OXFORD (GIGA)^n SCIENCE

GigaScience, 8, 2019, 1–10

doi: 10.1093/gigascience/giz085
Data Note

DATA NOTE

De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China

Jing Yang[1,2,†], Hafiz Muhammad Wariss[1,2,3,†], Lidan Tao[1,2], Rengang Zhang[4], Quanzheng Yun[4], Peter Hollingsworth[5], Zhiling Dao[1,2], Guifen Luo[1,2], Huijun Guo[6], Yongpeng Ma[1,2,*] and Weibang Sun[1,2,7,*]

[1]Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China; [2]Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China; [3]University of Chinese Academy of Sciences, Beijing, 100049, China; [4]Beijing Ori-Gene Science and Technology Co. Ltd, Beijing, 102206, China; [5]Royal Botanic Garden Edinburgh, 20a Inverleith Row, Edinburgh, UK; [6]Southwest Forestry University, Kunming, 650224, Yunnan, China and [7]Kunming Botanical Garden, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China

[*]**Correspondence address.** Yongpeng Ma, E-mail: mayongpeng@mail.kib.ac.cn http://orcid.org/0000-0002-7725-3677 and Weibang Sun, E-mail: wbsun@mail.kib.ac.cn
[†]These authors contributed equally to this work.

Abstract

Background: *Acer yangbiense* is a newly described critically endangered endemic maple tree confined to Yangbi County in Yunnan Province in Southwest China. It was included in a programme for rescuing the most threatened species in China, focusing on "plant species with extremely small populations (PSESP)". **Findings:** We generated 64, 94, and 110 Gb of raw DNA sequences and obtained a chromosome-level genome assembly of *A. yangbiense* through a combination of Pacific Biosciences Single-molecule Real-time, Illumina HiSeq X, and Hi-C mapping, respectively. The final genome assembly is ∼666 Mb, with 13 chromosomes covering ∼97% of the genome and scaffold N50 sizes of 45 Mb. Further, BUSCO analysis recovered 95.5% complete BUSCO genes. The total number of repetitive elements account for 68.0% of the *A. yangbiense* genome. Genome annotation generated 28,320 protein-coding genes, assisted by a combination of prediction and transcriptome sequencing. In addition, a nearly 1:1 orthology ratio of dot plots of longer syntenic blocks revealed a similar evolutionary history between *A. yangbiense* and grape, indicating that the genome has not undergone a whole-genome duplication event after the core eudicot common hexaploidization. **Conclusion:** Here, we report a high-quality *de novo* genome assembly of *A. yangbiense*, the first genome for the genus *Acer* and the family Aceraceae. This will provide fundamental conservation genomics resources, as well as representing a new high-quality reference genome for the economically important *Acer* lineage and the wider order of Sapindales.

Received: 17 March 2019; **Revised:** 17 May 2019; **Accepted:** 24 June 2019

Keywords: *Acer yangbiense*; PSESP; PacBio sequencing; genome assembly; genome annotation

Data Description

Background information

The genus *Acer* L., commonly known as maple, is one of the most important genus of trees and shrubs in the Northern Hemisphere [1–5]. *Acer* exhibits a classical pattern of biogeographic disjunction across Europe, Northern Africa, Asia, and North America, with the greatest species richness in Eastern Asia [2, 4, 6–12]. It is a wide-ranging genus comprising up to 129 species worldwide with maximum diversity in China, where ∼99 (61 endemic, 3 introduced) species are recognized [13]. The base chromosome number of *Acer* is x = 13, and cytological investigation indicates a range of ploidy levels including diploids, tetraploids, hexaploids, octoploids, and aneuploids [14]. *Acer* has long been of interest to botanists for its remarkable diversity, especially of leaves, fruits, and bark, and for its intercontinental disjunct distribution [7]. The colorful foliage of maples is a charismatic landscape feature, with vivid hues of red, yellow, and orange in the autumn. In addition to being ornamental, many species are sources of commercial products, such as maple syrup, furniture, and timber [15]. Maple has been found to contain a large number of phytochemicals that have antioxidant, antitumor, and anti-inflammatory activities [15–20].

Acer yangbiense Y. S. Chen & Q. E. Yang (Aceraceae, NCBI:txid1000413) is a newly described Chinese maple species (Fig. 1) [21]. It has a restricted distribution range of 2,200–2,500 m altitudes in the western valley of Cangshan Mountain, Yunnan Province, China. This species is facing a very high risk of extinction because of its small population size, poor reproduction, and habitat degradation [22]. The species was categorized as critically endangered (CR) by Gibbs and Chen in 2009 [23], and as only 5 individuals were recorded based on Qin et al. (2017) [24–26] in the first decade after its description. In 2016, further survey work recovered a total of 577 individuals from 12 localities [27]. This is the most accurate available population estimate of *A. yangbiense*.

A. yangbiense is classified as a “plant species with extremely small populations” (PSESP) by the Chinese government and included in the PSESP rescue plan [28, 29]. The concept of PSESP emphasizes species that face high risk of extinction, characterized by small remaining populations in restricted habitats and being subjected to severe human disturbance [28, 30]. It is targeted at species with <5,000 mature individuals in total and <500 mature individuals in each isolated population [31]. Genetic studies done by Yang et al. (2015) suggested that *A. yangbiense* was not genetically depauperate, but further parentage analysis indicated a high selfing rate in seedlings of *A. yangbiense* [25]. The current threatened status of *A. yangbiense* serves to emphasize that an effective conservation strategy is urgently required.

The generation of plant genome sequences and assemblies allows detailed insights into the evolutionary history of species and provides information to support sustainable conservation [32]. Here, we present a high-quality genome assembly of *A. yangbiense* as a valuable resource and reference for future population genomic studies. The availability of a fully sequenced and annotated genome is essential to resolve fundamental questions about *A. yangbiense* diversification and provide new insights into its demographic history, with important implications for future conservation efforts.

Plant material

Fresh young leaves were collected from *ex situ* conserved *A. yangbiense* at the Kunming Botanical Garden (KBG) of the Kunming Institute of Botany, Chinese Academy of Sciences. This tree was grown from seed in 2009, from seeds originally collected from Malutang, Yangbi County, Dali, Yunnan (Fig. 1) (25.7489 N latitude, 100.0064 E longitude, 2,474 m elevation). For genome library preparation, only leaf tissues were used; for transcriptome sequencing, samples were obtained from 5 different tissues: leaf buds, young leaves, young stems, roots, and fruits from healthy individuals planted in KBG in June and July 2018, respectively. All samples were collected with permission from KBG. For RNA samples, tissues were immediately transferred into liquid nitrogen and stored in dry ice until RNA extraction; for DNA samples, tissues were immediately stored in dry ice until DNA extraction.

PacBio SMRT sequencing

Genomic DNA with high quality and high molecular weight was extracted from fresh leaves using a cetyl trimethylammonium bromide protocol [33]. Libraries for single-molecule real-time (SMRT) Pacific Biosciences (PacBio) genome sequencing were constructed following the standard protocols of PacBio at Beijing Ori-Gene Science and Technology Co., Ltd (Beijing, China). Briefly, 50 μg of high-quality genomic DNA was sheared to ∼20 kb targeted size, followed by damage repair and end repair, blunt-end adapter ligation, and size selection. Finally, the libraries were sequenced on the PacBio Sequel platforms using S/P2-C2 sequencing chemistry (10 SMRT cells). A total of 6.3 million PacBio reads with ∼64 Gb sequencing data were generated, with an average read length of 10 kb. The longest read was 93 kb, and N50 was 16.8 kb (Supplementary Table S1).

Illumina sequencing

The Illumina libraries were constructed according to the standard manufacturer’s PCR-free protocol (Illumina). Short-insert libraries of 300–500 bp were prepared using 2 μg of whole-genomic DNA for Illumina sequencing. All the libraries were sequenced on Illumina HiSeq X platform with paired-end (PE) sequencing strategy. In total, 3 PCR-free libraries were generated, and Fastp v0.19.3 (fastp, RRID:SCR_016962) [34] was used to filter out low-quality reads and adapter sequences. A total of 624.149 million raw reads was generated. This produced ∼94.246 Gb (∼140× the assembled genome) of raw sequencing data, with an average cleaned read length of 148.5 bp (Supplementary Table S2).

Hi-C sequencing

The Hi-C library was prepared by Beijing Ori-Gene Science and Technology Co., Ltd (Beijing, China), with the standard procedure described as follows. A total of 700 ng of high molecular weight genomic DNA was cross-linked *in situ*, extracted, and then digested with a restriction enzyme. The sticky ends of the digested fragments were biotinylated, diluted, and then ligated to each other randomly. Biotinylated DNA fragments were enriched and sheared to a fragment size of 300–500 bp again for preparing the sequencing library, which was sequenced on a HiSeq X Ten platform (Illumina). A total of 740 million reads with

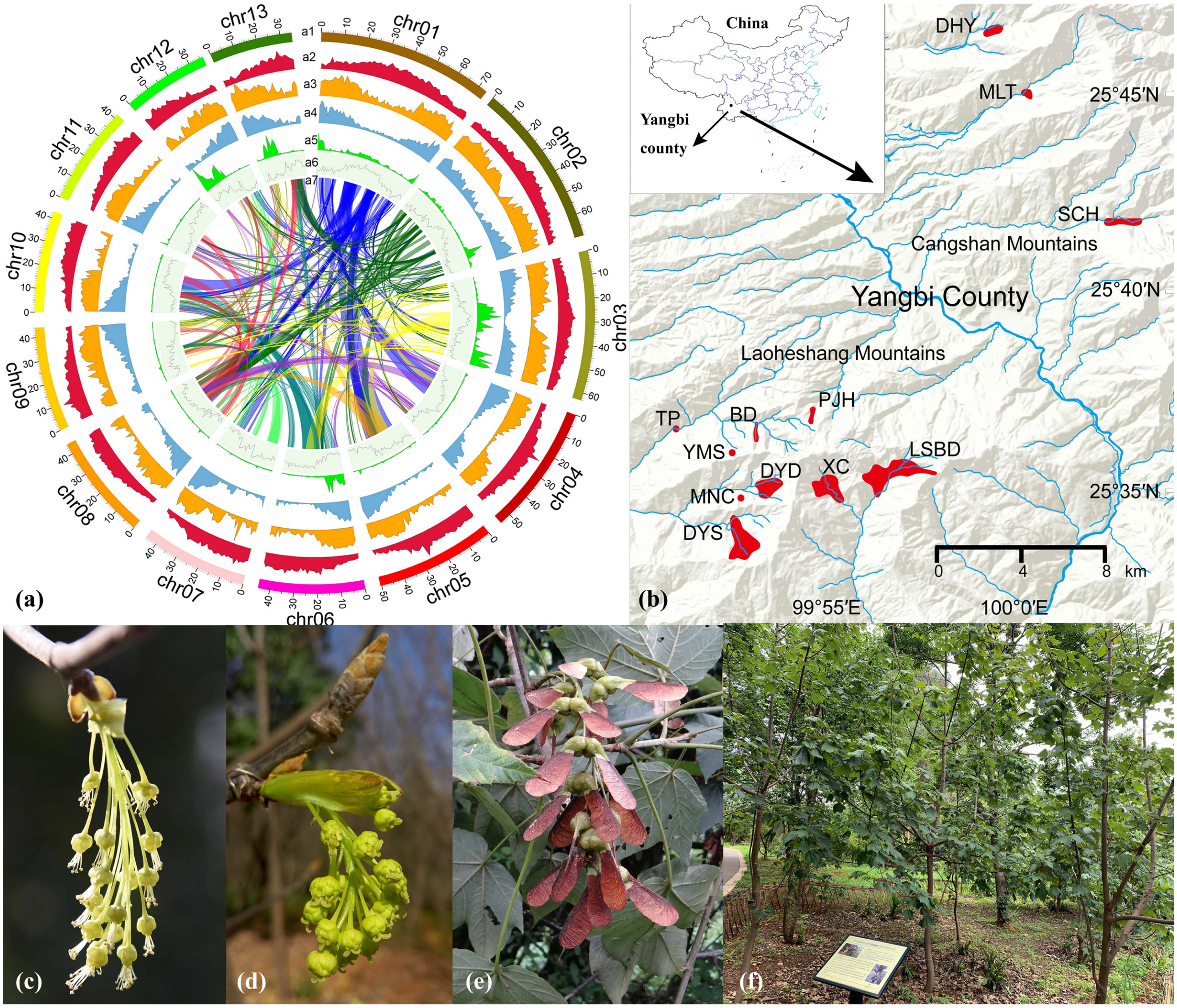

Figure 1: Images of *Acer yangbiense* chromosome assembly, distribution range, flowers, fruits, and *ex situ* conserved tree. **(a)** Genome features across 13 chromosomes. The tracks represent 13 assembled chromosomes (a1), Class I TE (LTRs, long interspersed nuclear elements, and short interspersed nuclear elements) density (a2), Class II TE (DNA and Heliron) density (a3), gene (messenger RNA) density (a4), heterozygous (single-nucleotide polymorphisms, insertions, and deletions) density (a5), GC content (a6), and genome rearrangement events of colinear blocks (a7). **(b)** Red-shaded regions denote distribution range of *A. yangbiense* in Yangbi county. Location codes: BD: Badahe; DHY: Dahuayuan; DYD: Diaoyudao; DYS: Dayingshan; LSBD: Luosibaidi; MLT: Malutang; MNC: Maoniuchang; PJH: Panjiahe; SCH: Sanchahe; TP: Taiping; XC: Xincun; YMS: Yangmeishu). **(c)** Staminate inflorescence. **(d)** Pistillate inflorescence. **(e)** Fruits. **(f)** *ex-situ* conserved tree.

∼110 Gb sequencing data were generated (∼170× the assembled genome) with an average read length of 149.8 bp (Supplementary Table S3). During preprocessing of the Illumina data, Fastp v0.19.3 [34] was used to remove the short reads, low-quality, and adapter sequences.

Estimation of genome size, heterozygosity, and repeat content

Three short fragment libraries were constructed by PCR-free method, and the whole-genome shotgun (WGS) short reads were generated using an Illumina HiSeq X Ten machine, which were filtered and corrected with Fastp v0.19.3 [34]. The genome size of *A. yangbiense* was estimated by the *k*-mer method [35] using sequencing data from the Illumina DNA library. First, Jellyfish v2 (Jellyfish, RRID:SCR_005491) [35] was used to count the occurrence of *k*-mers based on the processed data. Finally, gce v1.0.0 [36] was used to estimate the overall characteristics of the genome, such as genome size, repeat contents, and level of heterozygosity. In this study, 67,781,536,308 *k*-mers were generated, and the peak *k*-mer depth was 111 (Supplementary Fig. S1). The genome size was estimated to be ∼640 Mb, and repeat and heterozygosity rates were estimated to be 68.75% and 0.19%, respectively, based on *k*-mer individuals (Supplementary Table S4).

De novo assembly and chromosome construction

The *de novo* genome assembly was performed on full PacBio long reads using different assembly strategies to obtain a better genome assembly. Primary assembly v0.1 was generated from PacBio long reads by Canu v1.7 (Canu, RRID:SCR_015880) [37], assembly v0.2 by SMARTdenovo v1.0 [38], assembly v0.3 by Wtdbg v1.2.8 (WTDBG, RRID:SCR_017225) [39], assembly v0.4 after correction by Canu v1.7 [37] and SMARTdenovo v1.0 [38], as-

Table 1: *A. yangbiense* final genome assembly statistics

	Contig		Scaffold	
Characteristic	Size (bp)	Number	Size (bp)	Number
Total size	665,887,899 bp	562		280
N10	10,447,168 bp	5	73,781,861 bp	1
N50	5,479,097 bp	39	44,917,698 bp	6
N90	835,514 bp	154	36,383,401 bp	12
Maximum	17,438,070 bp		73,781,861 bp	
Minimum	7,640 bp		7,640 bp	
Mean	1,184,817 bp		2,378,171 bp	
Median	137,049 bp		50,985 bp	
Gap				282
GC content	35.96%			

Table 2: Summary of BUSCO evaluation of the gene prediction

Parameter	BUSCO groups (%)
Complete BUSCOs	1,375 (95.5)
Complete and single-copy BUSCOs	1,308 (90.8)
Complete and duplicated BUSCOs	67 (4.7)
Fragmented BUSCOs	29 (2.0)
Missing BUSCOs	36 (2.5)
Total BUSCO groups searched	1,440 (100)

sembly v0.5 after corrected by Canu v1.0 [37] and Wtdbg v1.2.8 [39], assembly v0.6 after corrected and trimmed by Canu v1.7 [37] and SMARTdenovo v1.0 [38], and assembly v0.7 after corrected and trimmed by Canu v1.7 and Wtdbg v1.2.8 [39] (Supplementary Table S5). The assembly (v0.4) from SMARTdenovo v1.0 [38] after Canu v1.7 [39] correction was chosen as the optimal assembly for further polishing and scaffolding. In this selected primary assembly (v0.4), the assembled genome size was 666 Mb distributed across 880 contigs with N50 of 2.3 Mb, L50 of 84, and maximum contig length of 11.9 Mb (Supplementary Table S5). The draft assembly was first polished with Pilon v1.22 (Pilon, RRID:SCR_014731) [40] based on the high-quality Illumina sequencing reads and then piped into the Hi-C assembly workflow. Clean Hi-C reads were mapped to the draft assembly with Juicer (Juicer, RRID:SCR_017226) [41], and then a candidate chromosome-length assembly was generated automatically using the 3d-DNA pipeline to correct mis-joins, order, orient, and anchor contigs from the draft assembly [42]. Manual review and refinement of the candidate assembly was performed in Juicebox Assembly Tools (JBAT) [43] for quality control and interactive correction. To reduce the influence of interactions of chromosomes and to further improve the chromosome-scale assembly, each chromosome was re-scaffolded with 3d-DNA [42] separately, and then manually refined with Juicebox [44]. With the modified 3d-DNA and JBAT workflow, 13 chromosomes (646,206,981 bp, ∼97.04%) were anchored with only 265 contigs (18,721,930 bp) unplaced. Finally, after gap filling with LR_GapCloser v1.1 (GapCloser, RRID:SCR_015026) [45] (based on PacBio long reads, running for 2 rounds), Pilon v1.22 (Pilon, RRID:SCR_014731) was used to polish the assembly (based on Illumina reads, running for 5 rounds), and Redundans v0.13 [46] was used to remove putative haplotigs, to obtain the final genome assembly (v1.1) (Supplementary Table S5). In this final genome assembly v1.1 (Supplementary Table S5), we achieved an assembled genome size of 666 Mb characterized by 562 contigs and 280 scaffolds (with contig N50 of 5.5 Mb and scaffold N50 of 45 Mb) (Table 1 and Supplementary Table S5).

Assessment of genome assembly

We evaluated the level of genome completeness of the final genome assembly using BUSCO (RRID:SCR_015008) [47] and the LTR Assembly Index (LAI) [48]. BUSCO analysis showed that 95.5% (90.8% complete and single-copy genes and 4.7% complete and duplicated genes) and 2.2% of the 1,440 expected embryophytic genes were identified as complete and fragmented genes, respectively (Table 2). In addition, a relatively high LAI score = 12.21 (categorized as reference level when $10 \leq \text{LAI} \leq 20$) showed that the assembly yielded high sequence continuity [48], agreeing with the BUSCO completeness. (Supplementary Table S5). The overall mapping rate of transcriptome data was 95.0% by HiSat2 v2.1.3 (HISAT2, RRID:SCR_015530) [49], showing good completeness of the assembly. The mapping of the whole Illumina short reads by BWA v0.7.17-r1188 (BWA, RRID:SCR_010910) [50] was 99.4%, which means almost all sequencing data were represented (covering 98.4% of the total genome length, among which, 97.9% with a coverage depth ≥5×, 97.7% with a coverage depth ≥10×, 97.4% with a coverage depth ≥20×, showing high coverage). Meanwhile, mapping of PacBio reads and bases by minimap2 v2.11-r797 [51] was 76.5% and 94.3%, respectively (covering 99.98% of the total length of genome, among which, 99.9% with a coverage depth ≥5×, 99.8% with a coverage depth ≥10×, and 99.4% with a coverage depth ≥20×). Both coverage rates of Illumina sequencing and PacBio sequencing were consistent and relatively high. The coverage depth distribution of the whole genome, as well as both gene regions of single-copy and duplicated BUSCO core genes, was plotted. The duplicated genes had the same depth distribution as the single-copy genes, indicating that the duplicated genes were not derived from unmerged haplotigs and thus there was almost no redundancy in the assembly (Supplementary Fig. S2). SAMtools (SAMtools, RRID:SCR_005227) [52] was used to detect variant sites. The heterozygosity rate was calculated by heterozygosity sites, and the error rate of single bases was calculated by homozygosity sites. The heterozygosity rate was ∼0.097%, while the error rate was ∼0.0037%. A guanine-cytosine (GC) depth analysis was conducted to assess potential contamination during sequencing and the coverage of the assembly, revealing that the genome had a mean GC content of 35.96% with no obvious GC bias (Supplementary Fig. S3). We searched all sequences of the genome assembly against the NCBI non-redundant nucleotide database with BLASTN to assess contamination, and the results suggested no potential contamination. When the Hi-C data were mapped to the final genome assembly using Juicer [41], the cluster results showed that there were 13 unambiguous chromosome scaffolds with no obvious chromosome assembly error (Supplementary Fig. S4).

DNA repeats annotation

To *de novo* identify and classify repeat families in the genome assembly, the software package RepeatModeler v1.0.8 (RepeatModeler, RRID:SCR_015027) [53] was used with 2 complementary computational methods for *de novo* identifying repeats within the genome: RECON v1.08 and RepeatScout v1.0.5 (RepeatScout, RRID:SCR_014653) [54]. Then, using the output data file from RepeatModeler as a custom repeat library, RepeatMasker v4.0.7 (RepeatMasker, RRID:SCR_012954) [55] was used to screen for repeats within the assembled genome. In summary, repeat sequences were estimated to account for 68.0% (452.81 Mb) of the *A. yangbiense* assembly, among which 17.32% were uncharacterized repeats. Long terminal repeats (LTRs) were dominant (250.98 Mb, 37.7%), with Copia (179.64 Mb) and Gypsy (66.18 Mb), the most abundant subtypes, representing 26.98% and 9.94% of the genome assembly, respectively. The results of repeat annotations are summarized in Supplementary Table S6.

Transcriptome assembly

Total RNA was extracted from the stem, roots, fruits, buds, and leaves using the Trizol reagent (Sangon Biotech, Shanghai) according to the manufacturer's instructions (Invitrogen). RNA quality was assessed on a Nanodrop-2000 spectrophotometer. The PE RNA sequencing libraries were prepared using the NEBNEXT Ultra RNA Library Prep Kit for Illumina, and 150 bp PE sequencing was performed on an Illumina HiSeq X Ten platform. A total of 252.03 million raw reads were generated (Supplementary Table S7). Using HiSat2 v2.1.0 [49], raw reads from RNA sequencing were aligned to the genome assembly. Then reference genome-guided transcriptome assemblies were constructed with StringTie v1.3.5 (StringTie, RRID:SCR_016323) [56] and Trinity v2.0.6 (Trinity, RRID:SCR_013048) [57], respectively. *De novo* assembly was generated using Trinity. After that, transcriptome assemblies were combined and further refined with CD-HIT v4.6 (CD-HIT, RRID:SCR_007105) [58]. In the end, a 138.40 Mb transcriptome with 82,766 unique transcripts was obtained as RNA sequencing evidence in genome annotation. The summary is provided in Supplementary Table S8.

Genome annotation

The MAKER2 genome annotation pipeline [59] was used to predict protein-coding genes. After the repetitive sequences were masked, AUGUSTUS v3.3.1 (AUGUSTUS, RRID:SCR_008417) [60] was used for *ab initio* gene prediction with model training based on 1,248 single-copy orthologs, which were predicted by BUSCO [47] from the genome assembly. Then, for evidence-based gene prediction, transcripts from RNA sequencing were aligned to the repeat-masked reference genome assembly with BLASTN (BLASTN, RRID:SCR_001598) and TBLASTX (TBLASTX, RRID:SCR_011823) from BLAST v2.2.28+ (NCBI BLAST, RRID:SCR_004870) [61]; protein sequences from *Arabidopsis thaliana* and *Dimocarpus longan* were aligned to the repeat-masked reference genome assembly with BLASTX (BLASTX, RRID:SCR_001653). After optimization with Exonerate v2.2.0 (Exonerate, RRID:SCR_016088) [62], MAKER package v2.31.9 (MAKER, RRID:SCR_005309) [59] was used to prepare gene model predictions. AED (Annotation Edit Distance) scores were generated for each of the predicted genes as part of the MAKER pipeline, in order to assess the quality of gene prediction. Non-coding RNAs in the genome were identified by searching from the Rfam database [63]. Gene sets were integrated into a non-redundant gene annotation, and its completeness was checked using BUSCO (the 1,440 single-copy orthologs from the embryophyta_odb9 database) [47].

From the assembled genome of *A. yangbiense* a total of 30,418 genes were annotated. Besides, 28,320 protein-coding genes were acquired; 25,572 of these had an AED $<$ 0.5 and a mean of 5.36 exons per gene. The mean lengths of gene region, transcript, and coding DNA sequence were 3,880, 1,455, and 1,308 bp, respectively (Supplementary Table S9). With regard to noncoding RNA, 734 noncoding RNA, 248 ribosomal RNA, and 1,116 transfer RNA sequences were identified by Rfam (Rfam, RRID:SCR_007891), RNAMMER (RNAmmer, RRID:SCR_017075) [64], and tRNAScan-SE (tRNAscan-SE, RRID:SCR_010835) [65], respectively.

Gene function annotation was assigned on the basis of sequence and domain conservation. For assignment based on sequence conservation, a BLAT (E-value threshold of 1e−5) (BLAT, RRID:SCR_011919) [66] search of the peptide sequences from several protein databases was performed, such as Swiss-Prot [67, 68], TrEMBL [67, 69], NR [70], Pfam [71], and eggnog [72]. For assignment based on domain conservation, InterProScan (InterProScan, RRID:SCR_005829) [73] was used to examine motifs and domains by matching against public databases, such as ProDom [74], PRINTS [75], Pfam, SMART [76], PANTHER [77], and PROSITE [78]. The highest proportions of annotation were 92.60% (database NR) by BLAT and 92.31% (database PANTHER) by InterProScan, and the unannotated proportions were 7.30% and 1.74%, respectively (Supplementary Table S10).

Identification of orthologous genes and phylogenetic tree construction

OrthoMCL v2.0.9 (Ortholog Groups of Protein Sequences, RRID:SCR_007839) [79] was used to identify orthologous and paralogous gene clusters in the assembled genomes of *A. yangbiense* and 14 related plant species (Supplementary Table S11), including *Arabidopsis thaliana* [80], *Theobroma cacao* [81], *Citrus grandis* [82], *Populus trichocarpa* [83], *Eucalyptus grandis* [84], *Vitis vinifera* [85, 86], *Coffea canephora* [87], *Beta vulgaris* [88], *Carica papaya* [89], *D. longan* [90], *Fragaria vesca* [91], *Medicago truncatula* [92], *Sclerocarya birrea* [93], and *Oryza sativa* [94]. Recommended settings were used for all-against-all BLASTP comparisons (BLAST+ v2.3.056) [61] and OrthoMCL [95] analysis.

A total of 29,892 OrthoMCL families including 379,261 genes were built on the basis of effective database sizes of all-vs-all BLASTP with an E-value of 10^{-5} and a Markov chain clustering default inflation parameter. In addition, 542 gene families with 1,793 genes were identified to be specific to the *A. yangbiense* genome when compared with the other 14 genomes (Supplementary Table S12). Furthermore, *A. yangbiense* and *D. longan* had the largest number of shared gene families (12,505) among the studied plants, supporting the closer relative relationships in the same family of Sapindaceae compared with other plant species (phylogeny of the angiosperms, APG IV) ([96] accessed 22 January 2019).

Phylogenetic analysis was performed using 854 orthologous protein-coding single-copy genes among the 15 genomes found by OrthoMCL [95]. These were then aligned with MUSCLE v3.8.31 (MUSCLE, RRID:SCR_011812) [97]. A maximum likelihood phylogenetic tree was then generated using the concatenated amino acid sequences in PhyML v3.0 with the default parameter (LG Model) [98]. The divergence time was estimated with r8s v1.81 [99] and calibrated against the divergence timing of Monocotyledoneae and Eudicotyledoneae (synchronously 135–130 million years ago [Mya]), of Pentapetalae (126–121 Mya), and Rosidae (123–115 Mya) [100]. The time-calibrated tree was further analysed together with these shared orthologous gene families among 15 plants by CAFE v4.0 [101], to detect expansion, contraction, and rapid evolution of those observed gene families.

The phylogenetic analysis identified the closest relationship of *A. yangbiense* to *D. longan*, with the divergence time between them estimated at ∼31.11 Mya (Fig. 2a). Moreover the close relationship among Sapindaceae, Anacardiaceae (*S. birrea*), and Rutaceae (*C. grandis*) was confirmed, supporting the placement of the 3 families within the order of Sapindales in APG IV (Fig. 2a). Using CAFE v4.0 [101], a total of 1,169 gene families were de-

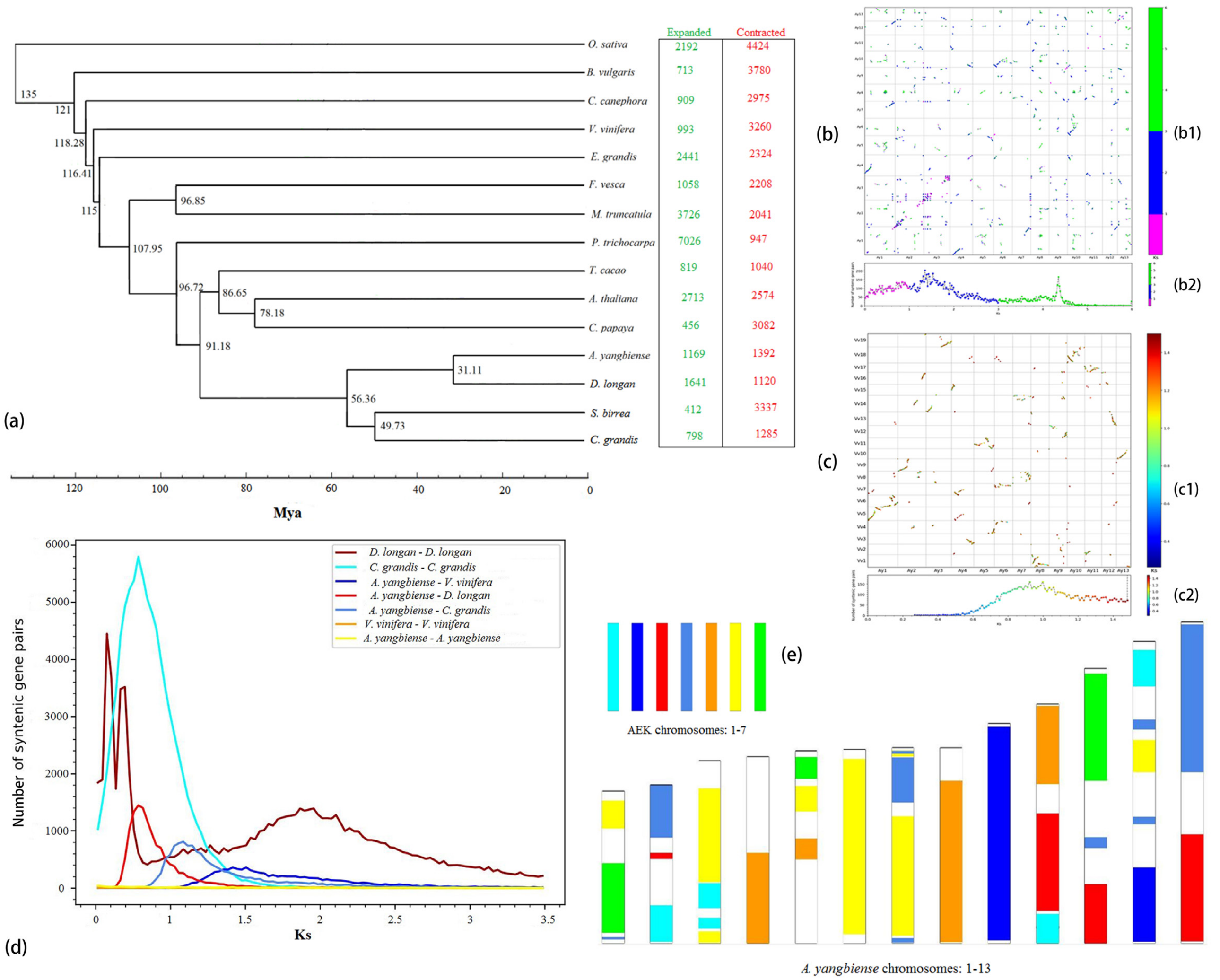

Figure 2: Genome evolution analysis of *A. yangbiense*. **(a)** Phylogenetic tree, divergence time, and profiles of gene families that underwent expansion or contraction. (b) Dot plots of syntenic blocks (b1) and corresponding Ks distribution histogram (b2) within *A. yangbiense*. (c) Dot plots of syntenic blocks (c1) and corresponding Ks distribution histogram (c2) between *A. yangbiense* and grape genome. **(d)** Synonymous substitution rate (Ks) distributions of syntenic blocks for *A. yangbiense* paralogs and orthologs with other eudicots are represented. **(e)** Comparison with ancestral eudicot karyotype (AEK) chromosomes reveals synteny. The syntenic AEK blocks are painted onto *A. yangbiense* chromosomes.

tected that have expanded, while 1,392 gene families were found to have contracted in *A. yangbiense*. The expanded gene families were enriched for 209 significant ($q < 0.05$) GO terms of 3 categories, i.e., BP (Biological Process), CC (Cellular Component), and MF (Molecular Function) (Supplementary Table S13), and 5 KEGG pathways (Supplementary Table S14) significant at $q < 0.05$. Alternatively the contracted gene families were enriched for 334 GO terms of the aforementioned 3 categories (Supplementary Table S15) and 14 KEGG pathways (Supplementary Table S16) involving several aspects of secondary metabolism, at $q < 0.05$. Additionally, functional enrichment analysis of rapidly evolving gene families reveals 218 significant GO terms (Supplementary Table S17) and 17 KEGG pathways (Supplementary Table S18), both at $q < 0.05$.

Genome evolution by synteny analysis

We performed synteny analysis of orthologous and paralogous genes previously identified by OrthoMCL [95] from *A. yangbiense* genomes, using MCScanX with default parameters, requiring ≥5 gene pairs per syntenic block [102]. The resulting dot plots were additionally used to assess characteristics of syntenic blocks by comparison within and between genomes (grape).

The Ks value was calculated to determine possible events of whole-genome duplication (WGD) and/or other duplications such as transposable elements (TE). First, protein sequences of those homologous colinear genes from *A. yangbiense* vs grape identified by MCScanX [102] were aligned against each other with MUSCLE (MUSCLE, RRID:SCR_011812) [97] to achieve the conserved protein sequences of each species, which were then converted into the corresponding codon alignments implemented in PAL2NAL [103]. Finally, Ks values were calculated by KaKs_Calculator [104] with YN model [105]. Based on the genome construction of the most recent ancestor of flowering plants, referred to as the ancestral eudicot karyotype (AEK) by Murat et al. [106], we compared the maple genome to AEK and then painted the syntenic AEK blocks onto *A. yangbiense* chromosomes.

A total of 999 colinear gene pairs on 139 colinear blocks were inferred within the *A. yangbiense* genome. There were 10,144 colinear gene pairs from 452 colinear blocks detected between *A. yangbiense* and grape (Supplementary Table S19). Dot plots of longer syntenic blocks between *A. yangbiense* and grape revealed a nearly 1:1 orthology ratio, indicating a similar evolution history to grape without undergoing a WGD event after the core eudicot common hexaploidization [107]. Synonymous substitution rate (Ks) distributions of syntenic blocks for *A. yangbiense* paralogs and orthologs with other eudicots also support the hypothesis of no recent WGD event (Fig. 2b–d). However, other than WGD, TE duplications might occur as the existence of short syntenic blocks in *A. yangbiense* (Fig. 2b). Furthermore, the genome painter image obtained by painting the syntenic AEK blocks onto *A. yangbiense* chromosomes illustrates that chromosomes 4, 6, 8, and 9 nearly exclusively contain the ancestral eudicot chromosome 2, 6, 5 without existence of inter-chromosomal segments (Fig. 2e). Such conserved gene content and order on these chromosomes in *A. yangbiense* could be due to the merged ancestral chromosome structures (e.g., multiple telomeres and centromeres on 1 chromosome) suppressing recombination and/or successive rearrangement, as was simultaneously inferred from the genome of *E. grandis* [84]. Last, we note that the genome of *A. yangbiense* has the potential to replace grape as the reference genome for studying recent WGD and chromosome evolution, especially for species within and/or close relatives to the order of Sapindales, due to high quality of genome assembly and no recent WGD, as well as less recombination of chromosomes in *A. yangbiense*.

Conclusion

We have presented a *de novo* genome assembly of *A. yangbiense* using a combination of PacBio (SMRT), Illumina HiSeq X, and Hi-C approaches and achieved a high-quality sequence assembly. The *A. yangbiense* genome that we have sequenced, assembled, and annotated here is the first genome for the genus *Acer* and the family Aceraceae. This critically threatened species genome will facilitate the genome assembly and resequencing of additional species. It will be an essential resource for further investigations of the demography, adaptability, and conservation genetics of this endangered species. Likewise, the novel genome data generated in the present study will provide a valuable resource for studying WGD and chromosome evolution particularly in the Sapindales.

Availability of supporting data and materials

The genome assembly, annotations, and other supporting data are available via the *GigaScience* database GigaDB [108]. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession VAHF00000000. The version described in this paper is version VAHF01000000. The raw sequence data have been deposited in the Sequence Read Archive under NCBI BioProject ID PRJNA524417.

Additional files

Figure S1. Frequency distribution of the 17-mer graph analysis used to estimate the size of the *A. yangbiense* genome.
Figure S2. Length distribution of PacBio subreads. Assessment of the distribution of genome reads (left) and BUSCO core region (right) coverage depth through PacBio-SMRT (lower) and Illumina sequencing data (upper).
Figure S3. Coverage depth of PacBio and Illumina sequencing data under different GCs. Assessment of the distribution of GC content and sequencing depth by PacBio-SMRT (left) and Illumina (right) under different GCs.
Figure S4. Hi-C map of final assembly of chromosomes. The distribution of links among chromosomes is exhibited by heat map based on HiCplotter. The heat map colors ranging from light yellow to dark red indicate the frequency of Hi-C interaction links from low to high (0–10).
Table S1. WGS-PacBio sequencing statistics
Table S2. WGS Illumina sequencing statistics
Table S3. Hi-C sequencing statistics
Table S4. *k*-mer survey statistics
Table S5. Statistics of all assemblies
Table S6. Repeat annotations of the *Acer yangbiense* genome assembly
Table S7. Summary of Illumina RNA sequencing data
Table S8. Summary of the transcriptome assemblies
Table S9. Gene annotation statistics of the *A. yangbiense* assembly
Table S10. Functional annotation of predicted genes in *A. yangbiense* genome
Table S11. Basic information with regard to genomes of 15 plants that were used for gene family analysis and phylogenetic tree construction
Table S12. Summary of the gene family analyses. Unique groups and genes, single-copy and duplicated groups and genes are summarized for the 15 plant genomes
Table S13. GO enrichment of expanded gene families. (A) “Category” is the Gene Ontology (GO) term ID; (B) “P value” is the over-represented *P*-value indicating that the observed frequency of a given term among analysed genes is equal to the expected frequency based on the null distribution; i.e., lower *P*-values indicate stronger evidence for overrepresentation; (C) “Q value” is the Benjamini-Hochberg adjusted *P*-value; (D) “numEPInCat” is the number of expanded gene families in the corresponding GO category; (E) “numInCat” is the number of detected gene families in the corresponding GO category; (F) “Term” is the GO term; (G) “Ontology” indicates which ontology the term comes from. Significant biological significance is at $q < 0.05$
Table S14. KEGG enrichment of expanded gene families. (A) “KO category” is the KEGG Orthology (KO) category ID; (B) “P value & Q value” have the same meaning as in Supplemental Table S13 (B) and (C); (D) “numEPInCat” is the number of expanded gene families in the corresponding KO category; (E) “numInCat” is the number of detected gene families in the corresponding KO category; (F) “Pathway” is the KEGG pathway; (G) “Class” indicates which KEGG class the pathway comes from. Significant biological significance is at $q < 0.05$
Table S15. GO enrichment of contracted gene families
Table S16. KEGG enrichment of contracted gene families
Table S17. GO enrichment of rapidly evolved gene families
Table S18. KEGG enrichment of rapidly evolved gene families
Table S19. Summary of colinear analysis within and between species

Abbreviations

AED: annotation edit distance; AEK: ancestral eudicot karyotype; APG: Angiosperm Phylogeny Group; BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; BWA: Burrows-Wheeler Aligner; Gb: gigabase pairs; GC: guanine-cytosine; gce: genome character estimation; GO: Gene Ontology; kb: kilobase pairs; KBG: Kunming Botanical Garden; KEGG: Kyoto Encyclopedia of Genes and Genomes; LAI: LTR Assembly Index; LTR: long terminal repeat; Mb: megabase pairs; Mya: million years ago; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; PE: paired end; PSESP: plant species with extremely small populations; SMRT: Single-Molecule Real-Time; TE: transposable element; WGD: whole-genome duplication; WGS: whole-genome shotgun.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by NSFC (National Natural Science Foundation of China)-Yunnan Joint Fund (Grant No. U1302262), Science and Technology Basic Resources Investigation Program of China (Grant No. 2017FY100100), National Key R&D Program of China (Grant No. 2017YFC0505200), Yunnan Science and Technology Talents and Platform Program for Key Laboratory Construction (Grant No. 2018DG004), Yunnan Science and Technology Innovation Team Program for PSESP (Plant Species with Extremely Small Populations) Conservation and Utilization (Grant No. 2019HC015), and STS Program of the Chinese Academy of Sciences "Full Cover Conservation Project of Native Plants in Southwestern China" (KFJ-3W-No1).

Authors' contributions

W.B.S. and Y.P.M. designed the study; L.D.T. and G.F.L. collected and prepared the materials; R.G.Z and Q.Z.Y conducted the experiments and data analysis. J.Y., H.M.W., and Y.P.M. wrote the manuscript; H.P., L.D.T., Z.L.D., H.J.G., and W.B.S. revised the manuscript. All authors read and approved the final draft.

Acknowledgements

The authors thank Dr. Alex Twyford from the University of Edinburgh for his help and comments from a technical perspective.

References

1. Ogata K. A systematic study of the genus *Acer*. Bull Tokyo Univ For 1967;**63**:89–206.
2. de Jong PC. Flowering and Sex Expression in Acer L. A Biosystematic Study, Wageningen, Netherlands. Veenman; 1976.
3. Wu ZY, Raven PH, Hong DY. Flora of China. Vol. 11: Oxalidaceae Through Aceraceae. St. Louis: Science Press, Beijing, and Missouri Botanical Garden Press; 2008.
4. van Gelderen DM, De Jong PC, Oterdoom HJ. Maples of the World. Timber Press; 1994.
5. Weakley A. Flora of the Southern and Mid-Atlantic States. University of North Carolina Herbarium; 2010.
6. Harris A, Chen Y, Olsen RT, et al. On merging *Acer* sections *Rubra* and *Hyptiocarpa*: Molecular and morphological evidence. PhytoKeys 2017;**86**:9–42.
7. Harris A, Frawley E, Wen J. The utility of single-copy nuclear genes for phylogenetic resolution of *Acer* and *Dipteronia* (Acereae, Sapindaceae). Ann Bot Fennici 2017;**54**(4–6): 209–22.
8. Wen J. Evolution of eastern Asian and eastern North American disjunct distributions in flowering plants. Annu Rev Ecol Syst 1999;**30**(1):421–55.
9. Wolfe JA, Tanai T. Systematics, phylogeny, and distribution of *Acer* (maples) in the Cenozoic of western North America. J Fac Sci Hokkaido Univ Ser 4 Geol Mineral 1987;**22**(1): 1–246.
10. Renner SS, Beenken L, Grimm GW, et al. The evolution of dioecy, heterodichogamy, and labile sex expression in *Acer*. Evolution 2007;**61**(11):2701–19.
11. Renner SS, Grimm GW, Schneeweiss GM, et al. Rooting and dating maples (*Acer*) with an uncorrelated-rates molecular clock: Implications for North American/Asian disjunctions. Syst Biol 2008;**57**(5):795–808.
12. Huang SF, Ricklefs RE, Raven PH. Phylogeny and historical biogeography of *Acer* I-Study history of the infrageneric classification. Taiwania 2002;**47**(3):203–18.
13. Xu TZ, Chen YS, Piet CDJ, et al. Flora of China. Vol. 11: Aceraceae. St. Louis: Science Press, Beijing, and Missouri Botanical Garden Press, 2008.
14. Contreras RN, Shearer K. Genome Size, ploidy, and base composition of wild and cultivated *Acer*. J Am Soc Hortic Sci 2018;**143**(6):470–85.
15. Bi W, Gao Y, Shen J, et al. Traditional uses, phytochemistry, and pharmacology of the genus *Acer* (maple): A review. J Ethnopharmacol 2016;**189**:31–60.
16. Ball DW. The chemical composition of maple syrup. J Chem Educ 2007;**84**(10):1647–50.
17. Gonzalez-Sarrias A, Li L, Seeram NP. Anticancer effects of maple syrup phenolics and extracts on proliferation, apoptosis, and cell cycle arrest of human colon cells. J Funct Foods 2012;**4**(1):185–96.
18. Perkins TD, van den Berg AK. Maple syrup-production, composition, chemistry, and sensory characteristics. Adv Food Nutr Res 2009;**56**:101–43.
19. Legault J, Girard-Lalancette K, Grenon C, et al. Antioxidant activity, inhibition of nitric oxide overproduction, and in vitro antiproliferative effect of maple sap and syrup from *Acer saccharum*. J Med Food 2010;**13**(2):460–68.
20. Park KH, Yoon KH, Yin J, et al. Antioxidative and anti-inflammatory activities of galloyl derivatives and antidiabetic activities of *Acer ginnala*. Evid Based Complement Alternat Med 2017;**2017**:1–8.
21. Chen YS, Yang QE, Zhu GH. *Acer yangbiense* (Aceraceae), a new species from Yunnan, China. Novon 2003;**13**(3):296–99.
22. Zhao LL. Genetic Diversity of the Critically Endangered Yanbi maple, *Acer yangbiense* (Aceraceae). MA Thesis. Graduate School of Chinese Academy of Sciences; 2011.
23. Gibbs D, Chen YS. The Red List of Maples. Botanic Gardens Conservation International; 2009. https://www.bgci.org/resources/bgci-tools-and-resources/the-red-list-of-maples/. Accessed 1 October 2018.
24. Qin HN, Yang Y, Dong SY, et al. Threatened species list of China's higher plants. Biodivers Sci 2017;**25**(7):696–744.
25. Yang J, Zhao LL, Yang JB, et al. Genetic diversity and conservation evaluation of a critically endangered endemic maple, *Acer yangbiense*, analyzed using microsatellite markers. Biochem Syst Ecol 2015;**60**:193–98.
26. Zhao LL, Sun WB, Yang JB. Development and characterization of microsatellite markers in the critically endangered species *Acer yangbiense* (Aceraceae). Am J Bot 2011;**98**(9):e247–e49.
27. Tao LD. Population Ecology Studies of two PSESP Plants, and the Reproductive Biology and SSR Primers of *Acer yang-*

biense. MA Thesis. University of Chinese Academy of Sciences; 2018.
28. Ma YP, Chen G, Grumbine RE, et al. Conserving plant species with extremely small populations (PSESP) in China. Biodivers Conserv 2013;**22**(3):803–09.
29. Sun WB, Yin Q. *Ex-situ* conserving the Yangbi maple *Acer yangbiense* in China. *Oryx* 2009;42:461-62.
30. Sun WB, Ma YP, Blackmore S. How a new conservation action concept has accelerated plant conservation in China. Trends Plant Sci 2019;**24**(1):4–6.
31. Sun WB. Words from the guest editor-in-chief. Plant Divers 2016;**38**(5):207–08.
32. Silva-Junior OB, Grattapaglia D, Novaes E, et al. Genome assembly of the pink ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly valued, ecologically keystone neotropical timber forest tree. GigaScience 2018;**7**(1):gix125.
33. Doyle JJ. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bull 1987;**19**:11–15.
34. Chen S, Zhou Y, Chen Y, et al. fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;**34**(17):i884–90.
35. Marcais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 2011;**27**(6):764–70.
36. Liu B, Shi Y, Yuan J, et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv 2013 . https://arxiv.org/abs/1308.2012.
37. 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**(5):722–36.
38. Jue R. smartdenovo: Ultra-fast de novo assembler using long noisy reads. https://github.com/ruanjue/smartdenovo. Accessed 1 October 2018.
39. Jue R. Redbean: A fuzzy Bruijn graph (FBG) approach to long noisy reads assembly. https://github.com/ruanjue/wtdbg-1.2.8. Accessed 1 October 2018.
40. Walker BJ, Abeel T, Shea T, et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;**9**(11):e112963.
41. Durand NC, Shamim MS, Machol I, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 2016;**3**(1):95–98.
42. Dudchenko O, Batra SS, Omer AD, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 2017;**356**(6333):92–95.
43. Dudchenko O, Shamim MS, Batra S, et al. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv 2018, doi:10.1101/254797.
44. Durand NC, Robinson JT, Shamim MS, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 2016;**3**(1):99–101.
45. Xu GC, Xu TJ, Zhu R, et al. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. Gigascience 2018;**8**(1):giy157.
46. Pryszcz LP, Gabaldon T. Redundans: An assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 2016;**44**(12):e113.
47. Simao FA, Waterhouse RM, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210–12.
48. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res 2018;**46**(21):e126–e26.
49. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nat Methods 2015;**12**(4):357–60.
50. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013. https://arxiv.org/abs/1303.3997.
51. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics 2018;**34**(18):3094–100.
52. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;**25**(16):2078–79.
53. Smit A, Hubley R. RepeatModeler Open-1.0. http://www.repeatmasker.org/RepeatModeler/. Accessed 1 October 2018.
54. Price AL, Jones NC, Pevzner PA. De novo identifcation of repeat families in large genomes. Bioinformatics 2005;**21**(Suppl 1):i351–8.
55. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 (2013-2015). http://repeatmasker.org. Accessed 1 October 2018.
56. Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;**33**(3):290–95.
57. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;**29**(7):644–52.
58. Fu L, Niu B, Zhu Z, et al. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;**28**(23):3150–52.
59. Cantarel BL, Korf I, Robb SM, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 2008;**18**(1):188–96.
60. Boratyn GM, Schaffer AA, Agarwala R, et al. Domain enhanced lookup time accelerated BLAST. Biol Direct 2012;**7**(1):12.
61. Stanke M, Diekhans M, Baertsch R, et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 2008;**24**(5):637–44.
62. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 2005;**6**(1):31.
63. Griffiths-Jones S, Bateman A, Marshall M, et al. Rfam: An RNA family database. Nucleic Acids Res 2003;**31**(1):439–41.
64. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 1997;**25**(5):955–64.
65. Lagesen K, Hallin P, Rodland EA, et al. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 2007;**35**(9):3100–08.
66. Kent WJ. BLAT-the BLAST-like alignment tool. Genome Res 2002;**12**(4):656–64.
67. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;**28**(1):45–48.
68. ExPASy Bioinformatics Resources Portal. http://www.expasy.ch/sprot. Accessed 1 December 2017.
69. UniProt. http://www.ebi.ac.uk/uniprot. Accessed 1 December 2017.
70. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov. Accessed 1 October 2018.
71. Punta M, Coggill PC, Eberhardt RY, et al. The Pfam protein families database. Nucleic Acids Res 2011;**40**(D1):D290–301.
72. Jensen LJ, Julien P, Kuhn M, et al. eggNOG: Automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 2007;**36**(suppl 1):D250–D54.

73. Jones P, Binns D, Chang HY, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics 2014;**30**(9):1236–40.
74. Corpet F, Gouzy J, Kahn D. Recent improvements of the ProDom database of protein domain families. Nucleic Acids Res 1999;**27**(1):263–67.
75. Attwood TK, Croning MD, Flower DR, et al. PRINTS-S: The database formerly known as PRINTS. Nucleic Acids Res 2000;**28**(1):225–27.
76. Schultz J, Copley RR, Doerks T, et al. SMART: A web-based tool for the study of genetically mobile domains. Nucleic Acids Res 2000;**28**(1):231–34.
77. Mi H, Lazareva-Ulitsky B, Loo R, et al. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 2005;**33**(suppl 1):D284–D88.
78. Sigrist CJ, De Castro E, Cerutti L, et al. New and continuing developments at PROSITE. Nucleic Acids Res 2012;**41**(D1):D344–D47.
79. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 2003;**13**(9):2178–89.
80. Cheng CY, Krishnakumar V, Chan AP, et al. Araport11: A complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J 2017;**89**(4):789–804.
81. Motamayor JC, Mockaitis K, Schmutz J, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol 2013;**14**(6):r53.
82. Wang X, Xu Y, Zhang S, et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. Nat Genet 2017;**49**(5):765–72.
83. Tuskan GA, Difazio S, Jansson S, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 2006;**313**(5793):1596–604.
84. Myburg AA, Grattapaglia D, Tuskan GA, et al. The genome of *Eucalyptus grandis*. Nature 2014;**510**(7505):356–62.
85. Jaillon O, Aury JM, Noel B, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 2007;**449**(7161):463–67.
86. Canaguier A, Grimplet J, Di Gaspero G, et al. A new version of the grapevine reference genome assembly (12X. v2) and of its annotation (VCost. v3). Genom Data 2017;**14**:56–62.
87. Denoeud F, Carretero-Paulet L, Dereeper A, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 2014;**345**(6201):1181–84.
88. Dohm JC, Minoche AE, Holtgrawe D, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature 2014;**505**(7484):546–49.
89. Ming R, Hou S, Feng Y, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 2008;**452**:991–96.
90. Lin Y, Min J, Lai R, et al. Genome-wide sequencing of longan (*Dimocarpus longan* Lour.) provides insights into molecular basis of its polyphenol-rich characteristics. Gigascience 2017;**6**(5):1–14.
91. Shulaev V, Sargent DJ, Crowhurst RN, et al. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 2011;**43**(2):109.
92. Young ND, Debelle F, Oldroyd GE, et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 2011;**480**(7378):520–24.
93. Chang Y, Liu H, Liu M, et al. The draft genomes of five agriculturally important African orphan crops. Gigascience 2019;**8**(3):giy152.
94. Ouyang S, Zhu W, Hamilton J, et al. The TIGR rice genome annotation resource: Improvements and new features. Nucleic Acids Res 2006;**35**(suppl 1):D883–D87.
95. Boetzer M, Henkel CV, Jansen HJ, et al. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2010;**27**(4):578–79.
96. Angiosperm Phylogeny Website. http://www.mobot.org/MOBOT/Research/APweb/welcome.html. Accessed 22 January 2019.
97. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;**32**(5):1792–97.
98. Guindon S, Dufayard JF, Lefort V, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol 2010;**59**(3):307–21.
99. Sanderson MJ. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 2003;**19**(2):301–02.
100. Magallon S, Gomez Acevedo S, Sanchez Reyes LL, et al. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol 2015;**207**(2):437–53.
101. De Bie T, Cristianini N, Demuth JP, et al. CAFE: A computational tool for the study of gene family evolution. Bioinformatics 2006;**22**(10):1269–71.
102. 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**(7):e49–e49.
103. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 2006;**34**(suppl 2):W609–W12.
104. Zhang Z, Li J, Zhao XQ, et al. KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 2006;**4**(4):259–63.
105. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 2000;**17**(1):32–43.
106. Murat F, Armero A, Pont C, et al. Reconstructing the genome of the most recent common ancestor of flowering plants. Nat Genet 2017;**49**(4):490–96.
107. Paterson AH, Wendel JF, Gundlach H, et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. Nature 2012;**492**(7429):423–27.
108. Yang J, Wariss HM, Tao LD, et al. Supporting data for “De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China.” GigaScience Database 2019. http://dx.doi.org/10.5524/100610.