SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

# A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*

Xuchen Yang[1,3], Minghui Kang[1,3], Yanting Yang[1,3], Haifeng Xiong[1], Mingcheng Wang[1], Zhiyang Zhang[1], Zefu Wang[1], Haolin Wu[1], Tao Ma[1], Jianquan Liu[1,2] & Zhenxiang Xi[1*]

**The deciduous Chinese tupelo (*Nyssa sinensis* Oliv.) is a popular ornamental tree for the spectacular autumn leaf color. Here, using single-molecule sequencing and chromosome conformation capture data, we report a high-quality, chromosome-level genome assembly of *N. sinensis*. PacBio long reads were *de novo* assembled into 647 polished contigs with a total length of 1,001.42 megabases (Mb) and an N50 size of 3.62 Mb, which is in line with genome sizes estimated using flow cytometry and the *k*-mer analysis. These contigs were further clustered and ordered into 22 pseudo-chromosomes based on Hi-C data, matching the chromosome counts in *Nyssa* obtained from previous cytological studies. In addition, a total of 664.91 Mb of repetitive elements were identified and a total of 37,884 protein-coding genes were predicted in the genome of *N. sinensis*. All data were deposited in publicly available repositories, and should be a valuable resource for genomics, evolution, and conservation biology.**

## Background & Summary

*Nyssa sinensis* Oliv., commonly known as Chinese tupelo, is a deciduous tree with ovate leaves, which turn brilliant red, orange, and yellow in autumn. It belongs to the family Nyssaceae within the order Cornales, and is native to southern China and Vietnam. The genus *Nyssa* comprises three species in North America (i.e., *N. aquatica*, *N. ogeche*, and *N. sylvatica*), three in eastern Asia (i.e., *N. javanica*, *N. sinensis*, and *N. yunnanensis*), and one in Costa Rica (i.e., *N. talamancana*)[1,2]. Thus, it is one of the plant genera that exhibit a classical disjunct distribution between eastern Asia and North America. In addition, the fossil record of *Nyssa* is very rich[1,3,4], making it ideal for studying the evolutionary history of Tertiary relict floras.

The genome of *Camptotheca acuminata* (Nyssaceae) has recently been sequenced using the Illumina platform[5], and the final assembly is 403.17 megabases (Mb) with a contig N50 size of 107.59 kilobases (kb). So far, this is the only sequenced genome within the order Cornales. Here, we utilized a combination of the PacBio long-read sequencing technology[6] and the high-throughput chromosome conformation capture (Hi-C) technique[7] to generate the genome sequence of *N. sinensis*. Long reads were *de novo* assembled into 647 polished contigs with a total length of 1,001.42 Mb and an N50 size of 3.62 Mb, which is in line with genome sizes estimated using flow cytometry and the *k*-mer analysis. These contigs were further clustered and ordered into 22 pseudo-chromosomes based on Hi-C data. Our results provide the first high-quality, chromosome-level genome assembly for the order Cornales, which should be a valuable resource for genomics, evolution, and conservation biology.

## Methods

**Sample collection and high-throughput sequencing.** We sampled a single individual of *N. sinensis* from the Kunming Botanical Garden, Yunnan, China. The total genomic DNA was extracted from fresh leaves using a modified CTAB method[8], and sequenced using the PacBio Sequel System (for genome assembly) and the Illumina HiSeq. 4000 System (for genome survey and base level correction after the assembly). Here, one library with an insertion size of 350 bp was prepared for the Illumina platform and 20-kb libraries were constructed for the PacBio platform according to the manufacturers' protocols. A total of 104.34 gigabases (Gb) of polymerase reads were generated using the PacBio platform, and a total of 104.19 Gb (coverage of 99.12×) of subreads

[1]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, 610065, China. [2]State Key Laboratory of Grassland Agro-Ecosystems, College of Life Sciences, Lanzhou University, Lanzhou, 730000, China. [3]These authors contributed equally: Xuchen Yang, Minghui Kang and Yanting Yang. *email: zxi@scu.edu.cn

| Library type | Platform | Read length | Clean reads | Clean base | Coverage | Application |
|---|---|---|---|---|---|---|
| Long reads | PacBio Sequel | 14,526 bp (N50) | 11,197,047 | 104.19 Gb | 99.12× | Genome assembly |
| Short reads | HiSeq. 4000 | 2 × 150 bp | 2 × 196,110,604 | 58.83 Gb | 55.97× | Genome survey and base level correction |
| Hi-C | HiSeq. 4000 | 2 × 150 bp | 2 × 423,362,084 | 126.81 Gb | 120.63× | Chromosome construction |
| RNA-Seq | HiSeq. 4000 | 2 × 150 bp | 2 × 57,866,710 | 17.36 Gb | — | Genome annotation |

**Table 1.** Summary of sequencing data generated in this study.

| *K*-mer | *K*-mer number | *K*-mer depth | Genome size | Heterozygosity rate | Repeat |
|---|---|---|---|---|---|
| 17 | 49,922,730,728 | 45 | 1,051.16 Mb | 0.87% | 56.92% |

**Table 2.** Summary of the *k*-mer analysis for estimating the genome size of *Nyssa sinensis*. 150-bp paired-end reads were generated using the Illumina platform, and a total of 58.83 Gb of reads were obtained after adapter trimming and quality filtering. The frequency of each *k*-mer was calculated and plotted in Fig. 1.

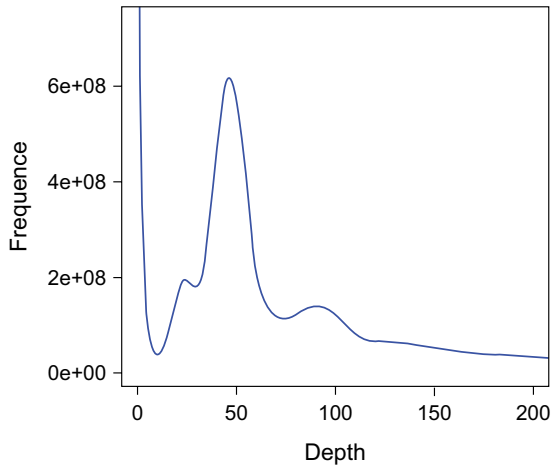


**Fig. 1** The *k*-mer analysis ($k = 17$) for estimating the genome size of *Nyssa sinensis*. The x-axis refers to the *k*-mer depth; the y-axis refers to the frequency of the *k*-mer for a given depth.

were obtained after removing adaptors in polymerase reads (Table 1). The N50 read length reached 22.26 kb and 14.53 kb for polymerase reads and subreads, respectively. A total of 58.92 Gb of 150-bp paired-end reads were generated using the Illumina platform, and a total of 58.83 Gb (coverage of 55.97×) of reads were obtained after adapter trimming and quality filtering (Table 1). In addition, the Hi-C library was constructed using young leaf tissue from the same individual of *N. sinensis*, and sequenced using the Illumina platform. A total of 126.81 Gb (coverage of 120.63×) of 150-bp paired-end reads were obtained after adapter trimming and quality filtering (Table 1), which were later applied to extend the contiguity of the genome assembly to the chromosomal level. Furthermore, leaves and flowers were collected from the same individual of *N. sinensis*, and RNA-Seq reads were generated for genome annotation using the Illumina platform. A total of 17.36 Gb of 150-bp paired-end reads were obtained after adapter trimming and quality filtering (Table 1).

**Genome size and heterozygosity estimation.** The genome size of *N. sinensis* was first estimated using the *k*-mer analysis with Jellyfish[9]. The 17-mer frequency of Illumina short reads followed a Poisson distribution, with the highest peak occurring at a depth of 45 (Fig. 1). The estimated genome size was 1,051.16 Mb, and the heterozygosity rate of the genome was 0.87% (Table 2). In addition, we performed flow cytometry analysis using *Vigna radiata* as the internal standard, and the genome size of *N. sinensis* was estimated at 992 Mb.

***De novo* genome assembly and pseudo-chromosome construction.** After the self-error correction step, the PacBio long reads were assembled into contigs using the hierarchical genome assembly process (HGAP)[10] as implemented in the FALCON assembler[11,12]. In addition, two rounds of polishing were applied to the assembled contigs using the Quiver algorithm[10] with the PacBio long reads, and another round of the genome-wide base-level correction was performed using Pilon[13] with the Illumina short reads. Finally, the Purge Haplotigs pipeline[14] was run to produce an improved, deduplicated assembly. The resulting genome assembly

| | FALCON assembly | Post Quiver | Post Pilon | Post Purge Haplotigs |
|---|---|---|---|---|
| Size of assembled contigs (bp) | 1,060,185,320 | 1,066,201,220 | 1,065,961,334 | 1,001,417,765 |
| No. of contigs (>100 bp) | 1,553 | 1,553 | 1,553 | 647 |
| Max. contig length (bp) | 29,948,976 | 30,066,399 | 30,063,645 | 30,063,645 |
| Contig N50 size (bp) | 3,447,630 | 3,466,154 | 3,466,018 | 3,624,455 |
| Contig N90 size (bp) | 525,139 | 527,948 | 527,985 | 1,008,072 |

**Table 3.** Summary of genome assemblies of *Nyssa sinensis* created at different stages of the assembly process.

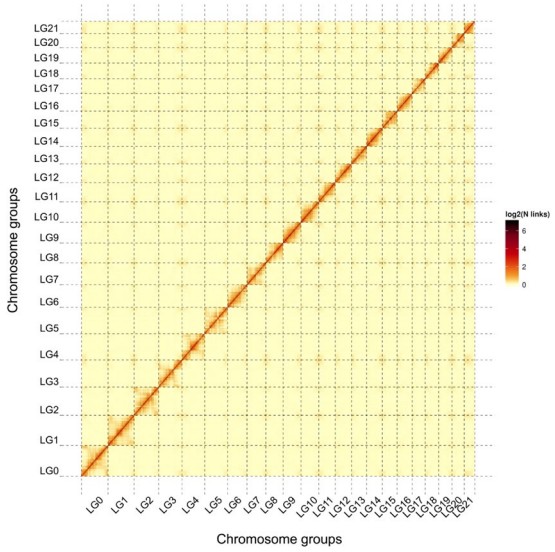


**Fig. 2** Interaction heat map of Hi-C links among chromosome groups for *Nyssa sinensis*. The assembled genome of *N. sinensis* was divided into 100-kb non-overlapping windows (or bins), and valid interaction links of Hi-C data were calculated between each pair of bins. The binary logarithm of each link number is coded using colors ranging from light yellow to dark red, indicating the frequency of Hi-C interaction links from low to high. LG0–LG21 represent the 22 chromosome groups inferred by LACHESIS.

of *N. sinensis* contained 1,001.42 Mb of sequences in 647 polished contigs with an N50 size of 3.62 Mb (contigs shorter than 100 bp were discarded; Table 3), and the overall GC-content was 35.98%.

Construction of pseudo-chromosomes followed the previous study[15] using the Hi-C library. Briefly, the clean Hi-C reads were mapped to the assembled contigs using the Burrows–Wheeler Aligner[16] (BWA), and only uniquely mapped read pairs were considered for downstream analysis. Duplicate removal, sorting, and quality assessment were performed using HiC-Pro[17]. The assembled contigs were then clustered, ordered, and oriented into pseudo-chromosomes using LACHESIS[18]. A total of 585 contigs spanning 1,000.96 Mb (i.e., 99.95% of the assembly) were clustered into 22 chromosome groups (Fig. 2), matching the chromosome counts in *Nyssa* ($n = 22$) based on cytological studies[19–21]. In addition, of the clustered contigs, 382 contigs spanning 968.49 Mb (i.e., 96.71% of the assembly) were successfully ordered and orientated (Online-only Table 1).

**The annotation of repetitive elements.** To annotate repetitive elements in the genome of *N. sinensis*, we utilized a combination of evidence-based and *de novo* approaches. The genome assembly was first searched using RepeatMasker (http://www.repeatmasker.org) against the Repbase database[22]. Next, a *de novo* repetitive element library was constructed using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/), which employed results from RECON[23] and RepeatScout[24]. This *de novo* repetitive element library was then utilized by RepeatMasker to annotate repetitive elements. Results from these two runs of RepeatMasker were merged together. A total of 664.91 Mb of repetitive elements (i.e., 66.40% of the assembly) were identified in the genome of *N. sinensis* (Table 4), including retroelements (32.51%), DNA transposons (11.23%), tandem repeats (2.95%), and unclassified elements (19.71%). Thus, the percentage of predicted repetitive elements in the genome of *N. sinensis* is much higher in comparison with that in the closely related species *C. acuminata* (i.e., 35.6%[5]).

Long terminal repeat (LTR) retrotransposons are prevalent in plant genomes[25]. In order to develop high-quality gene annotation, we additionally identified LTR retrotransposons in the genome of *N. sinensis* using a combination of four programs (i.e., LTR_FINDER[26], LTRharvest[27], LTR_retriever[25], and RepeatMasker). Here, LTR_FINDER and LTRharvest were used for initial identification of LTR retrotransposons; LTR_retriever was then used to filter out false positives and estimate the insertion time for each intact LTR retrotransposon; finally, RepeatMasker was used for annotation of LTR retrotransposons. Our results suggested that when comparing with

| Element type | No. of elements | Length occupied (bp) | Percentage of genome (%) |
|---|---|---|---|
| LTR | 433,015 | 284,127,511 | 28.37 |
| LINE | 83,199 | 40,234,938 | 4.02 |
| SINE | 9,696 | 1,168,096 | 0.12 |
| DNA | 326,322 | 112,436,370 | 11.23 |
| Satellite | 2,879 | 959,079 | 0.10 |
| Simple repeats | 72,249 | 28,565,346 | 2.85 |
| Unclassified | 582,325 | 197,421,711 | 19.71 |
| Total | 1,509,685 | 664,913,051 | 66.40 |

**Table 4.** Summary of repetitive elements annotated in the genome of *Nyssa sinensis*.

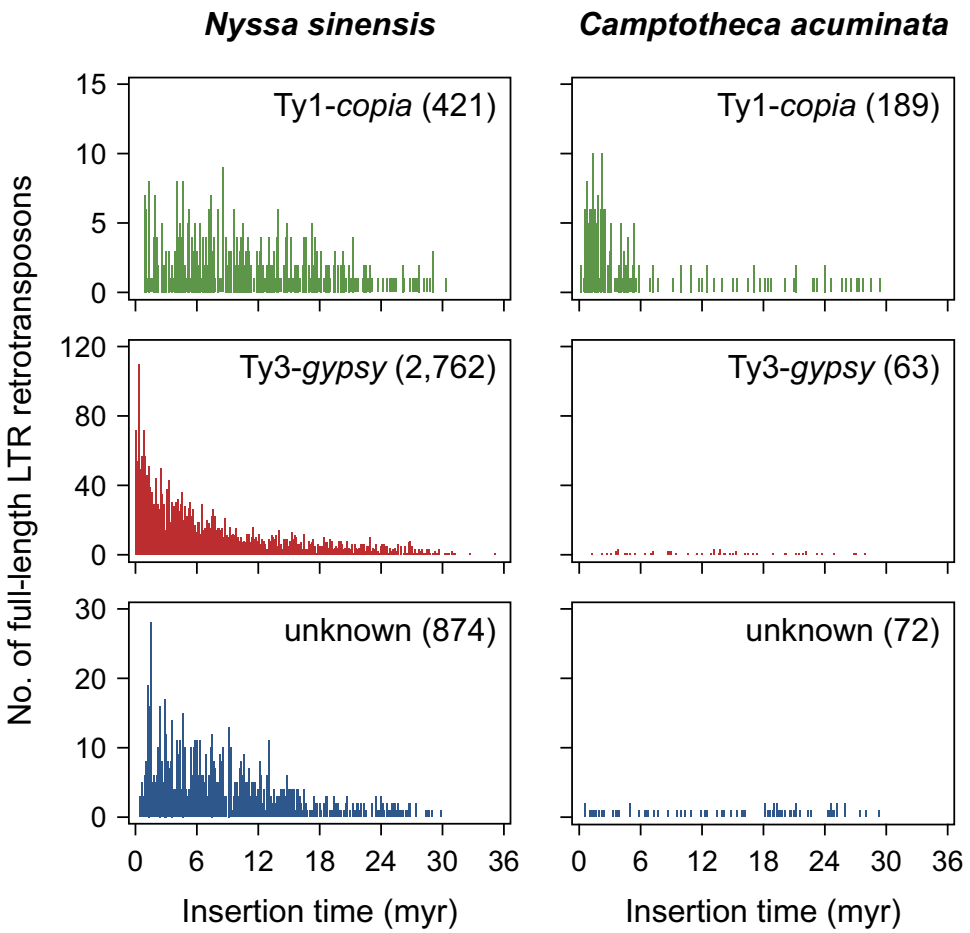


**Fig. 3** Estimated insertion time of full-length long terminal repeat (LTR) retrotransposons in the genomes of *Nyssa sinensis* and *Camptotheca acuminata*.

*C. acuminata*, LTR retrotransposons in the genome of *N. sinensis* had recently undergone a rapid proliferation, particularly the Ty3-*gypsy* family (Fig. 3).

**Protein-coding gene prediction and functional annotation.** The identification of protein-coding genes in the assembled genome of *N. sinensis* was based on transcriptome data and *ab initio* prediction. First, two strategies (i.e., *de novo* and genome-guided assembly) were applied to assemble RNA-Seq reads into transcripts using Trinity[28]. In order to use Trinity in genome-guided mode, RNA-Seq reads were first aligned to the assembled genome of *N. sinensis* using HISAT2[29]. These two transcriptome assemblies were then merged. To generate the initial gene models for training AUGUSTUS[30], our assembled transcripts were processed and utilized to identify open reading frames (ORFs) by the Program to Assemble Spliced Alignments[31] (PASA). AUGUSTUS was then utilized for *ab initio* gene prediction based on (i) a generalized hidden Markov model (HMM) and (ii) semi-Markov conditional random field (CRF). In addition, extrinsic evidence was incorporated into AUGUSTUS using a hints file, which was generated by aligning RNA-Seq reads to the hard-masked genome assembly with

| | |
|---|---|
| No. of protein-coding genes | 37,884 |
| No. of transcripts | 62,426 |
| Average exon size per transcript (bp) | 1,949 |
| Average coding sequence (CDS) size per transcript (bp) | 1,240 |
| Average intron size per transcript (bp) | 7,728 |
| Average exon number per transcript | 6.42 |
| Average exon size (bp) | 303 |

**Table 5.** Summary of protein-coding genes predicted in the genome of *Nyssa sinensis*.

| Database | No. of annotated genes | Percentage (%) |
|---|---|---|
| Swiss-Prot | 28,305 | 74.71 |
| TrEMBL | 33,656 | 88.84 |
| InterPro | 35,769 | 94.42 |
| GO | 32,293 | 85.24 |
| KEGG | 8,235 | 21.74 |
| Annotated | 36,185 | 95.52 |
| Total | 37,884 | 100.00 |

**Table 6.** Summary of functional annotation of protein-coding genes in the genome of *Nyssa sinensis*.

HISAT2. Lastly, untranslated regions (UTRs) and alternative splicing variations were annotated using PASA. A total of 37,884 protein-coding genes were predicted in the genome of *N. sinensis* (Table 5).

For functional annotation, our predicted protein-coding genes were searched against the Swiss-Prot and TrEMBL databases[32] using BLAST+[33] with and *E*-value threshold of 1e-05, as well as the InterPro database using InterProScan 5[34]. In addition, for predicted protein-coding genes, gene ontology (GO) annotations were performed using Blast2GO[35], and KEGG orthology (KO) identifiers were assigned using KEGG Automatic Annotation Server[36] (KAAS). A total of 36,185 genes (i.e., 95.52% of all predicted protein-coding genes) were successfully annotated by at least one database (Table 6).

## Data Records

PacBio Sequel long reads[37], Illumina paired-end reads[38], Hi-C reads[39], and RNA-Seq reads[40] have been deposited in NCBI Sequence Read Archive (SRA). The genome assembly and annotation of *N. sinensis* have been deposited in CoGe[41], Figshare[42,43], and GenBank[44].

## Technical Validation

**Total RNA quality assessment.** The quality of total RNA was evaluated using (i) agarose gel electrophoresis for RNA degradation and potential contamination, (ii) NanoDrop spectrophotometer for preliminary quantitation, and (iii) Agilent 2100 Bioanalyzer for RNA integrity and quantitation. Total RNA samples included in this study had an RNA integrity number (RIN) of 9.7–10 and an rRNA ratio of 1.5, which were then enriched for mRNA via an oligo(dT)–magnetic bead method.

**Quality filtering of Illumina data.** Illumina raw data were first filtered using Trimmomatic[45] to remove paired-end reads if either of the reads contained (i) adapter sequences, (ii) more than 10% of N bases, and (iii) more than 20% of bases with a Phred quality score less than 5.

**Assessing the completeness and accuracy of the genome assembly.** We first evaluated the completeness of the assembly using CEGMA[46] and BUSCO[47,48]. Out of the 248 core eukaryotic genes in CEGMA, 235 (94.8%) complete matches and 244 (98.4%) complete plus partial matches were found in the assembled genome of *N. sinensis*. In addition, 93.4% complete and 2.2% partial of the 1,440 plant-specific BUSCO genes were identified in the assembly. Second, the accuracy of the assembly was assessed using our Illumina short reads. In total, 94.51% of the filtered short reads (58.83 Gb, Table 1) were mapped to the assembled genome of *N. sinensis* using BWA, which covered 99.89% of the assembly. Furthermore, Single-nucleotide polymorphisms (SNPs) were called and filtered using SAMtools[49], and a total of 5,046,556 SNPs with a sequencing depth between 10× and 100× were identified, consisting of 5,040,788 heterozygous SNPs and 5,768 homozygous SNPs. The low rate of homozygous SNPs (0.0006% of the assembled genome) suggested the high accuracy of the assembly. Finally, the assembled genome of *N. sinensis* was divided into 10-kb non-overlapping windows, and the scatter plot of the sequencing depth versus the GC-content based on 10-kb windows indicated no contamination of foreign DNA in the assembly.

## Code availability

Sequencing data were generated using the software provided by sequencer manufacturers, and processed following the instruction manual of the software cited above. No custom codes were generated for this work.

Received: 18 July 2019; Accepted: 21 October 2019;
Published online: 25 November 2019

## References

1. Wen, J. & Stuessy, T. F. The phylogeny and biogeography of *Nyssa* (Cornaceae). *Syst. Bot.* **18**, 68–79 (1993).
2. Wang, N. *et al*. Phylogeny and a revised classification of the Chinese species of *Nyssa* (Nyssaceae) based on morphological and molecular data. *Taxon* **61**, 344–354 (2012).
3. Eyde, R. H. Fossil record and ecology of *Nyssa* (Cornaceae). *Bot. Rev.* **63**, 97–123 (1997).
4. Manchester, S. R., Grímsson, F. & Zetter, R. Assessing the fossil record of asterids in the context of our current phylogenetic framework. *Ann. Missouri Bot. Gard.* **100**, 329–363 (2015).
5. Zhao, D. *et al*. *De novo* genome assembly of *Camptotheca acuminata*, a natural source of the anti-cancer compound camptothecin. *GigaScience* **6**, gix065 (2017).
6. Eid, J. *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
7. van Berkum, N. L. *et al*. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39**, e1869 (2010).
8. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
9. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**, 764–770 (2011).
10. Chin, C.-S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
11. Pendleton, M. *et al*. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
12. Chin, C.-S. *et al*. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
13. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One* **9**, e112963 (2014).
14. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
15. Yin, D. *et al*. Genome of an allotetraploid wild peanut *Arachis monticola*: a *de novo* assembly. *GigaScience* **7**, giy066 (2018).
16. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
17. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
18. Burton, J. N. *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
19. Dermen, H. Cytological studies of *Cornus*. *J. Arnold Arbor.* **13**, 410–416 (1932).
20. Mehra, P. N. & Bawa, K. S. Chromosomal evolution in tropical hardwoods. *Evolution* **23**, 466–481 (1969).
21. Goldblatt, P. A contribution to cytology in Cornales. *Ann. Missouri Bot. Gard.* **65**, 650–655 (1978).
22. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11 (2015).
23. Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
24. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
25. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
26. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
27. Ellinghaus, D., Kurtz, S. & Willhoeft, U. *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
28. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
29. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
30. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
31. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
32. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
33. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
34. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
35. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
36. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–W185 (2007).
37. *NCBI Sequence Read Archive*, http://identifiers.org/ncbi/insdc.sra:SRX6405746 (2019).
38. *NCBI Sequence Read Archive*, http://identifiers.org/ncbi/insdc.sra:SRX6441717 (2019).
39. *NCBI Sequence Read Archive*, http://identifiers.org/ncbi/insdc.sra:SRX6441715 (2019).
40. *NCBI Sequence Read Archive*, http://identifiers.org/ncbi/insdc.sra:SRX6441716 (2019).
41. Yang, X. *et al*. A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*. *CoGe*, https://genomevolution.org/CoGe/GenomeInfo.pl?gid=55419 (2019).
42. Yang, X. *et al*. A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*. *figshare*, https://doi.org/10.6084/m9.figshare.8872700 (2019).
43. Yang, X. *et al*. A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*. *figshare*, https://doi.org/10.6084/m9.figshare.8872721 (2019).
44. Yang, X. *et al*. *Nyssa sinensis* isolate J267, whole genome shotgun sequencing project. *GenBank*, http://identifiers.org/ncbi/insdc:VIRR00000000 (2019).
45. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

46. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
47. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
48. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
49. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

## Acknowledgements

The authors thank Kai Chen, Jeffery DaCosta, Charles Davis, Christopher Grassa, Quanjun Hu, Leke Lyu, Xingxing Mao, Lei Zhang, and Yazhou Zhang for technical assistance and valuable discussions. This work was supported by the National Key Research and Development Program of China (2017YFC0505203) and the National Natural Science Foundation of China (31600172 and 31770232).

## Author contributions

J.L. and Z.X. designed research; X.Y., M.K. and Y.Y. performed research; X.Y., M.K., Y.Y., H.X., M.W., Z.Z., Z.W., H.W. and T.M. analyzed data; X.Y., J.L. and Z.X. wrote the manuscript; and all authors read, edited, and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to Z.X.