding rapid changes that took place at the base of the Eudicot tree. Finally, it can further genome-assisted improvement for cultivation and conservation efforts of the wheel tree.

Received: 25 May 2019; **Revised:** 22 August 2019; **Accepted:** 24 October 2019

Keywords: *Trochodendron aralioides*; chromosome-level genome assembly; Hi-C assembly; basal eudicot

Data Description

Introduction

The Trochodendraceae family (order Trochodendrales) includes only 2 species (*Trochodendron aralioides* and *Tetracentron sinense*), both of which are commercially used and widely cultivated. *T. aralioides* (wheel tree) is a native species of the forests of Japan (Honshu—southwards from Yamagata Prefecture, Shikoku, Kyushu, Ryukyu Islands) and Taiwan. Although its hardiness extends to lower temperatures, it is generally restricted to lower temperate montane mixed forests between 600 and 1,700 m in Japan. In Taiwan, the range is more extensive, occurring in broad-leaved evergreen forest (2,000–3,000 m) in the central mountain ranges and in northern Taiwan between 500 and 1,250 m forming monotypic stands [1]. Over the past century, it has been repeatedly reported from Korea [2–10] although these occurrences are not confirmed in online repositories (e.g., [11]). Properties of *Trochodendron* (e.g., mild-warm temperate range, restricted elevational intervals, and natural occurrence in small discontinuous populations) make it difficult to predict the sensitivity of *T. aralioides* to the effects of projected changes in climate [12]. The fossil record shows that both the species diversity and distribution of the family were much more extensive and continuous during the Eocene (50–52 million years ago [Mya]) to Miocene [13, 14]. Unique for basal eudicots, Trochodendraceae have secondary xylem without vessel elements, a property only found in *Amborella* and Winteraceae [15]. This raises interesting questions on the biological conditions or triggers giving rise to such anatomical reversals, and the evolutionary and ecological consequences inherent in them.

Here, we constructed a high-quality chromosome-level reference genome assembly for *T. aralioides* (NCBI:txid4407) using long reads from the Pacific Biosciences (PacBio) DNA sequencing platform and a genome assembly strategy taking advantage of the Canu assembler [16]. This assembly of *T. aralioides* genome is the first chromosome-level reference genome constructed for the Trochodendrales order, and the closest relative to core eudicots sequenced to date. The completeness and continuity of the genome will provide high-quality genomic resources for studies on floral evolution and the rapid divergence of eudicots.

Genomic DNA extraction, Illumina sequencing, and genome size estimation

High-quality genomic DNA was extracted from freshly frozen leaf tissue of *T. aralioides* (Fig. 1) using the Plant Genomic DNA Kit (Tiangen, Beijing, PR China), following the manufacturer's instructions. After purification, a short-insert library (300–350 bp) was constructed and sequenced on the Illumina NovaSeq platform (Illumina Inc., San Diego, CA, USA), according to manufacturer guidelines. A total of ∼124.6 Gb of raw reads were generated.

Sequencing adapters were then removed from the raw reads, and reads from non-nuclear origins (e.g., chloroplast, mitochondrial, bacterial, and viral sequences) screened by aligning them to the nr database (NCBI [17]) using megablast v2.2.26 with the parameters " -v 1 -b 1 -e 1e-5 -m 8 -a 13 ". The in-house script duplication_rm.v2 was used to remove the duplicated read pairs, and low-quality reads were filtered as follows:

1) reads with ≥10% unidentified nucleotides (N) were removed;
2) reads with adapters were removed;
3) reads with >20% bases having Phred quality <5 were removed.

After the removal of low-quality and duplicated reads, ∼124.3 Gb of clean data (Supplementary Table S1) were used for the genome size estimation.

The *k*-mer peak occurred at a depth of 51 (Fig. 2), and we calculated the genome size of *T. aralioides* to be 1.758 Gb, with an estimated heterozygosity of 0.86% and a repeat content of 69.31%. This estimate is slightly smaller than the previously reported size, based on cytometry estimate (1.868 Gb) [18]. The GC content was 39.58% (Supplementary Fig. S1). A first genome assembly, using the Illumina data and the assembly program SOAPdenovo (SOAPdenovo, RRID:SCR_010752) [19], was ∼1.324 Gb total length, with a contig N50 of 740 bp and a scaffold N50 of 1.079 kb. This first attempt to assemble the wheel tree genome was of low quality, likely due to its high genomic repeat content and high heterozygosity level.

PacBio sequencing

High molecular weight (HMW) genomic DNA (gDNA) was sheared using a g-TUBE device (Covaris, Brighton, UK) with 20-kb settings. Sheared DNA was purified and concentrated with AmpureXP beads (Agencourt, Bioscience Corp., Beverly, MA, USA) and then used for single-molecule real time (SMRT) bell sequencing library preparation according to manufacturer's protocol (Pacific Biosciences; 20-kb template preparation using BluePippin size selection). Size-selected and isolated SMRTbell fractions were purified using AmpureXP beads (Agencourt, Bioscience Corp., Beverly, MA, USA) and these purified SMRTbells were finally used for primer-and-polymerase (P6) binding according to manufacturer's protocol (Pacific Biosciences). DNA-Polymerase complexes were used for MagBead binding and loaded at 0.1 nM on-plate concentration in 35 SMRT cells. Single-molecule sequencing was performed on a PacBio Sequel platform, yielding a total of 177.80 Gb filtered polymerase read bases (Supplementary Table S1).

10X Genomics sequencing

Libraries were built using a Chromium automated microfluidic system (10X Genomics, San Francisco, CA, USA) that allows the combination of the functionalized gel beads and HMW gDNA with oil to form a "gel bead in emulsion (GEM)." Each GEM contains ∼10 molecules of HMW gDNA and primers with unique barcodes and P5 sequencing adapters. After PCR amplification, P7 sequencing adapters are added for Illumina sequencing. Data were processed as follows: First, 16-bp barcode sequences and the 7-bp random sequences are trimmed from the reads, as well as low-quality pairs. We generated a total of 186.95 Gb raw data, and 183.52 Gb clean reads (Supplementary Table S1).

Hi-C sequencing data

To build a Hi-C library [20], nuclear HMW gDNA from *T. aralioides* leaves was cross-linked, then cut with the DpnII GATC restriction enzyme, leaving pairs of distally located but physically interacted DNA molecules attached to one another. The sticky

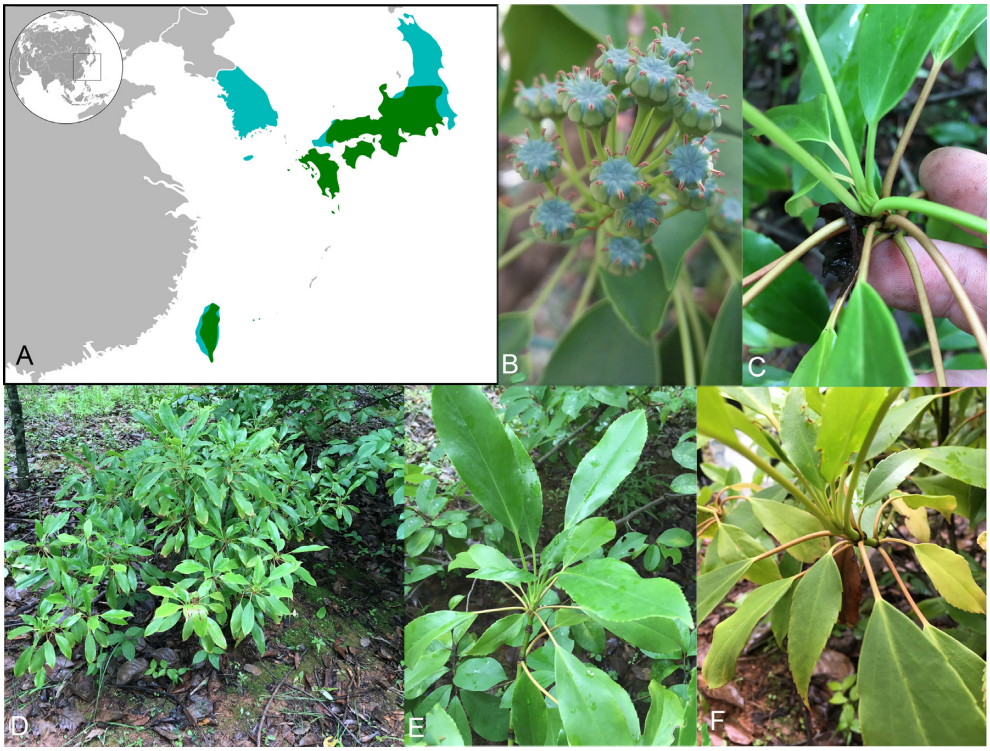

Figure 1: *Trochodendron aralioides* description. (A) Geographic distribution. Light blue: occurrence according to the Flora of China (at country level); green: occurrence according to Global Biodiversity Information Facility (GBIF); (B) flowers; (C) bud; (D) general habit; (E–F) stem and sprouting bud showing the wheel-like organization of leaves.

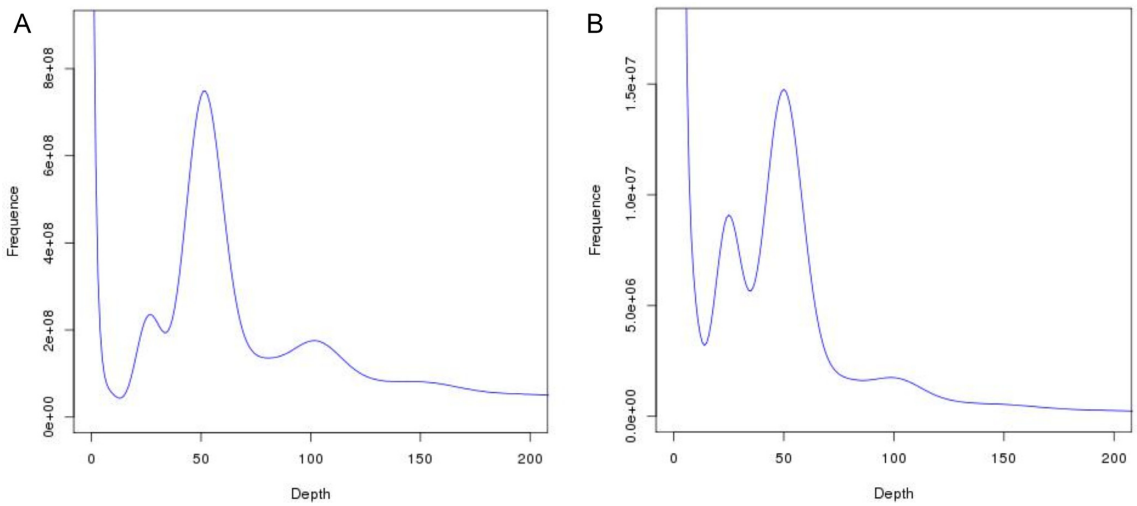

Figure 2: *k*-mer distribution of the *T. aralioides* genome. (A) *k*-mer depth and number frequency distribution; (B) *k*-mer depth and *k*-mer species number frequency distribution.

ends of these digested fragments were biotinylated and then ligated to each other to form chimeric circles. Biotinylated circles, that are chimeras of the physically associated DNA molecules from the original cross-linking, were enriched, sheared, and sequenced on an Illumina platform as described above. After adapter removal and filter of low-quality reads, there were a total of 193.90 Gb clean Hi-C reads. Sequencing quality assessment is shown in Supplementary Table S2.

De novo genome assembly

Short PacBio reads (<5 kb) were first used to correct the PacBio long reads using the "daligner" option in FALCON (Falcon, RRID:SCR_016089) [21], and to generate a consensus sequence. Following this error correction step, reads overlap was used to construct a directed string graph following Myers' algorithm [22]. Contigs were then constructed by finding the paths from the

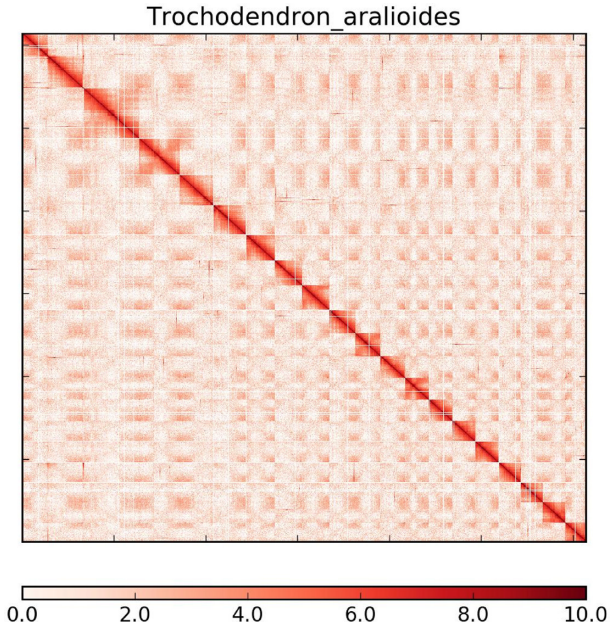

Figure 3: Hi-C interaction heat map for *T. aralioides* reference genome showing interactions between the 19 chromosomes.

string graph. Error correction of the preceding assembly was performed using the consensus-calling algorithm Quiver (PacBio Inc., Menlo Park, CA, USA) [23]. Reads were assembled and error-corrected with FALCON and Quiver to generate 4,226 contigs with a contig N50 length of 702 kb and total length of 1.607 Gb.

FragScaff [24] was used for 10X Genomics scaffolding, as follows:

1) Linked reads generated using the 10X Genomics library were aligned with BOWTIE v2 (Bowtie, RRID:SCR_005476) [25] against the consensus sequence of the PacBio assembly, to obtain Super-Scaffolds;
2) With increasing distance to consensus sequence, the number of linked reads supporting scaffold connection will decrease. Consensus sequences without linked read supports were then filtered and only the consensus sequence supported by linked reads was used for the subsequent assembly. FragScaff scaffolding resulted in 1,469 scaffolds, with a scaffold N50 length of 3.38 Mb.

To assess the completeness of the assembled *T. aralioides* genome, we performed a BUSCO analysis by searching against the plant (BUSCO, RRID:SCR_015008, version 3.0) [26]. Overall, 91.4% and 2.8% of the 1,440 expected genes were identified in the assembled genome as complete and partial, respectively (Supplementary Table S3). Overall, 94.2% (1,356) genes were found in our assembly. We also assessed the completeness of conserved genes in the *T. aralioides* genome by Core Eukaryotic Genes Mapping Approach (CEGMA, RRID:SCR_015055) [27]. According to CEGMA, 232 conserved genes in *T. aralioides* were identified, which have 93.55% completeness compared to the sets of CEGMA (Supplementary Table S3).

The Hi-C clean data were aligned against the PacBio reads assembly using BWA (BWA, RRID:SCR_010910) [28]. Only the read pairs with both reads aligned to contigs were considered for scaffolding. For each read pair, its physical coverage was defined as the total bp number spanned by the sequence of reads and the gap between the 2 reads when mapping to contigs. Per-base physical coverage for each base in the contig was calculated as the number of read pairs' physical coverage it contributes too. Misassembly can be detected by the sudden drop in per-base physical coverage in a contig.

Following the physical coverage of the resulting alignment, any misassembly was split to apply corrections. Using the clustering output, the order and orientation of each contig interaction was assessed on intensity of contig interaction and the position of the interacting reads. Combining the linkage information and restriction enzyme site, the string graph formulation was used to construct the scaffold graph using LACHESIS [29], and 1,469 scaffolds of our draft genome were clustered to 19 chromosomes (Figure 3, Supplementary Table S4). The *T. aralioides* genome information is summarized in Supplementary Table S5.

Repeat sequences in the wheel tree genome

Transposable elements (TEs) in the genome assembly were identified both at the DNA and protein level. RepeatModeler (RepeatModeler, RRID:SCR_015027) [30], RepeatScout (RepeatScout, RRID:SCR_014653) [31], and LTR_FINDER (LTR_Finder, RRID:SCR_015247) [32] were used to build a *de novo* TE library with default parameters. RepeatMasker [33] was used to map the repeats from the *de novo* library against Repbase [34]. Uclust [35] was then used with the 80-80-80 rule [36] to combine the results from the above software. At the protein level, RepeatProteinMask in the RepeatMasker package was used to identify TE-related proteins with WU-BLASTX searches against the TE protein database. Overlapping TEs belonging to the same type of repeats were merged.

Repeat sequences accounted for 64.2% of the *T. aralioides* genome, with 57.2% of the genome identified from the *de novo* repeat library (Table 2). Approximately 53.2% of the *T. aralioides* genome was identified as long terminal repeat (LTR) (most often TEs). Among them, LTRs were the most abundant type of repeat sequences, representing 53.249% of the whole genome. Long interspersed nuclear elements (LINEs) and DNA TE repeats accounted for 0.837% and 2.416% of the whole genome, respectively (Table 2, Supplementary Fig. S2).

The transfer RNA (tRNA) genes were identified by tRNAscan-SE (tRNAscan-SE, RRID:SCR_010835) [37] with the eukaryote set of parameters. The ribosomal RNA fragments were predicted by aligning them to *Arabidopsis thaliana* and *Oryza sativa* reference ribosomal RNA sequences using BlastN (BLASTN, RRID:SCR_001598) (E-value of 1E−10) [38]. The mitochondrial RNA and small nuclear RNA genes were predicted using INFERNAL (Infernal, RRID:SCR_011809) [39] by searching against the Rfam database (release 9.1) (Supplementary Table S6).

Gene Annotation

RNA preparation and sequencing

RNA sequencing (RNA-seq) was conducted for 4 tissue libraries (leaf, stem, bark, and bud) from the same individual as for the genome sequencing and assembly. A total of 8 libraries were constructed (Supplementary Table S7). Total RNA was extracted using the RNAprep Pure Plant Kit (Tiangen, Beijing, PR China) and gDNA contamination was removed with the RNase-Free DNase I (Tiangen, Beijing, PR China). RNA quality was determined based on the estimation of the ratio of absorbance at 260 nm/280 nm (OD = 2.0) and the RIN (value = 9.2) by using a Nanodrop ND-1000 spectrophotometer (LabTech, Wilmington, DE, USA) and a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), respectively. The complementary DNA libraries were

Table 1: Summary of *Trochodendron aralioides* genome assembly and annotation

Parameter	Draft scaffolds	Chromosome-length scaffolds based on Hi-C
Genome assembly		
Length of genome (bp)	1,623,741,898	1,530,107,441
Number of contigs	4,226	2,744
Contigs N50 (bp)	702,251	740,603
Number of scaffolds	1,469	19
Scaffold N50 (bp)	3,938,440	73,365,148
Genome coverage (X)	278.34	398.07
Number of contigs (>100 kb)	3,062	2,744
Total length of contigs (>100 kb)	1,567,464,199	1,523,319,687
Mapping rate of contigs	0.9779	
Genome annotation		
Protein-coding gene number		35,328
Mean transcript length (kb)		10,622.49
Mean exons per gene		5.09
Mean exon length (bp)		232.46
Mean intron length (bp)		2,308.46

Table 2: Detailed classification of repeat sequences identified in *Trochodendron aralioides*

	de novo + Repbase		TE proteins		Combined TEs	
Type	Length (bp)	% in genome	Length (bp)	% in genome	Length (bp)	% in genome
DNA	38,995,970	2.416	15,755,819	0.976	50,171,660	3.108
LINE	13,503,128	0.837	39,566,384	2.451	47,644,917	2.952
SINE	143,207	0.00887	0	0	143,207	0.00887
LTR	859,515,257	53.249	327,748,739	20.305	908,751,606	56.23
Unknown	13,395,729	0.83	0	0	13,395,729	0.83
Total	922,704,692	57.164	382,460,417	23.694	1,006,355,712	62.347

De novo + Repbase: annotations predicted *de novo* by RepeatModeler, RepeatScout, and LTR_FINDER; TE proteins: transposon elements annotated by RepeatMasker; combined TEs: merged results from approaches above, with overlap removed; unknown: repeat sequences RepeatMasker cannot classify.

constructed with the NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA), following the manufacturer's recommendations. Libraries were sequenced on an Illumina HiseqXTen platform (Illumina Inc., San Diego, CA, USA), generating 150-bp paired-end reads. A total of 26.7 Gb of clean RNA-seq sequences were produced, with ≥93.69% of the bases with quality >Q20 (Supplementary Table S7).

RNA clean reads were both assembled into 312,246 sequences using Trinity (Trinity, RRID:SCR_013048) [40] and then annotated, and mapped against the genomic sequence using tophat v2.0.8 (TopHat, RRID:SCR_013035) [41]; then cufflinks v2.1.1 (Cufflinks, RRID:SCR_014597) [42] was used to assemble transcripts into gene models. Finally, all annotation results were combined with EVidenceModeler (EVM) (EVidenceModeler, RRID:SCR_014659) [43] to obtain the final non-redundant gene set.

Annotation

Gene annotation was performed using 3 approaches:

- A homology-based approach, in which the protein sequences from *O. sativa*, *Aquilegia coerulea*, *Fraxinus excelsior*, *Nelumbo nucifera*, *Quercus robur*, and *Vitis vinifera* were aligned to the genome by using TblastN (TBLASTN, RRID:SCR_011822) [38] with an E-value cutoff by 1E−5. BLAST hits were conjoined with Solar software [44]. For each BLAST hit, Genewise (GeneWise, RRID:SCR_015054) [45] was used to predict the exact gene structure in the corresponding genomic regions.
- An *ab initio* gene prediction approach, using Augustus v2.5.5 (Augustus, RRID:SCR_008417) [46], Genescan v1.0 (GENSCAN, RRID:SCR_012902), GlimmerHMM v3.0.1 (GlimmerHMM, RRID:SCR_002654) [47], Geneid [48], and SNAP (SNAP, RRID:SCR_007936) [49] to predict coding genes on the repeat-masked *T. aralioides* genome.
- A transcriptome-based approach, in which RNA-seq data were mapped to the genome (see above).

All gene models predicted from the above 3 approaches were combined by EVM into a non-redundant set of gene structures. Then, we filtered out low-quality gene models, defined as follows: (i) coding region lengths ≤150 bp, (ii) models supported only by *ab initio* methods and with FPKM < 1.

We identified an average of 5.1 exons per gene (mean length of 10.622 kb) in the *T. aralioides* genome (Table 1). The gene number, gene length distribution, coding sequence (CDS) length distribution, exon length distribution, and intron length distribution were all comparable to those of selected angiosperm species (Supplementary Table S8, Supplementary Fig. S3).

Functional annotation was performed by blasting the protein-coding gene sequences against SwissProt and TrEMBL [50] using BLASTP (BLASTP, RRID:SCR_001010) (E-value 1E−5) [51]. The annotation information of the best BLAST hit from the databases was transferred to our gene set annotations. Protein domains were annotated by searching the InterPro (v32.0) and Pfam (V27.0) databases using InterProScan v4.8 (InterProScan, RRID:SCR_005829) [52] and Hmmer v3.1 (Hmmer, RRID:SCR_005305) [53], respectively. Gene Ontology (GO) terms for each gene

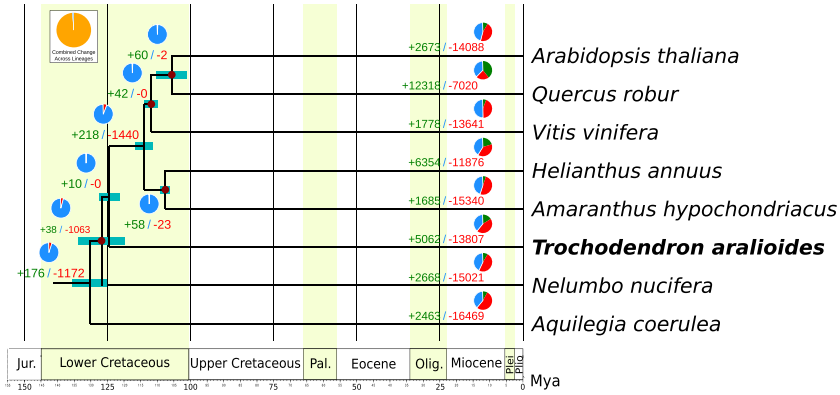

Figure 4: Phylogenetic tree and number of gene families displaying expansion and contraction among 12 plant species. The pie charts show the expanded (green), contracted (red), and conserved (blue) gene family proportions among all gene families. Estimated divergence time confidence interval are shown at each internal node as teal bars. Calibrated nodes are indicated by red dots (see text for details on calibration scheme).

were obtained from the corresponding InterPro or Pfam entry. The pathways in which the gene might be involved were assigned by blasting them against the KEGG database (release53), with an E-value cutoff of 1E−5. The genes successfully annotated in GO were classified into biological process (BP), cellular component (CC), and molecular function (MF). Ultimately, 95.4% (33,696 genes) of the 35,328 genes were annotated by ≥1 database (Supplementary Table S9).

Gene family identification and phylogenetic analyses of wheel tree

Orthologous relationships between genes of *Amaranthus hypochondriacus*, *Amborella trichopoda*, *Annona muricata*, *A. coerulea*, *A. thaliana*, *Helianthus annuus*, *Cinnamomum kanehirae*, *Musa acuminata*, *N. nucifera*, *O. sativa*, *Q. robur*, and *V. vinifera* were inferred through all-against-all protein sequence similarity searches with OrthoMCL [54] and only the longest predicted transcript per locus was retained (Supplementary Figs S4 and S5). This resulted in 31,290 orthologous groups with representatives in each species being identified. Among these gene families, 484 were single-copy orthologs, and used for phylogenomic reconstruction and positive selection analyses. The similarity among these orthologs ranged from 34.8% to 80.3% (mean: 60.64% ± 7.33%, distribution shown in Supplementary Fig. S6a).

For each gene family, an alignment was produced using Muscle (MUSCLE, RRID:SCR_011812) [55], and ambiguously aligned positions trimmed using Gblocks (Gblocks, RRID:SCR_015945) [56] with default parameters (-b3 8; -b4 10; -b5 n). A maximum likelihood (ML) tree was inferred using RAxML 7.2.9 (RAxML, RRID:SCR_006086) [57].

Divergence times between species were calculated using Markov Chain Monte Carlo methods, as implemented in PAML (PAML, RRID:SCR_014932) [58].

Node calibrations were defined from the TimeTree database [59] as follows:

- the divergence between *A. thaliana* and *Q. robur* (97–109 Mya);
- the divergence between *H. annuus* and *A. hypochondriacus* (107–116 Mya);
- the divergence between *A. thaliana* and *V. vinifera* (109–114 Mya);
- the divergence between *N. nucifera* and *V. vinifera* (116–127 Mya);
- the divergence between *M. acuminata* and *O. sativa* (90–115 Mya);
- the divergence between *A. thaliana* and *O. sativa* (140–200 Mya);
- the divergence between *A. trichopoda* and *O. sativa* (168–194 Mya).

The basal eudicot grade's most basal representative, namely, *A. coerulea*, diverged from other angiosperms during the lower Cretaceous 130.2 Mya (95% CI, 126.6–136.2 Mya), while *T. aralioides*, the most recently diverged basal eudicot, diverged from the core eudicots ∼124 Mya (95% CI, 121.8–125.8 Mya). Finally, the divergence between rosids and asterids, and thus the crown age of core eudicots, was reconstructed as 114.0 Mya (95% CI, 111.3–116.2 Mya).

To identify gene families that experienced a significant expansion or contraction during the evolution of the wheel tree, we used the likelihood model implemented in CAFE (CAFÉ, RRID:SCR_005983) [60], with default parameters. The phylogenetic tree topology and branch lengths were taken into account to infer the significance of change in gene family size in each branch (Fig. 4). The gene families that experienced the most significant expansions were mainly involved in pathogen/stress response (e.g., the cyanoamino-acid metabolism, $P = 2.35 \times 10^{-28}$; the plant-pathogen interaction map, $P = 2.29 \times 10^{-22}$; the tryptophan metabolism, $P = 5.85 \times 10^{-10}$) (Supplementary Table S10).

Positive selection

Positive selection is a major driver of biological adaptation. Using the protein-coding sequences, we calculated the number of synonymous substitutions per site (Ks) and nonsynonymous substitutions per site (Ka) and assessed the deviation from zero of the difference Ka − Ks. A Ka/Ks > 1 represents evidence of positive selection. The Ka/Ks ratio ranged from 0.0061 to 5.7689 (mean: 0.6976 ± 0.9173, see Supplementary Fig. S6b), with Ks ranging from 0.0030 to 0.6281 (mean: 0.1585 ± 0.0749). MUSCLE [55] was used to align the protein and nucleotide sequences; then we used Gblocks [56] with default parameters (-b3 8; -b4 10; -b5 n) to eliminate poorly aligned positions and divergent regions from the alignment. The maximum likelihood–based branch length test of the PAML package [58] was used for comparisons, and the ratio Ka/Ks was calculated over the entire length of the protein-coding gene. Using *T. aralioides* as the foreground branch and *A. hypochondriacus*, *H. annuus*, *N. nucifera*, and

A. coerulea as background branches, we identified 238 genes that were considered as candidate genes under positive selection (*P*-value < 0.01, false discovery rate < 0.05) using a maximum likelihood–based branch length test. The GO terms and KEGG pathways for these genes showed that positive selection was especially detected in cell metabolism (such as vitamins and amino acid biosynthesis; Supplementary Table S11).

Whole-genome duplication analysis

MCScan [61] was used to identify collinear segments within *Trochodendron* and between the *T. aralioides* and other angiosperm genomes. The sequences of the gene pairs contained in the (inter-) collinear segments of the genome were extracted and the codeml tool in the PAML package [58] was used to calculate the value of the 4-fold synonymous third-codon transversion rate (4dTv). The distribution of 4dTv can reflect whether genome-wide replication events occur in the evolutionary history of species, the relative time of genome-wide replication events, and the divergent events among species.

The 4dTv values of all paralogous gene pairs in *T. aralioides* were calculated, as well as those in *A. coerulea*, *H. annuus*, and *A. muricata* for comparisons. In addition, the 4dTv values were calculated for all ortholog gene pairs between *T. aralioides* and *A. coerulea*, *H. annuus*, and *A. muricata* to observe species divergence events. The peak of 4dTv distribution in the *T. aralioides* genome was around ∼0.1, whereas the peaks for interspecific comparisons with *A. coerulea* and *H. annuus* were around ∼0.3 and ∼0.2, respectively (Supplementary Fig. S7). All together, these 4dTv distributions indicate that a whole-genome replication event occurred in *T. aralioides* after its divergence from both other basal angiosperms and core eudicots.

Conclusions

We successfully assembled the genome of *T. aralioides* and report the first chromosome-level genome sequencing, assembly, and annotation based on long reads from the third-generation PacBio Sequel sequencing platform for basal eudicotyledons. The final draft genome assembly is ∼1.614 Gb, which is slightly smaller than the estimated genome size based on *k*-mer analysis (1.758 Gb) and on cytometry (1.868 Gb, [18]). With a contig N50 of 691 kb and a scaffold N50 of 73.37 Mb, the chromosome-level genome assembly of *T. aralioides* is the first high-quality genome in the Trochodendrales order. We also predicted 35,328 protein-coding genes from the generated assembly, and 95.4% (33,696 genes) of all protein-coding genes were annotated. We found that the divergence time between *T. aralioides* and its common ancestor with the core eudicots was ∼124.2 Mya. The chromosome-level genome assembly together with gene annotation data generated in this work will provide a valuable resource for further research on floral morphology diversity, on the early evolution of eudicotyledons, and on the conservation of this iconic tree species.

Availability of Supporting Data and Materials

Raw reads were deposited to EBI (project PRJEB32669), and supporting data and materials are available in the *GigaScience* GigaDB database [62].

Additional Files

Supplementary Figure S1. GC content analysis of *T. aralioides* genome based on Illumina reads for genome size survey.

Supplementary Figure S2. Divergence distribution of transposable elements in the genome of *T. aralioides*. Units in Kimura substitution level (CpG adjusted).

Supplementary Figure S3. Gene characteristics in *T. aralioides* and other angiosperms. From left to right and top to bottom: lengths of messenger RNA; lengths of exons in coding regions; number of exons per gene; lengths of introns in genes; lengths of coding regions (CDS). Aco: *Aquilegia caeruleus*; Fex: *Fraxinus excelsior*; Nun: *Nelumbo nucifera*; Osa: *Oryza sativa*; Qro: *Quercus robur*; Tar: *Trochodendron aralioides*; Vvi: *Vitis vinifera*.

Supplementary Figure S4. Comparing orthogroups between *T. aralioides* and other angiosperm species. Aco: *Aquilegia caeruleus*; Ahy: *Amaranthus hypochondriacus*; Amu: *Annona muricata;* Ath: *Arabidopsis thaliana*; Atr: *Amborella Trichopoda*; Cka: *Cinnamomum micranthum*; Han: *Helianthus annuus*; Mac: *Musa acuminata*; Nnu: *Nelumbo nucifera*; Osa: *Oryza sativa*; Qro: *Quercus robur*; Tar: *Trochodendron aralioides*; Vvi: *Vitis vinifera*.

Supplementary Figure S5. Orthologous gene families across 4 angiosperm genomes (*Trochodendron aralioides*, *Annona muricata*, *Amaranthus hypochondriacus*, and *Aquilegia coerulea*).

Supplementary Figure S6. Characteristics of the 484 orthologs found between *T. aralioides* and selected angiosperms (see text for details). (a) Distribution of the similarity among each ortholog; genes are counted vertically (red bars) for bins of 1% unit (horizontally). (b) Distribution of the genes' Ka/Ks ratio; genes are counted vertically (red bars) for bins of 0.1 Ka/Ks unit (horizontally).

Supplementary Figure S7. Distribution of 4dTv values in several species pair comparisons, highlighting potential whole-genome duplications (WGD). Aco: *Aquilegia caeruleus*; Amu: *Annona muricata;* Han: *Helianthus annuus*.

Abbreviations

4dTv: 4-fold synonymous third-codon transversion rate; BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; CDS: coding sequence; CEGMA: Core Eukaryotic Genes Mapping Approach; Gb: gigabase pairs; GC: guanine-cytosine; gDNA: genomic DNA; GO: Gene Ontology; HMW: high molecular weight; KEGG: Kyoto Encyclopedia of Genes and Genomes; LINE: long interspersed nuclear element; LTR: long terminal repeats; Mb: megabase pairs; Mya: million years ago; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PAML: Phylogenetic Analysis by Maximum Likelihood; RAxML: Randomized Axelerated Maximum Likelihood; SINE: short interspersed nuclear repeat; SMRT: single-molecule real time; SNAP: Scalable Nucleotide Alignment Program; TE: transposable element; tRNA: transfer RNA.

Competing Interests

The authors declare that they have no competing interests.

Funding

Genome sequencing, assembly, and annotation were conducted by the Novogene Bioinformatics Institute, Beijing, China;

mutual contract No. C101SC17090417. This work was supported through the Bagui Scholarship team funding under Grant No. C33600992001, the Guangxi Province One Hundred Talent program and Guangxi University to J.S.S., and by funding from the China Postdoctoral Science Foundation under Grant No. 2015M582481 and No. 2016T90822 to D.D.H., and the National Science Foundation of China under Grant No. 31470469 to K.F.C..

Authors' Contributions

Conceptualization: J.S.S., D.D.H.; funding: J.S.S., C.K.F.; sample collecting and treatment; D.D.H., F.P.Z.; investigation: J.S.S., D.D.H.; writing original draft: J.S.S., D.D.H.; review and editing: J.S.S., D.D.H., F.P.Z., C.K.F.

References

1. Li HL, Chaw SM, Trochodendraceae. In Flora of Taiwan, 2nd ed.; Huang, TC, Editorial Committee of the Flora of Taiwan , Eds.; Department of Botany, National Taiwan University: Taipei, Taiwan, 1996; **Volume 2**, 504–5.
2. Ohwi J. Flora of Japan. Washington, DC: Smithsonian Institution; 1965.
3. Hogan S. Trees for all seasons: broadleaved evergreens for temperate climates. Portland, OR: Timber Press; 2008.
4. Mabberley DJ. Mabberley's Plant-Book: A Portable Dictionary of Plants, Their Classification and Uses. 3rd ed. Cambridge University Press; 2008.
5. Hilliers J, Coombes A. The Hilliers Manual of Trees and Shrubs. Devon, UK: David & Charles; 2002.
6. Phillips R, Rix M. The Botanical Garden: Trees and Shrubs. Richmond Hill, ON, Canada: Firefly Books; 2002.
7. Jacobson AL. North American Landscape Trees. Berkeley, CA: Ten Speed Press; 1996.
8. Krussman G. Trochodendron. Manual of Cultivated Trees and Shrubs. Portland, OR: Timber Press; 1985.
9. Bean WJ. Trochodendron. Trees and Shrubs Hardy in the British Isles. 8th ed. John Murray; 1980.
10. Rehder A. Manual of Cultivated Trees and Shrubs. New York, NY: Macmillan; 1940.
11. Plants of the World Online. http://plantsoftheworldonline.org/. Accessed 1 March, 2019.
12. Larsen T, Brehm G, Navarrete H, et al. Range shifts and extinctions driven by climate change in the tropical Andes: synthesis and directions. In: Climate Change and Biodiversity in the Tropical Andes. Inter-American Institute of Global Change Research and Scientific Committee on Problems of the Environment; 2011:47–67.
13. Manchester SR, Pigg KB, Kvaček Z, et al. Newly recognized diversity in trochodendraceae from the Eocene of western North America. Int J Plant Sci 2018;**179**:663–76.
14. Manchester SR, Pigg KB, Devore ML. Trochodendraceous fruits and foliage in the Miocene of western North America. Fossil Imprint 2018;**74**:45–54.
15. Cronk QCB, Forest F. The evolution of angiosperm trees: from palaeobotany to genomics. In: Comparative and Evolutionary Genomics of Angiosperm Trees. Cham: Springer; 2017:1–17.
16. Koren S, Walenz BP, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;**27**:722–36.
17. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov. Accessed 15 January 2019.
18. Hanson L, Brown RL, Boyd A, et al. First nuclear DNA C-values for 28 angiosperm genera. Ann Bot 2003;**91**:31–8.
19. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 2012;**1**:18.
20. Lieberman-Aiden E, Van Berkum NL, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;**326**:289–93.
21. Chin C-S, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 2016;**13**:1050–4.
22. Myers EW. The fragment assembly string graph. Bioinformatics 2005;**21**(suppl_2):ii79–85.
23. Chin CS, Alexander DH, Marks P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 2013;**10**:563–6.
24. Adey A, Kitzman JO, Burton JN, et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. Genome Res 2014;**24**(12):2041–9.
25. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;**9**:357–66.
26. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**:3210–2.
27. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007;**23**:1061–7.
28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;**25**:1754–60.
29. Burton JN, Adey A, Patwardhan RP, et al. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol 2013;**31**:1119–54.
30. Smit AFA, Hubley R. RepeatModeler Open-1.0. 2018. http://www.repeatmasker.org . Accessed 15 January 2019.
31. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics 2005;**21**:i351–8.
32. Xu Z, Wang H. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007;**35**:W265–8.
33. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 2009;**25**:4–10.
34. Smit A, Hubley R, Green P. RepeatMasker Open-4.0.6 2013–2015. 2017. http://www.repeatmasker.org. Accessed 15 January 2019.
35. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;**26**:2460–1.
36. Wicker T, Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 2007;**8**:973.
37. Lowe TM, Eddy SR. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1996;**25**:955–64.
38. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;**10**:421–9.
39. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013;**29**:2933–5.
40. Grabher MG, Haas BJ, Yassour M, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol 2013;**29**:644.

41. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 2013;**14**:R36.
42. Ghosh S, Chan CK. Analysis of RNA-Seq data using TopHat and cufflinks. Methods Mol Biol 2016;**1374**:339–61.
43. Haas BJ, Salzberg SL, Zhu W, et al. Automated eukaryotic gene structure annotation using EVidence modeler and the program to assemble spliced alignments. Genome Biol 2008;**9**:1.
44. Yu X, Zheng H, Wang J, et al. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics 2006;**88**:745–51.
45. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res 2004;**14**:988–95.
46. Stanke M, Keller O, Gunduz I, et al. AUGUSTUS: Aa initio prediction of alternative transcripts. Nucleic Acids Res 2006;**32**:W309–12.
47. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 2004;**20**:2878–9.
48. Blanco E, Parra G, Guigó R. Using geneid to identify genes. Curr Protoc Bioinformatics 2007;**18**:4–7.
49. Johnson AD, Handsaker RE, Pulit SL, et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;**24**:2938–9.
50. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res 2014;**43**:D204–8.
51. Gish W, States DJ. Identification of protein coding regions by database similarity search. Nat Genet 1993;**3**:266–6.
52. Quevillon E, Silventoinen V, Pillai S, et al. InterProScan: protein domains identifier. Nucleic Acids Res 2005;**33**:W116–20.
53. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011;**39**:W29–37.
54. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 2003;**13**:2178–89.
55. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;**32**(5):1792–7.
56. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 2000;**17**:540–52.
57. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;**30**:1312–3.
58. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;**24**:1586–91.
59. Timetree. http://www.timetree.org/. Accessed 15 January 2019.
60. De Bie T, Cristianini N, Demuth JP, et al. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 2006;**22**:1269–71.
61. Wang Y, Tang H, Debarry JD, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 2012;**40**:e49.
62. Strijk JS, Hinsinger D, Zhang F, et al. Supporting data for "*Trochodendron aralioides*, the first chromosome-level draft genome in Trochodendrales and a valuable resource for basal eudicot research." GigaScience Database 2019. http://dx.doi.org/10.5524/100657.