DNA Research, 2022, **30**, 1–11
https://doi.org/10.1093/dnares/dsac051
Advance access publication 15 December 2022
Resource Article: Genomes Explored

OXFORD

Resource Article: Genomes Explored

Karyotype evolution of the Asterids insights from the first genome sequences of the family Cornaceae

Congcong Dong[1,#], Shang Wang[2,#], Han Zhang[1], Jianquan Liu[1,2,], and Minjie Li[1,*]

[1]State Key Laboratory of Herbage Improvement and Grassland Agro-ecosystems, College of Ecology, Lanzhou University, Lanzhou, China
[2]Key Laboratory of BioResource and EcoEnvironment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, China

*To whom correspondence should be addressed. Email: lmj@lzu.edu.cn (M.L.)
#Equal contributions to this work.

Abstract

Cornaceae is a core representative family in Cornales, the earliest branching lineage in the Asterids on the life tree of angiosperms. This family includes the only genus *Cornus*, a group of ~55 species. These species occur widely in Northern Hemisphere and have been used as resources for horticultural ornaments, medicinal and industrial manufacturing. However, no any genome sequences are available for this family. Here, we reported a chromosomelevel genome for *Cornus controversa*. This was generated using high-fidelity plus Hi–C sequencing, and totally ~771.80 Mb assembled sequences and 39,886 protein-coding genes were obtained. We provided evidence for a whole-genome duplication event (WGD) unique to *C. controversa*. The evolutionary features of this genome indicated that the expanded and unique genes might have contributed to response to stress, stimulus and defense. By using chromosome-level syntenic blocks shared between eight living genomes, we found high degrees of genomic diversification from the ancestral core-eudicot genome to the present-day genomes, suggesting an important role of WGD in genomic plasticity that leads to speciation and diversification. These results provide foundational insights on the evolutionary history of Cornaceae, as well as on the Asterids diversification.

Key words: Asterids, *Cornus controversa*, comparative genomics, karyotype evolution

1. Introduction

Whole-genome duplication (WGD) has played a dramatic role in angiosperm (or flowering plant) diversification.[1] The genome size and chromosome number of angiosperms fluctuate remarkably, even among close related species, spanning 1,000-fold and 50-fold, respectively.[2–4] WGD and mobile elements are two major contributors of genomic evolution. Recurring ancient WGD events caused wholescale chromosomal rearrangement (such as translocations, inversions, duplications, and deletions) and condensations (or fusions) and gene loss that have led to genomic differences in synteny and collinearity between lineages having the most recent common ancestor (MRCA).[4] In addition, mobile elements nearly reshuffle heterochromatic regions and largely break collinearity in these regions.[5] Advances in genome DNA sequencing technologies have provided opportunities to detect WGD events by identifying chromosome-like synteny blocks shared between derived extant genomes and further to infer ancestral karyotype evolution.[6,7] This is helpful in estimating the role of WGD in genomic plasticity that leads to speciation and diversification.

The Asterids is one of the major clades of angiosperms, including nearly 1/3 of flowering plants. However, little is known for its success of early diversification. Cornaceae comprising only *Cornus* L.[8] is a core representative family in Cornales, the earliest branching lineage in the Asterids on the phylogenetic tree of flowering plants.[9] There are about 55 trees or shrub dogwoods,[8] which are one of major members of the boreal and tropical forests in the Northern Hemisphere. The dogwoods underwent rapid diversification and have large-scale morphological heterogeneity, such as variables in fruits, inflorescences, and chromosomes.[10,11] This genus is horticulturally important, with many species widely cultivated for their showy blossoms and brightly coloured fruits. Additionally, some dogwoods produce special oil for industrial usage and some are used for medicines in China.[12–14] Previous molecular systematic studies recovered that the dogwoods can be classified into about 10 subgenera, corresponding to the four groups identified by inflorescence differences.[10] However, molecular barcoding markers cannot fully resolve evolutionary relationships of the dogwoods, leading to systematic controversy and uncertainty to remain debated.[8,10–12,15,16] Genome sequence unavailability of the dogwoods has complicated success to resolve their evolutionary histories and made a knowledge gap for improving the accuracy of phylogenetic inference and precision of molecular dating using genome-scale data.

In this study, we chose *C. controversa* Hemsl. with a chromosome number of $2n = 2x = 20$[17] for genome-sequencing. This deciduous canopy tree occurs frequently in evergreen broad-leaved forest and coniferous broad-leaved mixed forest at elevations of 250–2,600 m across

Received 30 July 2022; Revised 25 November 2022; Accepted 12 December 2022

China and adjacent countries of Asia.[18] It diverged from other dogwoods very early with a likely origin since the Miocene.[19] This species has been widely used for horticultural landscape, medicinal and industrial resources.[12–14,18] Based on high-fidelity (HiFi) and chromosome conformation capture (Hi–C) sequencing, we assembled a high-quality, chromosome-level genome for this species, which is firstly reported for the family Cornaceae. This genome data provided an opportunity to understand the evolution of genome structure variations and gene families of *C. controversa* by performing comparative genomic analyses. The genomic evolutionary characteristics revealed that the species-specific and significantly expanded gene families might have contributed to its widespread adaption in the Northern Hemisphere. Furthermore, ancestral karyotype reconstructions of the extant Cornales/Asterids make it possible to investigate evolutionary forces that drive the success of early diversification of the Asterids. These reconstructed MRCA genomes have identified the genomic fraction conserved across sampled extant genomes following genomic changes and can be useful to infer their evolutionary relationships and understand the degrees of genomic plasticity from the MRCA.

2. Materials and methods

2.1. Sample collection and genome sequencing

We collected fresh leaves from one individual plant of *C. controversa* that naturally distributes in Emei Mountain of Sichuan, China (E: 103.35, N: 29.56) and immediately stored them in liquid nitrogen. The total genomic DNA for genome sequencing was isolated from leaves according to a modified CTAB method.[20] The quality of the extracted DNA was assessed using the ultraviolet-visible spectrophotometer. The Illumina libraries with an insert size of 350 bp were prepared on PCR-free DNBSEQ platform (BGI, Beijing, China) and sequenced on paired-end (PE) 150-bp format system. A total of 97.49 gigabases (Gb) (coverage of ~126×) of clean PE reads were obtained. High-quality high-fidelity (HiFi) sequencing was carried out on PacBio Sequel II platform with Sage ELF libraries, sheared with 15 kilobases (kb) fragments using Megaruptor 3 (Diagenode). A total of 44.87 Gb (coverage of ~58×) HiFi clean reads were generated. Chromosome conformation capture, 3C (Hi-C) techniques were used to anchor HiFi contigs to the 10 pseudo-chromosomes according to the custom procedure.[21] The Hi-C sequencing library was constructed using NEB Next Ultra II DNA library Prep Kit (New England Biolabs, England). Fragments between 400 and 600 bp were sequenced on Illumina platform with PE 150-bp format. A total of 74.36 Gb (coverage of ~96×) of Hi–C raw reads were yielded (Supplementary Table S1).

2.2. Estimation of genome size and heterozygosity

Genome size survey for *C. controversa* was performed in GenomeScope based on the k-mer statistics[22] using the short reads data. The 33kmer frequency of Illumina short reads was used to construct k-mer depth distributions by Jellyfish.[23] A total of 76,655,126,044 k-mers followed a bimodal distribution, with a primary peak observed at a depth of 89.70 and an additional peak at nearly half of the major depth. The estimated genome size of *C. controversa* was 768.81 Mb (Table 1 and Supplementary Fig. S1), and genomic heterozygosity was estimated to be 1.09%.

2.3. *De novo* genome assembly and chromosome construction

Adaptors and low-quality reads in the raw Illumina short reads were filtered using SOAP nuke v2.1.6[24] with default parameters. The high-quality HiFi reads were used to *de novo* assemble the contigs with HIFIASM v0.15.4r347.[25] Then three rounds of contig polish were carried out in Nextpolish1.4.1.[26] To confirm the accuracy of assembly, the BWA-MEM2 v2.0pre2[27] was carried out to map the short clean reads to the contigs. We then reassembled these polished contigs to form scaffolds by using the HiC reads. In brief, BurrowsWheeler Aligner[28] (BWA) was used to obtain uniquely mapped read pairs by mapping the clean Hi–C reads to the assembled contigs. HiCPro v3.0.0[29] was then applied to collect the valid interaction pairs by comparing and filtering. 3D-DNA v180114[30] was conducted to cluster, sort, and orientate the contigs to generate a chromosome-level genome (Supplementary Fig. S2). To estimate the assembly quality and structure of this chromosomal genome, the genome was then cut into numerous 1 kb bins, and HiCPlotter (https://github.com/kcakdemir/HiCPlotter) was executed to plot the interaction heat map between any two bins using the number of HiC read pairs between those two bins. We carried out Benchmarking Universal SingleCopy Orthologs (BUSCO) v3.0.2[31] with embryophyta_odb10 to assess the integrity of the genome assembly.

2.4. Repetitive sequence annotation

To structurally annotate repetitive sequences in the *C. controversa* genome, we predicted repetitive elements with the application of RepeatModeler open2.0[32] and RepeatMasker open4.0.7.[33] RECON and RepeatScout in RepeatModeler were applied to discover repetitive elements, and then extract and classify the consensus repeat models to construct a repeat library. RepeatMasker was used to perform a homology method based on a repeat search throughout the *C. controversa* genome. Finally, the same repeat classes in the two approaches were overlapped according to the coordinates in the genome. Transposable elements (TEs) could have an impact on gene expression and function, we thus identified genes with TE insertion via overlapping the coordinates between genes and TEs

Table 1. The assembly and annotation information of the *Cornus controversa* genome

Genomic feature	Value
Estimated genome size (Mb)	768.81
Assembly size (Mb)	771.80
Number of contigs	107
Contig N50 (Mb)	62.64
GC content (%)	35.36
Genome complete BUSCO (%)	97.9
Protein complete BUSCO (%)	94.9
Pesudochromosome number	10
Sequences anchored to chromosomes (Mb)	756.59
Number of protein-coding gene	39,886
Average length of per genes (bp)	4,257.89
Number of CDS	183,344

throughout the genome. Then GO enrichment was carried out in agriGO v2.0[34] for functional annotation of these genes.

Long-terminal repeat (LTR) retrotransposons were initially predicted by using LTR Finder v1.02[35] and LTR harvest.[36] LTR_retriever was then used to filter the false LTRs by checking the structure and sequence features according to (1) target site duplications, (2) terminal motifs, and (3) LTR Pfam domains. RepeatMasker was used to annotate LTRs based on the non-redundant LTR library and the insertion time of LTRs was computed by LTR_retriever.[37]

2.5. Gene prediction and function annotation

A combination of *ab initio* prediction and protein homologous mapping was used to annotate the protein-coding genes of *C. controversa*. For *de novo* prediction, the repeatmasked genome of *C. controversa* was used to predict the protein-coding genes by performing Augustus v.3.2.3,[38] GenScan,[39] and GlimmerHMM v.3.0.4.[40] For Augustus, we fed the pipeline only with the genome sequences of *C. controversa*, and used the parameter of "--species=arabidopsis" for training set done automatically (http://bioinf.uni-greifswald.de/augustus/). And both of GenScan and GlimmerHMM were also trained on the *Arabidopsis thaliana* (L.) Heynh. genome. For protein homologous mapping, the gene models of *C. controversa* were determined by using GeMoMa v1.6.4[41] to hit the combined protein-coding genes of six species genomes including *A. thaliana*, *Camellia sinensis* (L.) Kuntze, *Nyssa sinensis* Oliv., *N. yunnanensis* W. Q. Yin ex H. N. Qin & Phengklai, *Rhododendron simsii* Planch., and *Vitis vinifera* L. (Supplementary Table S2). The consensus gene sets were generated using Evidence Modeler v1.1.1[42] by overlapping the *ab initio* and homology-mapping based results.

Functional annotation was carried out using NCBI Blast+ v 2.2.28 (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) (*E*-value ≤ 1e−5) by searching all predicted protein-coding genes of *C. controversa* against the public databases of COG,[43] KOG,[43] NCBI-NR,[44] SwissProt,[45] and TrEMBL.[45] The best hits of Blast were then used to perform prediction of gene functions. InterProScan v.5.28[46] and HMMER v3.1b2[47] were used to identify protein domains and motifs by searching the protein sequence against the databases of InterPro and Pfam. Gene Ontology (GO) terms for each gene were achieved using Blast2GO v4.1 (http://www.blast2go.org/) from InterPro or Pfam entries. Ortholog assignment and pathway mapping were performed in KASS-KEGG Automatic Annotation Server (http://www.genome.jp/tools/kaas) using BLAST (*E*-value ≤ 1e−5) method.

2.6. Gene family and phylogenetic analysis

A total of 12 species, including *C. controversa*, *Cercidiphyllum japonicum* Siebold & Zucc. ex J. J. Hoffm. & J. H. Schult. bis (available at https://bigd.big.ac.cn/?xml:lang=en), *V. vinifera* (ENA, PRJEB37020), *Rhododendron delavayi* Franch. (NCBI, PRJNA659608), *Actinidia rufa* (Siebold & Zucc.) Planch. ex Miq. (NCBI, PRJDB8483), *Camptotheca acuminata* Decne. (available at https://doi.org/10.6084/m9.figshare.12570599 and https://doi.org/10.6084/m9.figshare.12570614), *Davidia involucrata* Baill. (NCBI, PRJNA596897), *Lactuca sativa* L. (NCBI, PRJNA173551), *Apium graveolens* L. (NCBI, PRJNA593940), *Salvia splendens* Sellow ex Schult. (DDBJ/ENA/GenBank, PNBA00000000), *Coffea canephora* Pierre ex A. Froehner (NCBI, PRJEB4211), and *Catharanthus roseus* (L.) G. Don (NCBI, JQHZ00000000), were selected to identify the orthologous groups (Fig. 2a). The longest transcript for each gene was used to represent the gene. Sequence similarities were determined using All-vs.-all gene alignments in BLAST (*E*-value ≤ 1e−5). Gene family memberships were constructed using Markov Chain Clustering (MCL) in OrthoMCL v 2.0.9.[48] Single-copy orthologous groups were then extracted from OrthoMCL results, and their protein-coding sequences were aligned using MAFFT v7.4.02.[49] RAxML[50] was used to construct a maximum likelihood (ML) tree with *C. japonicum* plus *V. vinifera* as outgroup. The robustness of each node was tested by running 1,000 bootstrap analyses. We used concatenated CDS alignments to estimate species divergence time with MCMCTree in PAML v4.9.[51] The parameters were set to 10,000 burn-in, 20,000 MCMC runs and sampling frequency every 1,000 runs. Two independent runs were performed to check the convergence. Based on Bayesian relaxed molecular clock approach, we used the divergence time, 106.0–118.9 million years ago (Mya) between *C. acuminata* and *L. sativa* (retrieved from the TimeTree database, available at http://www.timetree.org/) to calibrate this phylogenetic tree.

Expansion or contraction of gene families was identified using CAFÉ v3.1[52] in the above inferred time tree of the 12 sequenced genomes, with the *p*-value set to 0.05 and the parameter λ value for each branch estimated using random searching. Functional annotation of genes that had undergone significant expansion was performed using GO enrichment in agriGO v2.0.[34]

Plant resistance (*R*) genes are a gene group that plays an important role in pathogen recognition pathways.[53] *R* genes usually contain two conserved domains, that is, nucleotide-binding site (NBS) domain and leucinerich repeat (LRR) domain. *R* genes can be further divided into three classes according to the domain types in the Nterminal region, namely CNL (CCNBSLRR), RNL (RPW8NBSLRR), and TNL (TIR-NBSLRR).[54] We thus performed the hidden Markov model (HMM) and BLAST searching to identify *R* genes in the *C. controversa*, *C. acuminata*, *D. involucrata*, and *A. thaliana* genomes. In brief, we first performed HMMSCAN in HMMER v3.1b2 to search the HMM profile of the NB-ARC domain (Pfam no. PF00931) against each protein sequence file of the four genomes. We then used BLASTp to search the protein sequences of the NB-ARC domain against the protein sequence file of each genome. After merging all hits identified from HMMSCAN and BLAST analyses and filtering the redundant ones, we obtained the *R* gene sequences of these four genomes. Furthermore, we carried out Pfam and coiled-coil (CC) analyses to detect LRR, RPW8, TIR, zf-BED, and CC domains in these *R* gene sequences. Paircoil2[55] (the threshold set of 0.025) and COILS software (http://www.ch.embnet.org/software/COILS_form.html) were used for identifying CC domain.

2.7. Demographic history

To explore whether the population size (*Ne*) expansion has contributed to the widespread distribution of *C. controversa* and *C. acuminata* compared to *D. involucrata* which has a narrow geographic distribution,[56] a Pairwise Sequentially Markovian Coalescent (PSMC) model v0.6.5-r67[57] (https://github.com/lh3/psmc) was carried out to estimate their historical dynamics in *Ne*. The Illumina short clean reads of each species were mapped to the corresponding assembled genomes to get the consensus sequences by using BWA-MEM2

v2.0pre2[27] and SAMtools v1.9.[58] PSMC analysis was performed by setting the parameters of '–N25 – t15 – r5 – p "4 + 25 × 2 + 4 + 6"', with the generation time of 15 years and a mutation rate of 3.65×10^{-8} per site per year for all species.

2.8. Whole-genome duplication analysis

To provide evidence for the lineage-specific whole-genome duplicate (WGD) events, we selected three species, including *C. controversa*, *D. involucrata*, and *V. vinifera* to find collinear blocks both between species and within each species by using MCScanX[4] with default parameters. For each block, at least five collinear gene pairs were contained. Synonymous substitution rates per gene (Ks) of each collinear gene pair were calculated using WGDI[59] with a parameter of '-ks' by YN00. We then extracted the median Ks values of each collinear block to plot the density distribution curve of Ks probability using the kernel smoothing function in MATLAB, by setting default parameters of Ks density and bandwidth. Multi-peak fitting curve was estimated using the vcftools in MATLAB, with the coefficient (R^2) set to 0.95. To validate whether the identified WGD events were shared among the families in the Cornales, we further examined the gene duplication events across the phylogeny of the above mentioned 12 species. First, according to the results of gene family clustering by OrthoMCL, we selected the clusters with more than four genes contained for the following analysis. Next, we reconstructed an independent gene tree for each cluster using IQ-TREE (http://www.iqtree.org/) by setting the parameters of '-m MFP -bb 1000 -bnni -nt AUTO -redo'. With the phylogeny of the 12 selected angiosperms used as a reference of species tree, we used NOTUNG[60] algorithm to obtain the gene trees that were consistent with the species-tree, and then to count gene duplication events for all internal nodes and terminal taxa in each tree with the parameter setting of '--threshold 50%'. Gene duplication events were determined by two measures: (i) two child branches need have genes from at least one common species; (ii) bootstrap value at each internal node was greater than 50%. The final result of gene duplication events was summarized and presented on the species tree. We finally scaled the time of WGD in the *C. controversa* genome using the time of the γ event of 115–130 Mya.[61] Additionally, we performed MCScanX to identify the duplication gene types of all genes in the genome of *C. controversa*, and then extracted the duplication genes resulted from the recent WGD for the GO enrichment in agriGO v2.0.

2.9. Ancestral karyotype reconstruction

To understand how the genomes have evolved within Superasterids/Asterids/Cornales, we reconstructed their ancestral karyotypes by selecting eight extant species genomes, including *C. acuminata*, *C. controversa*, *D. involucrata*, *A. rufa*, *L. sativa*, *S. splendens*, and *Amaranthus hybridus* L., and with *C. japonicum* from the family Cercidiphyllum used as outgroup. The operation procedure was generally according to WGDI,[59] and the specific workflow was consisted of three main steps: (i) polyploidy inference using Dotplot, Collinearity extraction, and Ks distributions; (ii) hierarchical inference of genomic homology; (iii) subgenomic and ancestral genome reconstruction and other evolutionary scenarios. In brief, we first processed the genomic and annotation data of the eight species according to the WGDI format requirements and then carried out the protein sequence alignment using the MMseqs2 (https://github.com/soedinglab/MMseqs2), a blastplike software method. Afterward, the homologous chromosome fragment identification was performed using a dynamic programming algorithm based on gene synteny, and the Ks values of homologous gene pairs were also calculated. The resulting syntenic blocks and Ks values were used to infer the occurrence of genomic duplications. Since the haploid chromosome number of *C. japonicum* (x = 19) was close to that of the ancestral core eudicots (x = 21), we thus mapped the *C. japonicum* genome to the other seven species to infer the karyotype evolution of their MRCA.

3. Results

In this study, we combined three different sequencing strategies to *de novo* produce a high-quality, chromosome-level genome of *C. controversa*, including ~97.49 Gb Illumina PE short reads, ~44.87 Gb HiFi reads, and ~74.36 Gb Hi–C reads (Supplementary Table S1). The estimated genome size of *C. controversa* was 768.81 Mb (Table 1 and Supplementary Fig. S1). The *de novo* assembled *C. controversa* genome was 771.80 Mb in length and contained 107 contigs. Based on ~96× Hi–C reads, 55 out of 107 contigs (~756.59 Mb, 98.03% of the original assembly) were anchored to 10 pseudo-chromosomes (Fig. 1 and Table 1). The contig N50 length of the assembled genome was 62.64 Mb, which is much higher than many recently reported genomes using PacBio sequencing (e.g. Chen Y., Li M.J., and Kang M.H.[62–64]). The overall GC-content of the *C. controversa* genome was 35.36% (Table 1 and Supplementary Table S3). The maximum and minimum chromosome lengths were 129.13 Mb and 48.29 Mb, respectively (Supplementary Table S4 and Supplementary Fig. S2). The assembled quality of the *C. controversa* genome was assessed according to the following aspects: (i) more than 98.56% Illumina short reads could be properly mapped to the genome assembly (Supplementary Table S3); (ii) a total of 1,580 (97.90%) orthologs and 1,481 (91.80%) single-copy orthologs specific in *C. controversa* were found to be complete in BUSCO assessments (Table 1 and Supplementary Table S5). These results indicated that a highly contiguous, complete, and accurate genome of *C. controversa* was reported.

A total of 39,886 protein-coding genes accounting for 94.90% in complete BUSCO (Table 1 and Supplementary Table S6) were identified in the *C. controversa* genome, with the average gene length of 4,257.89 bp, the average CDS length of 1,103.77 bp, the average exon length of 239.95 bp, and the average intron length of 877.16 bp, respectively (Supplementary Table S7). The mean exon counts per gene were 4.60 in the *C. controversa* genome. These results suggested a conversed evolution of protein-coding genes in the *C. controversa* genome, when compared with the closely related species (*C. acuminata* and *D. involucrata*) genomes (Supplementary Table S7). However, when compared to *C. acuminata* and *D. involucrata*, a rather shorter gene length was observed in *C. controversa* mainly due to fewer exons and introns in each gene, plus a shorter average intron length per gene (Supplementary Table S7). In addition, all these protein-coding genes were aligned with public annotated protein database for homology: COG (28.30%), GO (28.32%), KEGG (17.99%), KOG (43.00%), SwissProt (53.80%), TrEMBL (84.01%), and NCBINR (79.90%; Supplementary Table

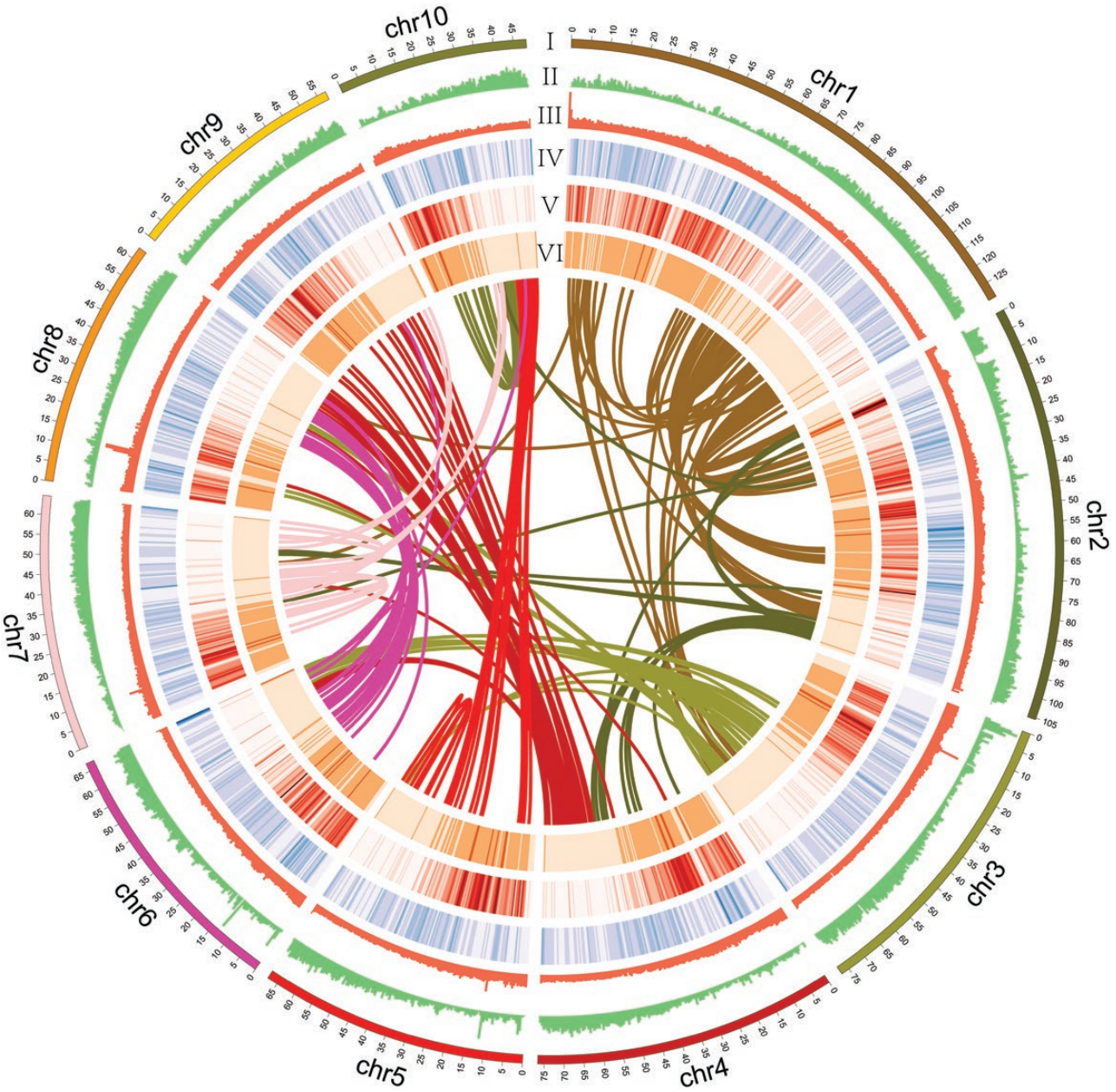

Figure 1. Overview of the chromosome features of *Cornus controversa* in 500 kb sliding windows for each pseudochromosome. Tracks from outside to inside respectively correspond to (i) pseudochromosome number; (ii) gene density and (iii) GC content. Links in the core connect syntenous genes of *C. controversa*; (iv) *Copia* density (low to high, from undertint to dark blue), (v) *Gypsy* density (low to high, from undertint to dark red), and (vi) total repeat density (low to high, from undertint to dark red). Links in the core connect syntenous genes of *C. controversa*.

S8). Functional annotation using InterProScan and HMMER showed that 95.93% of these protein-coding genes possessed conserved protein domains. Finally, 80.16% of the protein-coding genes (31,973) were functionally annotated.

Transposable elements (TE) were found to constitute 52.23% of the *C. controversa* genome, in which LTRs were most abundant, accounting for 26.62% (201,404,289 bp) of the assembled genome (Supplementary Table S9). The time of LTR insertion in the *C. controversa* genome was estimated to begin at 5.30 million years ago (Mya), and reached to a peak at 1.99 Mya, which both were much older than those in the genomes of *C. acuminata* (began to burst at 4.00 Mya and reach to a peak at 0.24 Mya) and *D. involucrata* (began to burst at 4.00 Mya and reach to a peak at 0.70 Mya; Fig. 2b). The GO enrichment analysis of genes with TE insertion were mainly correlated with macromolecular methylation/modification, stress response and DNA repair (Supplementary Figs. S3–S5).

3.1. Phylogenetic analysis and gene family evolution

We selected 12 species from the Rosanae to infer phylogenetic relationships with *V. vinifera* and *C. japonicum* used as outgroup. A total of 39,886 *C. controversa* genes were clustered into 17,948 gene families, of which, 7,176 (44.83%) gene families were shared with the other 12 species and 1,494 ones were unique to *C. controversa* (Supplementary Table S10). The GO enrichment analysis revealed that those unique genes were functionally involved in responding to

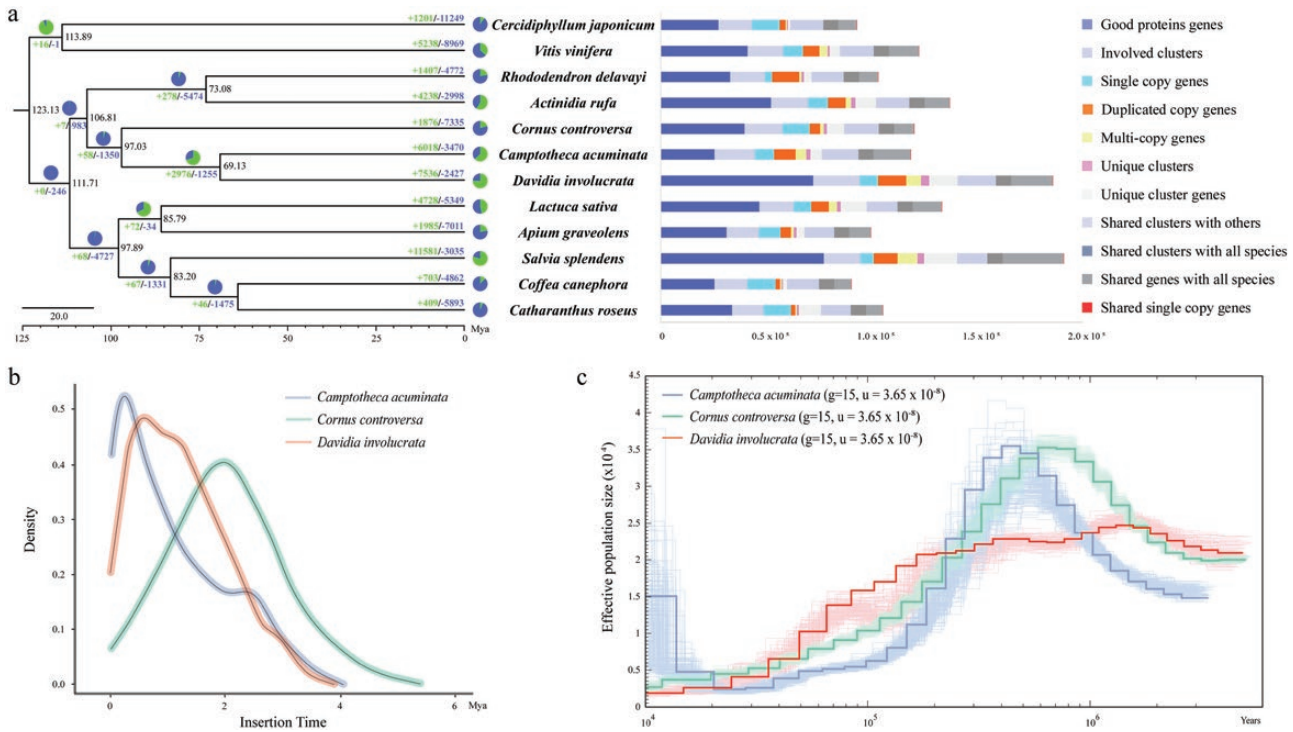

Figure 2. Genomic evolutionary history of *Cornus controversa*. (a) Phylogenetic tree for *C. controversa* and other 11 angiosperms. All branch bootstrap values are 100. The pie chart and the corresponding positive and negative numbers represent the expanding and contracting gene families; the corresponding proportions among the total changes are shown using the same colours in the pie charts. The estimated divergence time (million years ago, Mya) is indicated at each node. The histograms represent the gene investigation of this phylogenetic tree based on software of orthoMCL. (b) Insertion time of long-terminal repeat (LTR) retrotransposons in *C. controversa*, *Davidia involucrata*, and *Camptotheca acuminata*. Mya, million years ago. (c) Demographic histories for *C. controversa*, *D. involucrata*, and *C. acuminata* inferred using pairwise sequentially markovian coalescent (PSMC) model.

stress and stimulus and defense response (Supplementary Figs. S6 and S7).

We selected 70 single-copy genes across the sampled 12 species to infer a phylogenetic tree using RAxML.[50] We found that *C. controversa* was sister to the Nyssaceaes species, that is, *C. acuminata* and *D. involucrata*. Then, all of them constituted the Cornales and were sister to *A. rufa* + *R. delavayi* (Fig. 2a). The divergence time estimation indicated that *C. controversa* diverged from the Nyssaceae clade at 97.03 Mya. The divergence of the Cornales—Ericales and the Lamiids—Campanulids occurred at 111.71 Mya (Fig. 2a and Supplementary Fig. S8).

The gain and loss of gene families play critical roles in plant evolution. Our computational analysis of 43,056 gene families among the 12 selected species found a total of 1,876 and 7,335 gene families to be expanded and contracted in *C. controversa*, respectively (Fig. 2a). Based on the MCScanX result, we further classified the 6,326 genes of 1,876 expanded gene families into five classes according to their origins and locations: singleton (0.71%, 45), dispersed (32.03%, 2,026), proximal (7.73%, 489), tandem (36.93%, 2,336), and WGD/segmental (22.61%, 1,430) (Supplementary Table S11). This likely suggested that WGD and tandem duplication had played an important role in contributing to the expansion of gene families in *C. controversa*. GO enrichment analysis revealed that the expanded genes in *C. controversa* were mainly related to molecular and cellular functions (Supplementary Fig. S7 and Supplementary Tables S12–S13).

3.2. Demographic history

We employed the pairwise sequentially Markovian coalescent (PSMC) method to infer dynamic histories of the effective population size (*Ne*) of *C. controversa*, *C. acuminata*, and *D. involucrata*. Our PSMC result showed that *C. controversa* and *D. involucrata* had a similar *Ne* at 3 Mya, which was slightly larger than that of *C. acuminata* (Fig. 2c). Since then, *C. acuminata* and *C. controversa* experienced a rapid increase of *Ne* during 0.5–3.0 Mya; meanwhile, a relatively stable *Ne* was observed in *D. involucrata*. The first sharp decline of *Ne* in *C. controversa* and *C. acuminata* occurred from 0.5 to 0.02 Mya, while a relatively gradual decline of *Ne* occurred in *D. involucrata* at the same time. Following the end of Last Glacial Maximum (LGM), the *Ne* of *D. involucrata* and *C. controversa* continued to decline with the *Ne* of *C. controversa* larger than that of *D. involucrata*. However, *C. acuminata* recovered its *Ne* during 0.02–0.01 Mya, making it much larger than those of *D. involucrata* and *C. controversa* (Fig. 2c).

3.3. Whole-genome duplication analysis

To investigate the whole-genome duplication (WGD) history in the *C. controversa* genome, we established the distribution of synonymous substitutions per synonymous site (Ks) using syntenic paralogs. We found *C. controversa*, *C. acuminata*, and *D. involucrata* experienced independent WGD after the shared γ (whole genome triplication) event in all core-eudicots.[61] The species-specific WGD in *C. controversa* occurred at the peak of Ks of ~0.478, which was slightly earlier

than the divergence between *C. controversa* and *D. involucrata* (Ks peak: ~0.440), but slightly later than *C. controversa* and *C. acuminata* (Ks peak: ~0.524) (Fig. 3a). When performing intergenomic analysis between *V. vinifera*, *C. controversa, C. acuminata*, and *D. involucrata*, we observed a clear syntenic depth ratio of 1:2 of the large collinear blocks within the intergenomic analysis of *V. vinifera* and *C. controversa*, and a syntenic depth ratio of 2:2 was identified in the intergenomic analyses of Cornales (Supplementary Figs. S9–S11 and Fig. 3b and c). To validate whether these three species shared a recent WGD, we performed gene duplication analysis. A total of 16,591 clusters (35.5% of all orthologous clusters) were obtained among the 12 selected angiosperms, and 15,364 gene trees were retained through topology reconciliation between gene tree and species tree. Our results showed that 2,542 genes in *C. controversa*, 9,015 genes in *C. acuminata* and 10,716 genes in *D. involucrata* experienced duplication, respectively. However, only 460 genes experienced duplication at the nearest common ancestor node of *C. controversa* and *C. acuminata* + *D. involucrata*, and 3,532 genes at the nearest common ancestor node of *C. acuminata* and *D. involucrata* (Supplementary Fig. S12). As a consequence, the ratios of the Cornales (0.18 of 460/2,542, 0.05 of 460/9,015 and 0.04 of 460/10,716) were largely less than those of the Nyssaceae family (0.39 of 3,532/9,015 and 0.33 of 3,532/10,716), suggesting an independent WGD event for *C. controversa*, and a recently shared WGD between *C. acuminata* and *D. involucrata*.[61] This conclusion was also supported by the Ks distribution[62] and syntenic depth ratios (Fig. 3). When scaling the time of the WGD event in *C. controversa* by the core-eudicot γ event of 115–130 Mya,[61] we found *C. controversa* encountered polyploidization approximately at 45.38–51.30 Mya. Additionally, we identified 4,620 duplicated genes resulted from the recent WGD in the *C. controversa* genome, and they showed functions associated with biosynthetic and metabolic processes (Supplementary Fig. S13).

3.4. Evolution of resistance-related (*R*) genes

We totally identified 275 *R* genes in the *C. controversa* genome, including 65 CNL genes, 1 RNL gene and 47 TNL genes (Supplementary Table S14). The counts of *R* genes in three closely related species were similar, with 270 and 316 *R* genes respectively identified in *C. acuminata* and *D. involucrata*, but the counts of *R* genes in these three species were highly higher than that in *A. thaliana* (Supplementary Table S14). Furthermore, *D. involucrata* had nine TNL genes, which was much lower than those of *C. controversa* (47), *C. acuminata* (23), and *A. thaliana* (75) (Supplementary Table S14), while *D. involucrata* had much more CNL genes than the other threes (Supplementary Table S14). We used *R* genes having complete domains from these four species to construct a phylogenetic tree. The resulting topology showed that all *R* genes were classified into CNL, RNL, and TNL groups (Supplementary Fig. S14).

3.5. Ancestral karyotypes of the modern Asterids

Ancestral karyotype reconstruction is greatly helpful in estimating the role of polyploidization in genomic plasticity that leads to speciation and diversification. We thus did that by refining the ancestral core-eudicot karyotype (AEK) as a post-γ AEK with 21 protochromosomes.[7] A comparison of the *C. japonicum* genome and the post-γ AEK revealed a chromosome fusion event, which derived *C. japonicum* with 19 chromosomes.[65] We inferred the ancestral Superasterids karyotype (ASK) by defining the synteny and collinearity of ancestral genes among the genomes of *C. japonicum* and *A. hybridus* (Supplementary Fig. S15). The inferred ASK was refined with 19 protochromosomes and two chromosome fusion events were arisen from the post-γ AEK to the ASK (Fig. 4). We then reconstructed the ancestral Cornales karyotype (ACK) by performing a comparison of the extant Cornales genomes and the *C. japonicum* genome (Supplementary Figs. S16–S18). The reconstructed ACK was refined with 14 protochromosomes and seven chromosome condensations from the post-γ AEK to the ACK were identified (Fig. 4). On the basis of this analysis, the ancestral Asterids karyotype (AAK) was also refined with 14 protochromosomes and no clear chromosome condensation occurred from the AAK to the ACK (Fig. 4). However, three chromosome fusions from the ACK to the ancestral Nyssaceae karyotype (ANK) were observed and the ANK was refined with 12 protochromosomes (Fig. 4). The degrees of genomic arrangements and chromosomal fusion changes were significantly higher in *C. controversa* than those in the Nyssaceae lineage (both *C. acuminata* and *D. involucrata*) (Fig. 4). To validate the precise nature of AAK, pairwise comparisons of the genomes between *C. japonicum* and the other Asterids including Ericales (*A. rufa*), Lamiales (*S. splendens*), and Asterales (*L. sativa*) were also performed. The AAK with 14 protochromosomes was robustly supported by full deconvolution of the completely conserved synteny and paralogy between the pairwise genomes of *C. japonicum* and the other three Asterids families (Supplementary Figs. S19–S21). Based on these analyses (Fig. 4), the lineage-specific polyploidization events were also observed. An independent WGD event was respectively identified in *A. rufa*, *Amaranthus hybridus*, *C. controversa*, and the family Nyssaceae. Additionally, whole-genome triplication (WGT) in *L. sativa* and three independent WGD events in *S. splendens* were also detected (Fig. 4). These results agreed with previous reports. The evolutionary relationships among the sampled eight species were also inferred based on the results of protein sequence alignments and homology dotplots, which were consistent with that inferred from the orthologous genes (Fig. 2a).

4 Discussion

In addition to the multipurpose functions in ecology and economy, dogwoods also have a key position in the evolution of the Asterids, while the lack of a chromosome-level reference genome is a serious constraint to boost relevant research. In this study, we report the genome sequences for one dogwood plant, *C. controversa*, with a combination of Illumina, HiFi, and Hi-C sequencing technologies. The genome assembly assessments suggest a highly-quality chromosome genome toward continuity and gene annotation. This is the first report of the genome sequences for the family Cornaceae. The genomic sequences are critical to investigate genome evolution of *C. controversa* by performing comparative genomic analyses, as well as to study karyotype evolution of the Asterids by comparing chromosome-level synteny blocks.

The assembly genome *C. controversa* is 771.80 Mb in length, and more than 98.54% of the assembled sequences were anchored on 10 pseudochromosomes. Compared to other two closely related species, the genome size of *C. controversa* is ~2.2fold greater than that of *C. acuminata*

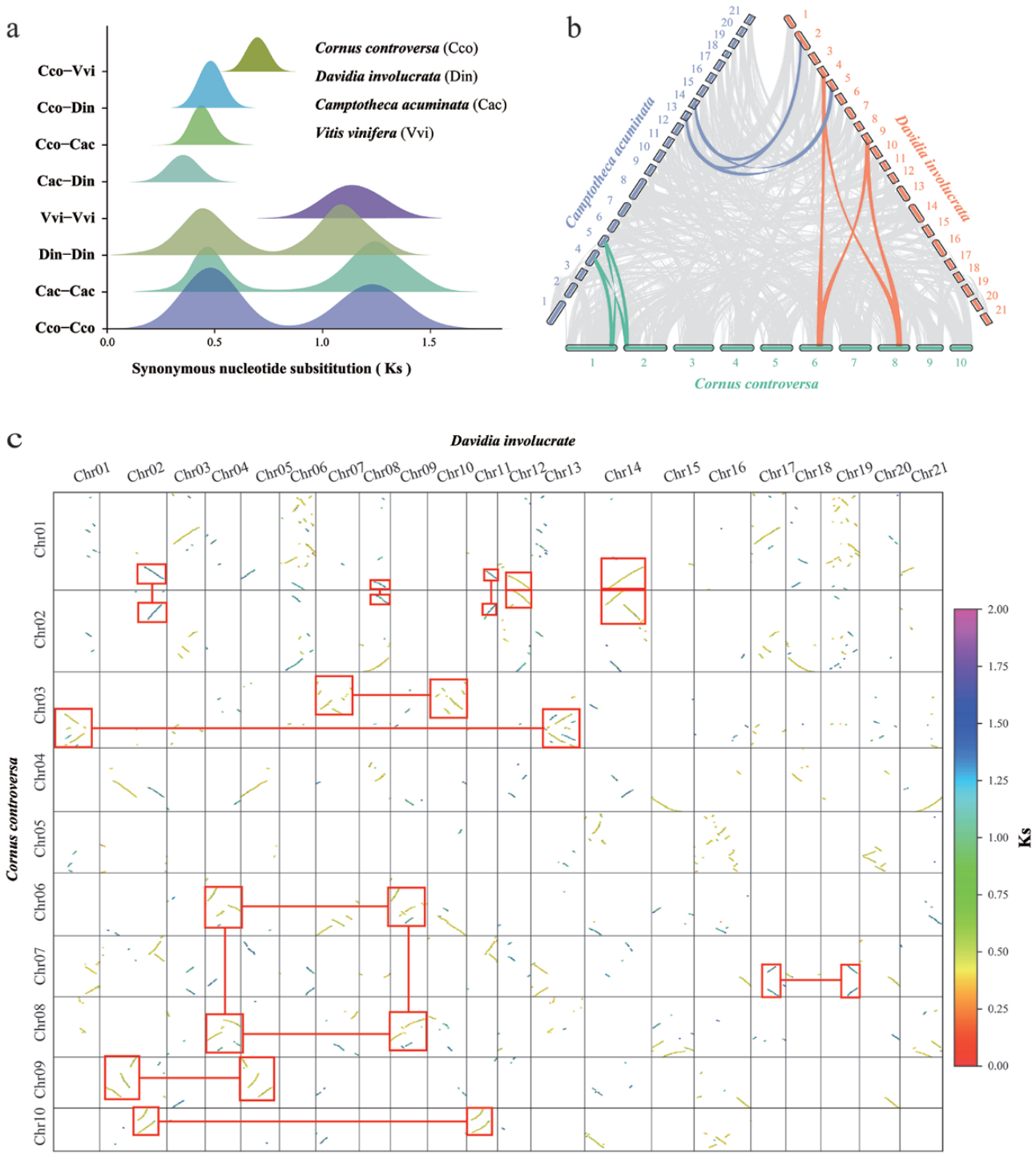

Figure 3. Polyploidization analysis of the *Cornus controversa* genome. (a) Distribution of synonymous nucleotide substitutions (Ks) among *C. controversa* and other three species. (b) Collinear relationships between *C. controversa*, *Davidia involucrata*, and *Camptotheca acuminata*. All banded lines in the background indicate syntenic blocks between the genomes spanning more than 15 genes; some of the 2:2 syntenic blocks are highlighted with bright lines. (c) Syntenic block dotplot between *C. controversa* and *D. involucrata*. A syntenic depth ratio of 2:2 is shown using squares and solid lines.

(414.95 Mb),[64] but is smaller than that of *D. involucrata* (1,169 Mb)[62] (Supplementary Table S7). WGD events and TE bursts are two most common aspects for the genome size differences.[66] We found *C. controversa* and *C. acuminata*—*D. involucrata*[62] experienced one independent recent WGD except the shared γ event of all core-eudicots, respectively. The gene number in *C. controversa* (39,886) is close to that of *D. involucrata* (42,554), but greatly higher than that of *C. acuminata* (27,940). Given the similar average length per gene between *C. acuminata* and *D. involucrata*, but rather shorter in *C. controversa* (Supplementary Table S7), it is conclusive that WGD cat not fully explained their genome size differences. The genome assemblies and annotations found that the TE sequences occupy 66.00% (771.59 Mb) of the genome

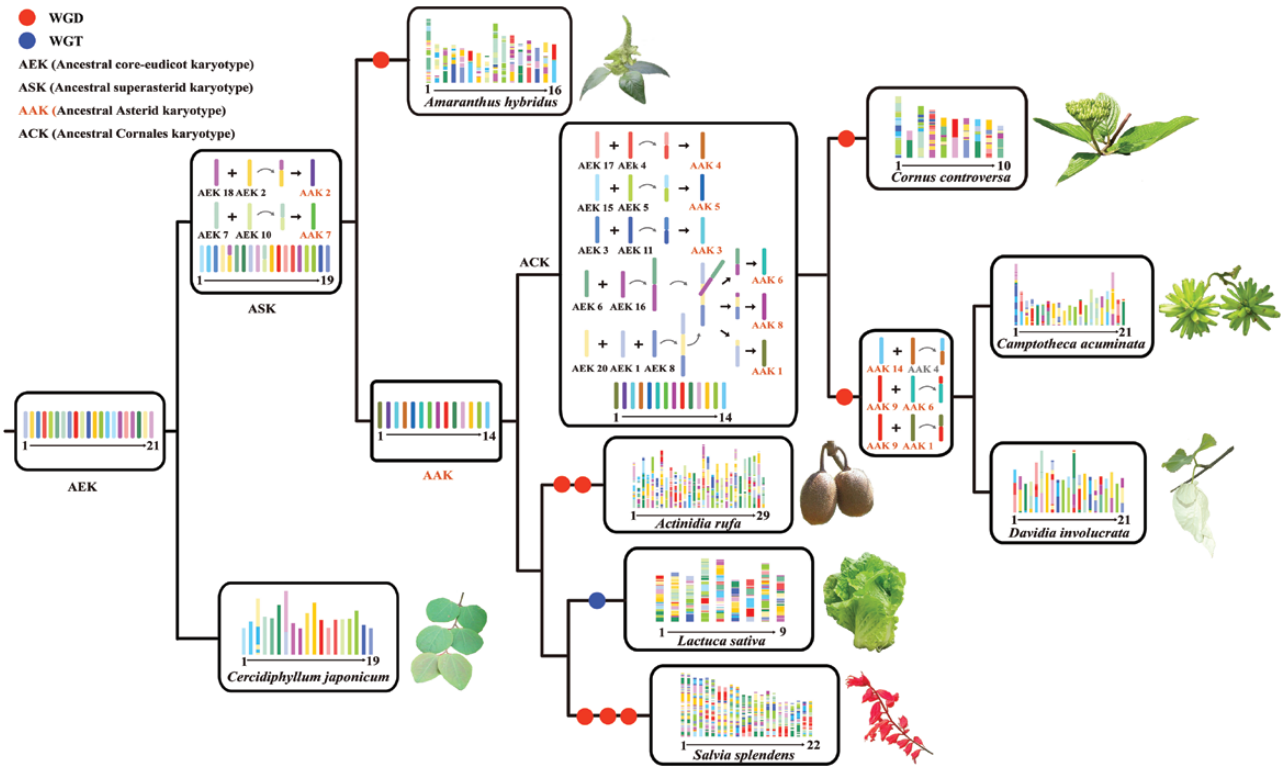

Figure 4. Ancestral karyotypes of Superasterids/Asterdis/Cornales from the ancestral core-eudicot karyotype (AEK). The known AEK is illustrated with 21-colour code. WGD, whole genome duplication. WGT, whole genome triplication. The chromosome fusions and fissions that shaped the modern karyotypes from the AEK are shown on the tree branches. The protochromosome number of most recent common ancestor and chromosome number of extant sampled species are indicated using Roman numbers from left to right along with an arrow. ASK: the ancestral Superasterids karyotype; AAK: the ancestral Asterids karyotype; ACK, the ancestral Cornales karyotype; ANK: the ancestral Nyssaceae karyotype.

sequences in *D. involucrata*,[62] which is much larger than those in *C. controversa* (390.32 Mb, 52.23%) and *C. acuminata* (156.37 Mb, 37.68%). In summary, the genome size expansion in *D. involucrata* and *C. controversa* compared to *C. acuminata* likely stem from the TE bursts, especially the LTR expansion which has a profound impact on variations of gene length and number.

Our PSMC-based demographic histories of the three Cornales species recovered contrasting demographic histories for the endangered relict *D. involucrata* and the widespread *C. acuminata* and *C. controversa*. Effective population size (*Ne*) of *D. involucrata* failed to recover at the end of the last glaciation maximum (LGM, 26.5–19 kaBP)[67] may have contributed much to its population collapse.[62] Despite a smaller *Ne* of *C. controversa* at the end of the LGM, its widespread distribution may suggest substantial genetic variability and enough adaptive potential for recovery. The genome evolutionary features of *C. controversa* also provide support for this hypothesis when considering the expanded and unique genes that are functionally correlated to response to stress, stimulus, and defense. However, this hypothesis need further evidence from population genetics.

In this study, we inferred the ancestral Superasterids karyotype (ASK, 19 protochromosomes), the ancestral Asterids karyotype (AAK, 14 protochromosomes), and the ancestral Cornales karyotype (ACK, 14 protochromosomes), providing an overview of genome evolution from the ancestral core-eudicot (AEK, a post-γ AEK with 21 protochromosomes) to the eight present-day genomes. Although no shared paleopolyploidization events were found in the ASK, AAK, and ACK, the relatively recent WGD events have independently happened in *C. controversa*, Nyssaceae, *Actinidia rufa*, *Lactuca sativa*, *Salvia splendens*, and *Amaranthus hybridus* (Fig. 4). The early diversification of Asterids mostly during the last 117–70 million years of evolution (Fig. 2a), in the context of the era of Cretaceous–Paleocene (K–Pg) mass species extinction,[68] might have benefited from the genomic plasticity inherited through the shared polyploidization events of eudicots (γ).[7,69] The post-polyploidization compartments from the AEK likely acting as a reservoir of genomic plasticity for speciation and diversification have been specialized at chromosomal structures, and therefore present diverse degrees of synteny and collinearity among different species (Fig. 4).[7,69] Furthermore, the genomic plasticity inherited from the relatively recent polyploidization may have derived lineage-specific diversification, in the context of increasingly environmental changes (such as cooling and aridification) during the following 60 million years of evolution.[69] Our inferred evolutionary framework of ASK, AAK, and ACK, with a comparison of them with the present-day extant species, advances our understanding and provides fundamental insights on the genomic plasticity to speciation and diversification. These inferred ancestral karyotypes recovered that polyploidization has caused different degrees of genomic changes at the species, subgenome, and even gene levels (Fig. 4), as previous report.[7,70,71] Overall, our inferred ancestral karyotypes and evolutionary frameworks are useful in understanding the success of the Asterids diversification.

Funding

The authors thank the Supercomputing Center of Lanzhou University for computation support. This work was supported by the National Key Research and Development Program of China (2021YFD2200202), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB31010300), and the National Natural Science Foundation of China (31901074 and 31590821).

Authorship

J.Q.L. and M.J.L. devised the study and supervised all parts of the project. C.C.D. performed genome assembly, annotation, the phylogenetic analysis; C.C.D. and H.Z. performed whole-genome duplication analysis; S.W and C.C.D. carried out ancestral karyotype reconstruction. M.J.L. and C.C.D. drafted the paper and J.Q.L. helped in revision.

Conflict of Interest

All authors did not have any conflict of interest.

Data Availability

All raw reads sequences (short reads, long reads, and HiC reads) and assembly genome file of *C. controversa* used in this study have been submitted to GenBank of NCBI (https://www.ncbi.nlm.nih.gov/) under the BioProject accession number PRJNA778449. The accession numbers of short reads and Hi-C reads are SRR22199186 and SRR22199183, and the accession number of long reads is SRR22199184 and SRR22199185. The accession number of assembly genome is JANPWI000000000. The annotation file can be available from https://figshare.com/s/3bc489c07312dc699196.

References

1. Mandáková, T. and Lysak, M.A. 2018, Postpolyploid diploidization and diversification through dysploid changes, *Curr. Opin Plant Biol.*, **42**, 55–65, doi:10.1016/j.pbi.2018.03.001.
2. Bennett, M.D. and Smith, J.B. 1991, Nuclear dna amounts in angiosperms, *R. Soc.*, **334**, 309–45, doi:10.1098/rstb.1991.0120.
3. Bowers, J.E., Chapman, B.A., Rong, J. and Paterson, A.H. 2003, Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events, *Nature*, **422**, 433–8, doi:10.1038/nature01521.
4. Tang, H., Bowers, J.E., Wang, X., et al. 2008, Synteny and collinearity in plant genomes, *Science*, **320**, 486–8, doi:10.1126/science.1153917.
5. Bowers, J., Arias, M.A., Asher, R., et al. 2005, Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses, *Proc. Natl. Acad. Sci. USA*, **102**, 13206–11, doi:10.1073/pnas.0502365102.
6. Salse, J. 2016, Ancestors of modern plant crops, *Curr. Opin Plant Biol.*, **30**, 134–42, doi:10.1016/j.pbi.2016.02.005.
7. Murat, F., Armero, A., Pont, C., Klopp, C. and Salse, J. 2017, Reconstructing the genome of the most recent common ancestor of flowering plants, *Nat. Genet.*, **49**, 490–6.
8. Xiang, Q.Y. and Thomas, D.T. 2008, Tracking character evolution and biogeographic history through time in Cornaceae—does choice of methods matter, *J. Syst. Evol.*, **46**, 349–74, doi:10.3724/SP.J.1002.2008.08056.
9. Byng, J.W., Chase, M.W., Briggs, B., et al. 2016, An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV, *Bot. J. Linn. Soc.*, **181**, 1–20, doi:10.1111/boj.12385.
10. Xiang, Q.Y., Thomas, D.T., Zhang, W.H., et al. 2006, Species level phylogeny of the genus cornus (cornaceae) based on molecular and morphological evidence—implications for taxonomy and tertiary intercontinental migration, *Taxon.*, **55**, 9–30, doi:10.2307/25065525.
11. Feng, C.M., Xiang, Q.Y. and Franks, R.G. 2011, Phylogeny-based developmental analyses illuminate evolution of inflorescence architectures in dogwoods (*Cornus* s. l., Cornaceae), *New Phytol.*, **191**, 850–69, doi:10.1111/j.1469-8137.2011.03716.x.
12. Fan, C.Z. and Xiang, Q.Y. 2001, Phylogenetic relationships within *Cornus* (Cornaceae) based on 26S rDNA sequences, *Am. J. Bot.*, **88**, 1131–8, doi:10.2307/2657096.
13. Zhang, J. 2008, Cultivation technique of *Cornus controversa*, *Gansu. Arg.*, **8**, 80–2.
14. Dai, C.C., Liu, X. and Song, L. 2009, Research on major medical component in different parts of *Bothrocaryum controversum*, *J. Anhui. Agr. Sci.*, **37**, 5490–1, doi:10.13989/j.cnki.0517-6611.2009.12.078.
15. Eyde, R.H. 1988, Comprehending *Cornus*: puzzles and progress in the systematics of the dogwoods, *Bot. Rev.*, **54**, 233–351.
16. Xiang, Q.Y., Soltis, D.E. and Soltis, P.S. 1998, Phylogenetic relationships of Cornaceae and close relatives inferred from matK and rbcL sequences, *Am. J. Bot.*, **85**, 285–97, doi:10.2307/2446317.
17. Li, R.J. and Shang, Z.Y. 2002, Karyotypes of five species of *Cornus (s.l.)* (Cornaceae) from China, *Acta. Phytotaxon Sin.*, **40**, 357–63.
18. Xiang, Q. Y. and Boufford, D. E. 2005, Flora of China, In: Wu Z. Y., Raven P. H., Hong D. Y. (eds.) *Cornaceae*, *vol.* **14**. Science Press, Beijing, pp. 206–221.
19. Xiang, Q.Y., Manchester, S.R., Thomas, D.T., Zhang, W. and Fan, C. 2005, Phylogeny, biogeography, and molecular dating of *cornelian cherries* (*Cornus*, Cornaceae)— tracking tertiary plant migration, *Evolution*, **59**, 1685–700, doi:10.1111/j.0014-3820.2005.tb01818.x.
20. Doyle, J.J. and Doyle, J.L. 1987, A rapid DNA isolation procedure for small quantities of fresh leaf tissue, *Phytochem. Bull.*, **19**, 11–5.
21. Louwers, M., Splinter, E., van Driel, R., de Laat, W. and Stam, M. 2009, Studying physical chromatin interactions in plants using Chromosome Conformation Capture, *Nat. Protoc.*, **4**, 1216–29, doi:10.1038/nprot.2009.113.
22. Vurture, G.W., Sedlazeck, F.J., Nattestad, M., et al. 2017, GenomeScope: fast reference-free genome profiling from short reads, *Bioinformatics*, **33**, 2202–4, doi:10.1093/bioinformatics/btx153.
23. Marçais, G. and Kingsford, C. 2011, A fast, lockfree approach for efficient parallel counting of occurrences of kmers, *Bioinformatics*, **27**, 764–70, doi:10.1093/bioinformatics/btr011.
24. Box, D., Ehnebuske, D., Kakivaya, G., et al. 2000, Simple Object Access Protocol (SOAP), *Encycl. Genet., Genoicms Proteomics Informatics*, **14**, 303–5.
25. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. and Li, H. 2021, Haplotype-resolved denovo assembly using phased assembly graphs with hifiasm, *Nat. Methods*, **18**, 170–175. doi:10.1038/s41592-020-01056-5.
26. Hu, J., Fan, J.P., Sun, Z.Y. and Liu, S.L. 2020, NextPolish: a fast and efficient genome polishing tool for long-read assembly, *Bioinformatics*, **36**, 2253–5, doi:10.1093/bioinformatics/btz891.
27. Vasimuddin, M., Misra, S., Li H. and Aluru, S. 2019, Efficient architecture-aware acceleration of BWA-MEM for multicore systems, 2019 IEEE, International Parallel and Distributed Processing Symposium (IPDPS), pp. 314–324. doi: 10.1109/IPDPS.2019.00041.
28. Li, H.and Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, **25**, 1754–60, doi:10.1093/bioinformatics/btp324.
29. Steven, W., Philip, E., Mayra, F.M., et al. 2015, HiCUP: pipeline for mapping and processing Hi-C data, *F1000Research*, **4**, 1310, doi:10.12688/f1000research.7334.1.

30. Dudchenko, O., Batra, S.S., Omer, A.D., et al. 2017, De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds, *Science*, **356**, 92–5, doi:10.1126/science.aal3327.
31. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with singlecopy orthologs, *Bioinformatics*, **31**, 3210–2, doi:10.1093/bioinformatics/btv351.
32. Price, A.L., Jones, N.C. and Pevzner, P.A. 2005, De novo identification of repeat families in large genomes, *Bioinformatics*, **21**, i351351–i358, doi:10.1093/bioinformatics/bti1018.
33. Tarailo-Graovac, M. and Chen, N. 2004, Using RepeatMasker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics*, **5**, 1–14, doi:10.1002/0471250953.bi0410s25.
34. Tian, T., Yue, L., Hengyu, Y., et al. 2017, agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update, *Nucleic Acids Res.*, **45**, 122–9, doi:10.1093/nar/gkx382.
35. Zhao, X. and Hao, W. 2007, LTR_FINDER: an efficient tool for the prediction of fulllength LTR retrotransposons, *Nucleic Acids Res.*, **35**, 265–8, doi:10.1093/nar/gkm286.
36. Ellinghaus, D., Kurtz, S. and Willhoeft, U. 2008, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, *BMC Bioinf.*, **9**, 18, doi:10.1186/1471-2105-9-18.
37. Ou, S. and Jiang, N. 2018, LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons, *Plant Physiol.*, **176**, 1410–22, doi:10.1104/pp.17.01310.
38. Stanke, M., Keller, O., Gunduz, I., et al. 2006, AUGUSTUS: ab initio prediction of alternative transcripts, *Nucleic Acids Res.*, **34**, W435–W439, doi:10.1093/nar/gkl200.
39. Burge, C. and Karlin, S. 1997, Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, **268**, 78–94, doi:10.1006/jmbi.1997.0951.
40. Majoros, W., Pertea, M. and Salzberg, S. 2004, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics*, **20**, 2878–8789, doi:10.1093/bioinformatics/bth315.
41. Keilwagen, J. and Hartung, F.J. 2019, GeMoMa: homology-based gene prediction utilizing intron position conservation and RNAseq data, *Methods Mol. Biol.*, **1962**, 161–77, doi:10.1007/978-1-4939-9173-0_9.
42. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, *Genome Biol.*, **9**, R7, doi:10.1186/gb-2008-9-1-r7.
43. Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. 2000, The COG database: a tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Res.*, **28**, 33–6, doi:10.1093/nar/28.1.33.
44. Judith, T., Christine, A., Jean-Charles, G., et al. 2015, Data for comparative proteomics of ovaries from five non-model, crustacean amphipods, *Data in Brief*, **5**, 1–6, doi:10.1016/j.dib.2015.07.037.
45. Amos, B. and Rolf, A. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL, *Nucleic Acids Res.*, **28**, 45–8, doi:10.1093/nar/28.1.45.
46. Zdobnov, E.M. and Rolf, A. 2001, InterProScan—an integration platform for the signaturerecognition methods in InterPro, *Bioinformatics*, **17**, 847–8, doi:10.1093/bioinformatics/17.9.847.
47. Wheeler, T.J. and Eddy, S.R. 2013, nhmmer: DNA homology search with profile HMMs, *Bioinformatics*, **29**, 2487–9, doi:10.1093/bioinformatics/btt403.
48. Li, L., Stoeckert, C.J. and Roos, D.S. 2003, OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res.*, **13**, 2178–89, doi:10.1101/gr.1224503.
49. Kazutaka, K. and Daron, M.S. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol.*, **30**, 772–80, doi:10.1093/molbev/mst010.
50. Stamatakis, A. 2014, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, **30**, 1312–3, doi:10.1093/bioinformatics/btu033.
51. Puttick, M.N. 2019, MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees, *Bioinformatics*, **35**, 5321–2, doi:10.1093/bioinformatics/btz554.
52. De Bie, T., Cristianini, N., Demuth, J.P. and Hahn, M.W. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, **22**, 1269–71, doi:10.1093/bioinformatics/btl097.
53. Lozano, R., Hamblin, M.T., Prochnik, S. and Jannink, J.-L. 2015, Identification and distribution of the NBSLRR gene family in the Cassava genome, *BMC Genom.*, **16**, 360, doi:10.1186/s12864-015-1554-9.
54. Xiang, L., Liu, J., Wu, C., et al. 2017, Genomewide comparative analysis of NBSencoding genes in four *Gossypium* species, *BMC Genom.*, **18**, 292. dio:10.1186/s12864-017-3682-x.
55. McDonnell, A.V., Jiang, T., Keating, A.E. and Berger, B. 2006, Paircoil2: improved prediction of coiled coils from sequence, *Bioinformatics*, **22**, 356–8, doi:10.1093/bioinformatics/bti797.
56. Qin, H. and Chamlong, P. 2007, Flora of China. In: Wu Z. Y, Raven P. H. and Hong D. Y. (eds.) *Nyssaceae*, *vol.* **13**. Science Press, Beijing, pp. 300–303.
57. Li, H. and Durbin, R. 2011, Inference of human population history from individual wholegenome sequences, *Nature*, **475**, 493–6, doi:10.1038/nature10231.
58. Danecek, P., Bonfield, J.K., Liddle, J., et al. 2021, Twelve years of SAMtools and BCFtools, *GigaScience*, **10**, 1–4, doi:10.1093/gigascience/giab008.
59. Sun, P., Jiao, B., Yang, Y., et al. 2022, WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes, *Mol Plant*, **15**, P1841–51, doi:10.1101/2021.04.29.441969.
60. Stolzer, M., Lai, H., Xu, M., et al. 2012, Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees, *Bioinformatics*, **28**, i409409–i415, doi:10.1093/bioinformatics/bts386.
61. Jiao, Y.N., Leebens-Mack, J., Ayyampalayam, S., et al. 2012, A genome triplication associated with early diversification of the core eudicots, *Genome Biol.*, **13**, R3, doi:10.1186/gb-2012-13-1-r3.
62. Chen, Y., Ma, T., Zhang, L., et al. 2020, Genomic analyses of a 'living fossil': the endangered dovetree, *Mol. Ecol. Resour.*, **20**, 756–69, doi:10.1111/1755-0998.13138.
63. Li, M.J., Yang, Y.Z., Xu, R.P., et al. 2021, A chromosome-level genome assembly for the Tertiary relict plant *Tetracentron sinense* Oliv. (Trochodendraceae), *Mol. Ecol. Resour.*, **21**, 1–14, doi:10.1111/1755-0998.13334.
64. Kang, M.H., Fu, R., Zhang, P.Y., et al. 2021, A chromosome-level *Camptotheca acuminata* genome assembly provides insights into the evolutionary origin of camptothecin biosynthesis, *Nat. Commun.*, **12**, 3531, doi:10.1038/s41467-021-23872–23879.
65. Zhu, S., Chen, J., Zhao, J., et al. 2020, Genomic insights on the contribution of balancing selection and local adaptation to the longterm survival of a widespread living fossil tree, *Cercidiphyllum japonicum*, *New Phytol.*, **228**, 1674–89, doi:10.1111/nph.16798.
66. Wendel, J.F., Jackson, S.A., Meyers, B.C. and Wing, R.A. 2016, Evolution of plant genome architecture, *Genome Biol.*, **17**, 37, doi:10.1186/s13059-016-0908-1.
67. Clark, P.U., Dyke, A.S., Shakun, J.D., et al. 2009, The last glacial maximum, *Science*, **325**, 710–4.
68. Moghe, G.D. and Shiu, S.H. 2014, The causes and molecular consequences of polyploidy in flowering plants, *Ann. N. Y. Acad. Sci.*, **1320**, 16–34, doi:10.1111/nyas.12466.
69. Wu, S.D., Han, B.C. and Jiao, Y.N. 2020, Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms, *Mol. Plant*, **13**, 59–71, doi:10.1016/j.molp.2019.10.012.
70. Ren, G.P., Jiang, Y.Y., Li, A., et al. 2021, The genome sequence provides insights into salt tolerance of *Achnatherum splendens* (Gramineae), a constructive species of alkaline grassland, *Plant Biotechnol. J.*, **20**, 116–28, doi:10.1111/pbi.13699.
71. Pont, C. and Salse, J. 2017, Wheat paleohistory created asymmetrical genomic evolution, *Curr. Opin. Plant Biol.*, **6**, 29–37, doi:10.1016/j.pbi.2017.01.001.