The draft genome assembly of the critically endangered *Nyssa yunnanensis*, a plant species with extremely small populations endemic to Yunnan Province, China

Weixue Mu[1], Jinpu Wei[1], Ting Yang[1], Yannan Fan[1], Le Cheng[2], Jinlong Yang[2], Ranchang Mu[3], Jie Liu[3], Jianming Zhao[3], Weibang Sun[4], Xun Xu[1,7], Xin Liu[1], Radoje Drmanac[1,6*], Huan Liu[1,5*]

1. State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen 518083, China
2. BGI-Yunnan, BGI-Shenzhen, Kunming, 650106, China
3. Forestry Bureau of Ruili, Yunnan Dehong, Ruili 678600, China
4. Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650204, Yunnan, China
5. Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark.
6. Complete Genomics Inc., 2904 Orchard Pkwy, San Jose, CA 95134 USA.
7. Guangdong Provincial Key Laboratory of Genome Read and Write, Shenzhen 518083, China

ORCIDs:

Weixue Mu: 0000-0002-2675-261X; Jinpu Wei: 0000-0002-1463-0236; Ting Yang: 0000-0002-2872-4954; Yannan Fan: 0000-0003-3308-6878; Jinlong Yang: 0000-0003-1323-7755; Xun Xu: 0000-0002-5338-5173; Xin Liu: 0000-0003-3256-2940; Huan Liu: 0000-0003-3909-0931

Abstract

Nyssa yunnanensis is a deciduous tree species in the family Nyssaceae within the order Cornales. As only eight individual trees and two populations have been recorded in China's Yunnan province, this species has been listed among China's national Class I protection species since 1999 and also among 120 PSESP (Plant Species with Extremely Small Populations) in the *Implementation Plan of Rescuing and Conserving China's Plant Species with Extremely Small Populations(PSESP)* (2011-2-15). Here, we present the draft genome assembly of *N. yunnanensis*. Using 10X Genomics linked-reads sequencing data, we carried out the *de novo* assembly and annotation analysis. The *N. yunnanensis* genome assembly is 1475 Mb in length, containing 288,519 scaffolds with a scaffold N50 length of 985.59 kb. Within the assembled genome, 799.51 Mb was identified as repetitive elements, accounting for 54.24% of the sequenced genome, and a total of 39,803 protein-coding genes were predicted.

With the genomic characteristics of *N. yunnanensis* available, our study might facilitate future conservation biology studies to help protect this extremely threatened tree species.

Data Description

Nyssa yunnanensis (NCBI: txid161873), which belongs to the family Nyssaceae, is an extremely threatened range-restricted tree species among the Critically Endangered (CR) in the IUCN Red List of Threatened Species [1], as well as a national key protected species under Class I protection in China [2]. *Nyssa yunnanensis* is also listed as one of 120 PSESP (Plant Species with Extremely Small Populations) in the *Implementation Plan of Rescuing and Conserving China's Plant Species with extremely Small Populations (PSESP)* (2011-2-15) and as critically endangered in the Threatened Species List of China's Higher Plants [3, 4] (Figure 1).

Figure 1. Photograph of *Nyssa yunnanensis* from Ruili, Yunnan Province, China.

It is a canopy tree species able to reach 30 meters in height and it is functionally

dioecious, consisting of two individual species. One species bears staminate flowers

while the other type bears flowers that are morphologically normal but produce

inaperturate and inviable pollen grains. *Nyssa yunnanensis* does not appear to exhibit

parthenogenesis [5]. A survey of the population and ecological characteristics of *N.*

yunnanensis suggested that this species is at high risk of extinction, as only two

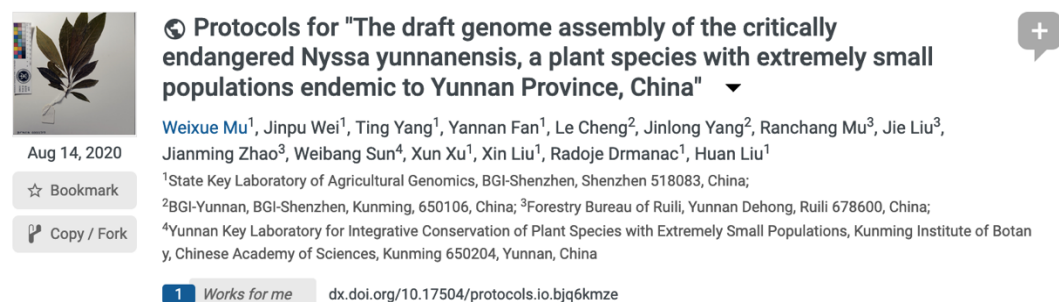
natural populations and eight individual trees have been recorded in Yunnan, China.

This scarcity might be caused by both ecological and human factors [6]. In 2009, an

integrated PSESP conservation strategy was initiated for *N. yunnanensis*. After more

than seven years of implementation, the natural populations are now securely

protected. Along with the development of propagation technologies and production of

vigorous seedlings, three new populations as well as four ex situ germplasm

collections of *N. yunnanensis* have now been established [7]. Phylogenetic studies of the six *Nyssa* species (*N. yunnanensis, N. javanica, N. sinensis, N. shangszeensis, N. shweliensis and N. wenshanensis*) recognized by the Flora of China have been conducted. Based on morphological and molecular evidence, the results suggest that only *N. sinensis, N. yunnanensis and N. javanica* should be recognized as species [8]. Recent advances in whole genome sequencing technology have provided valuable genomic resources to help us better understand the origin and evolutionary history of endangered species and to improve conservation strategies [9]. *Acer yangbiense,* another plant species with extremely small populations endemic to Yunnan Province, was sequenced in 2019. The *A. yangbiense* genome has a total length of 666 Mb with 13 chromosomes and a scaffold N50 size of 45 Mb [10]. The recently published genomes of *Nyssa sinensiss* and *Camptotheca acuminata* are the only two genome assemblies that have been sequenced within the Nyssaceae family. The *N. sinensiss* genome is 1,001.42 Mb in length with an N50 scaffold size of 3.62 Mb [11], and the *C. acuminate* genome is 403.17 Mb in length with an N50 scaffold size of 1753 Kb [12].

Although *N. yunnanensis* is not the first species sequenced in the Nyssaceae family, a detailed understanding of this endangered species' genomic makeup along with other information, such as population structure and reproductive biology, is urgently required to improve the PSESP conservation strategy for its continued survival.

Methods

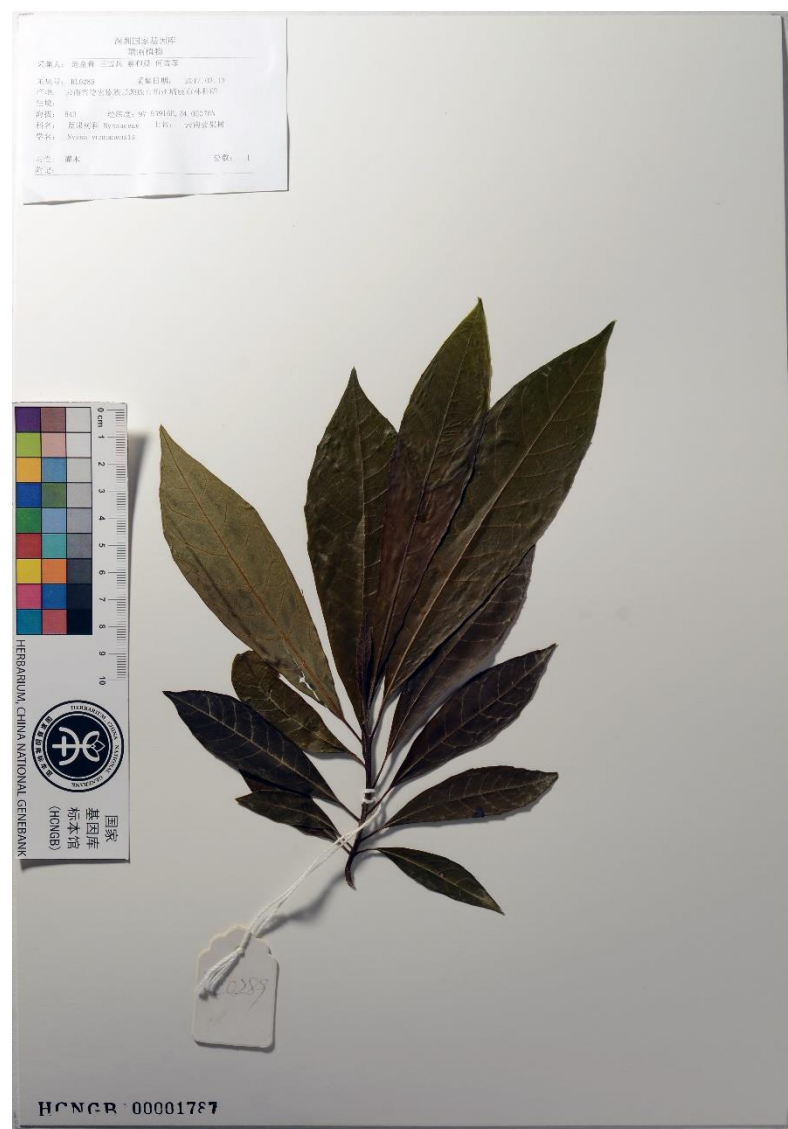
A protocol collection gathering together methods for DNA extraction and with DNBSEQ-G50 and 10X library construction and sequencing is available via protocols.io (Figure 2).

Figure 2. Protocol collection for the draft genome assembly of the critically endangered *Nyssa yunnanensis*, a plant species with extremely small populations endemic to Yunnan Province, China.

Plant material

We selected and sampled a 70 cm high individual tree of *Nyssa yunnanensis* from Ruili, Yunnan province, China (97°56′20.99″ N, 24°03′02.72″ E, altitude 843 m). Fresh young leaves were collected then immediately transferred into liquid nitrogen and stored in dry ice until DNA and RNA extraction. Voucher specimens and images were collected and stored in the CNGB herbarium (Figure 3). The extracted DNA is now stored in the BGI-sample center (voucher RL0289 and RL1182).

Figure 3. Photograph of the voucher specimens of *Nyssa yunnanensis*, stored in the CNGB herbarium (voucher RL0289)

HCNGB 00001787

DNA extraction and sequencing

Total genomic DNA was extracted from leaf tissues of *N. yunnanensis* using a
modified CTAB method [13]. Quality control was done using a NanoDrop One UV-
Vis spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit 3.0 Fluorometer
(Thermo Fisher Scientific, USA). A Sage Science Pippin Pulse electrophoresis system
was used to evaluate the molecular weight of the DNA and high-molecular-weight
(HMW) gDNA with a length of around 50 kb was obtained for further sequencing.
The HMW gDNA was then loaded onto a Chromium Controller chip with 10X
Chromium reagents and gel beads, and the rest of the library preparation procedures
were carried out according to the manufacturer's protocol [14]. Subsequently, the

sequencing was performed on a DNBSEQ-G50 (previously known as BGISEQ-500, RRID:SCR_017979) platform at BGI-Shenzhen (BGI Co. Ltd., Shenzhen, China) according to the manufacturer's instructions [15]. Using the whole-genome shotgun sequencing strategy, a total of ~163.66 Gb of raw data (150 bp, paired-end) was eventually generated, covering about 100× the sequencing depth of the 1.64Gb estimated genome size. All of the newly generated raw 10X Genomics reads were trimmed and filtered for adapter sequences and low-quality reads using Trimmomatic v. 0.38 (Trimmomatic, RRID:SCR_011848) [16] with the parameters "ILLUMINACLIP:adapter.fa:2:35:4 HEADCROP:5 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:50".

RNA extraction and sequencing

Total RNA was extracted from young leaves of the same individual *N. yunnanensis* tree using a CTAB-pBIOZOL method [17]. The purity, concentration, and integrity of RNA samples were measured on a NanoDrop One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA), a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA) and with Agilent 2100 Bioanalyzer on-chip electrophoresis (Agilent Technologies, Inc; Santa Clara, CA) [18], respectively, to ensure that the samples qualified for transcriptome sequencing. The cDNA library was prepared using the TruSeq RNA Sample Preparation Kit v2 (Illumina, San Diego, CA, USA), and then sequenced on a DNBSEQ-G50 platform at BGI-Shenzhen (BGI Co. Ltd., Shenzhen, China), resulting in ~8.85 Gb of raw transcriptome data (150 bp, paired-end). All of the raw sequence reads were further filtered using SOAPfilter v2.2 (SOAP, RRID:SCR_000689) with the parameters "-y -q 33 -i 200 -g 1 -M 2 -Q 20" to remove the adapters and low-quality reads.

Genome size estimation

The 1.64 Gb genome size of *N. yunnanensis* was estimated using the 21 k-mer counts of clean reads from the 10X Genomics library. First, K-mer frequency distribution

analyses were performed using the kmer_freq_hash software within the gce v1.0.0 package (GCE, RRID:SCR_017332) based on the clean 10X Genomics data with the parameters "-k 21 -l reads.lst -t 8". Then, gce software within the same package was used to estimate the overall characteristics of the genome, including genome size, repeat proportions, and level of heterozygosity [19].

De novo genome assembly

De novo assembly was carried out using Supernova v2.0.0 software (Supernova assembler, RRID:SCR_016756) [20] with the "--lanes=1 --localcores=48 --localmem=350 --maxreads 691040000" parameters. Raw 10X Genomics linked read data without trimming were used as the software recommended. Then, the gaps within the scaffolds were filled by GapCloser v1.12 (GapCloser, RRID:SCR_015026) [21] with the parameters "-l 150 -t 32" based on the clean 10X Genomics pair-end reads.

Genome evaluation

The completeness of the *N. yunnanensis* assembly was estimated using two strategies. First, we assessed completeness using BUSCO (BUSCO; v3.0.2, RRID:SCR_015008) [22] with the parameters "-l Embryophyta_odb10 -m geno -c 8 -f". Second, all the clean transcriptome reads and clean 10X Genomics reads were mapped back to the final genome assembly using BWA-MEM (BWA, version 0.7.16, RRID:SCR_010910) [23] with default parameters.

Repeat annotation

Repetitive elements were identified using both homology-based and *de novo* predictions in the *N. yunnanensis* genome assembly. For homology-based prediction, RepeatMasker v3.3.0 (RepeatMasker , RRID:SCR_012954) and RepeatProteinMasker v3.3.0 [24] were applied with parameters "-nolow -no_is -norna -engine ncbi -parallel 1 -lib RepeatMaskerLib.embl.lib " and "-engine ncbi -noLowSimple -pvalue 0.0001", respectively. The *N. yunnanensis* genome

sequence was aligned against the known repeats database, Repbase v16.10 [25], at both DNA and protein levels to identify the known repetitive elements. For *de novo* prediction, RepeatModeler v1.0.5 (RepeatModeler, RRID:SCR_015027) [26] was first executed to build a *de novo* repeat library using the *N. yunnanensis* genome assembly with parameters "-engine ncbi -name mydb". Then, RepeatMasker v3.3.0 [24] was employed to align the *N. yunnanensis* genome sequences against the *de novo* repeat library with parameters "-nolow -no_is -norna -engine ncbi -parallel 1 -lib final.library" to identify the repetitive elements. LTR_FINDER v1.05 (LTR_Finder, RRID:SCR_015247) [27] was used with parameters "-w 2 -s Athal-tRNAs.fa" for *ab initio* LTR retrotransposon finding and Tandem Repeats Finder v4.07 [28] was used with parameters "2   7   7   80   10   50   2000   -d   -h" to identify tandem repeats.

Gene prediction

The *N. yunnanensis* genome with repetitive regions masked was used to predict more genes. Protein-coding genes were predicted based on *de novo* prediction, homology search, and RNA evidence. For *de novo* prediction, Genemark-ES v4.21 (GeneMark, RRID:SCR_011930) [29] was used to carry out self-training with the default settings. To search for homologs, protein sequences of *Camptotheca acuminate* and *Arabidopsis thaliana* were used as references. For RNA evidence, a *de novo* approach was used. All of the clean RNA reads were assembled into inchworm contigs to function as expressed sequence tag evidence using Trinity v2.0.6 (Trinity, RRID:SCR_013048) [30] with the parameters "--min_contig_length 100 --min_kmer_cov 2 --inchworm_cpu 6 --group_pairs_distance 200 --no_run_chrysalis". MAKER-P v2.31 (MAKER, RRID:SCR_005309) [31] was used to perform the prediction based on the evidence above. The first round of MAKER-P was run with the "protein2genome" and "est2genome" parameter set to "1" to obtain evidence-supported gene models. SNAP [32] was then applied to train these gene models. Then, MAKER-P was run for the second round with default parameters to generate

the final consensus gene set. The search tool tRNAscan-SE v1.23 (tRNAscan-SE,

RRID:SCR_010835) [33] was used for identifying tRNA genes. The rRNA sequences

of *Arabidopsis thaliana* and *Oryza sativa* were BLAST against the *N. yunnanensis*

assembly using BLASTN (BLASTN, RRID:SCR_001598) (E-value ≤ 1e-05) to

identify rRNA genes. MicroRNAs and snRNAs were detected by searching the

sequences against the Rfam database [34] using INFERNAL (Infernal,

RRID:SCR_011809) [35] software.

Functional annotation

The predicted gene models were further functionally annotated by querying the

protein sequences against those in the public databases of Swiss-Prot [36], NCBI non-

redundant (NR),   KEGG [37], and TrEMBL with BLASTP (BLASTP,

RRID:SCR_001010) with the parameters "-e 1e-05 -a 5 -m 8 -F F" . InterProScan

v5.21 (InterProScan, RRID:SCR_005829) [38] was further used to search for the

protein motifs and domains against public domain databases including the PFAM,

PANTHER, PRINTS, PROSITE, ProDom, and SMART databases with the

parameters "-goterms -f tsv -appl Pfam -appl PRINTS -appl ProSiteProfiles -appl

ProSitePatterns -appl ProDom -appl SMART".

Results & Discussion

Assembly and annotation of the *N. yunnanensis* genome

We assembled the draft genome assembly of the highly endangered tree species *N.*

yunnanensis with DNBSEQ-G50 data from a 10X Genomics linked-reads library. The

final genome assembly was 1.475 Gb in length, which is close to the estimated

genome size of 1.64 Gb, with a scaffold N50 of 985.59 Kb and a contig N50 of 32.33

Kb (Table 1). The *N. yunnanensis* genome size we assembled was also close to the

estimated genome size of 1.23 Gb based on the raw data produced [39] for the

Digitization of the Ruili Botanical Garden project [40]. The GC content of the *N.*

yunnanensis assembly was 42.18% excluding gaps, and a total of 54.24% of the

assembly was composed of repetitive elements (Table 2). We ultimately obtained 39,803 protein-coding genes and successfully annotated 96.57% of the *N. yunnanensis* gene loci (Table 3). Non-coding genes were also annotated, identifying 175 microRNA (miRNA), 1,130 transfer RNA (tRNA), 1,502 ribosomal RNA (rRNA) and 3,106 small nuclear RNA (snRNA) genes (Table 4).

Data validation and quality control

The BUSCO analysis showed that up to 1244 (90.5%) of the expected 1375 conserved plant orthologs were detected as complete in the *N. yunnanensis* assembly and 81.9% of them were identified as complete and single-copy genes (Table 5). The RNA mapping showed that 98.95% of the reads could be successfully mapped back to the assembled genome and 83.74% of them were properly paired. The DNA mapping resulted in a 96.88% mapping rate and 86.74% of them were properly paired. These results demonstrate the high completeness of the *N. yunnanensis* assembled genome.

Potential for reuse

Here we report a draft genome assembly of the PSESP plant species *N. yunnanensis.* The completeness assessment carried out by reads mapping and BUSCO assessment indicated the high completeness of this draft assembly. As part of the 10KP (10,000 Plants) Genome Sequencing Project [41], the sequencing data and the well-annotated draft assembly generated in this study can be used for future phylogenetics and comparative genomics analyses, such as resolving the controversial phylogenetic relationships within the Nyssa genus. In particular, due to the extremely small population structure of *N. yunnanensis,* the genomic resources released in this study will support further research on the conservation biology of this highly endangered species as well as other PSESP species.

Abbreviations

10KP: 10,000 plant genome project; BUSCO: Benchmarking Universal Single-Copy Orthologs; CR: Critically Endangered; CTAB: cetyl-triethylammonium bromide; HMW: high-molecular-weight; PSESP: Plant species with extremely small populations;

Availability of supporting data

The 10X Genomics clean reads and RNA-seq clean reads are deposited in NCBI under the BioProject accession PRJNA438407, with SRA accession number SRX8345787 and SRX8373586. These reads are also deposited in the CNGB Nucleotide Sequence Archive (CNSA) with accession number CNP0001048. Genome assembly, protein-coding genes, and repeat annotations are deposited in the *GigaScience* GigaDB repository [42].

Funding

This work was supported by funding from the National Key R&D Program of China (Grant No. 2019YFC1711000), the Shenzhen Municipal Government of China (Grant No. JCYJ20170817145512476), the Shenzhen Municipal Government of China Peacock Plan (No. KQTD2015033017150531), the Guangdong Provincial Key Laboratory of Genome Read and Write (Grant No. 2017B030301011), and the Yunnan Innovation Team Program for conservation and utilization of PSESP(Plant Species with Extremely Small Populations) Ggrant No. 2019HC015). This work was also supported by the China National GeneBank.

Competing interests

The authors W.M, J.W, T.Y, Y.F, L.C, J.Y, X.X, X.L, R.D and H.L are BGI-Shenzhen employees.

Author contributions

H.L, R.D, W.S, X.L and X.X conceived and supervised the study; J.W, L.C, J.Y, R.M, J.L and J.Z prepared the samples; W.M, Y.F and T.Y analyzed the results; W.M wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Subject Areas

Genetics and Genomics; Plant Genetics, Botany

References

1.   IUCN Red List of Threatened Species. http://www.iucnredlist.org/.
2.   List of National Key Protected Wild Plants. http://www.forestry.gov.cn/yemian/minglu1.htm.
3.   Sun WB, Yang, J., Dao, Z.,. Study and Conservation of Plant Species with Extremely Small Population (PSESP) in Yunnan Province, China. Science Press, Beijing, China. 2019.
4.   Yang J, Cai L, Liu D, Chen G, Gratzfeld J and Sun W. China's conservation program on Plant Species with Extremely Small Populations (PSESP): Progress and perspectives. Biological Conservation. 2020;244   doi:10.1016/j.biocon.2020.108535.
5.   Sun B-L, Zhang C-Q, Lowry PP and Wen J. Cryptic Dioecy in *Nyssa Yunnanensis* (Nyssaceae), A Critically Endangered Species from Tropical Eastern Asia. Annals of the Missouri Botanical Garden. 2009;96 4:672-84, 13.
6.   Chen W, Shi F, Yang W, Zhou Y and Chen H. Population status and ecological characteristics of *Nyssa yunnanensis*. Journal of Northeast Forestry University. 2011;39 9:17-9, 61.
7.   Yang W-Z, Zhang S-S, Wang W-B, Kang H-M and Ma N. A sophisticated species conservation strategy for *Nyssa yunnanensis*, a species with extremely small populations in China. Biodiversity and Conservation. 2017;26 4:967-81. doi:10.1007/s10531-016-1282-8.
8.   Wang N, Milne RI, Jacques FMB, Sun B-L, Zhang C-Q and Yang J-B. Phylogeny and a revised classification of the Chinese species of Nyssa (Nyssaceae) based on morphological and molecular data. TAXON. 2012;61 2:344-54. doi:10.1002/tax.612006.
9.   Supple MA and Shapiro B. Conservation of biodiversity in the genomics era. Genome Biol. 2018;19 1:131. doi:10.1186/s13059-018-1520-3.
10.   Yang J, Wariss HM, Tao L, Zhang R, Yun Q, Hollingsworth P, et al. De novo genome assembly of the endangered *Acer yangbiense*, a plant species with extremely small populations endemic to Yunnan Province, China. Gigascience. 2019;8 7 doi:10.1093/gigascience/giz085.
11.   Yang X, Kang M, Yang Y, Xiong H, Wang M, Zhang Z, et al. A chromosome-level genome assembly of the Chinese tupelo *Nyssa sinensis*. Sci Data. 2019;6 1:282. doi:10.1038/s41597-019-0296-y.
12.   Zhao D, Hamilton JP, Pham GM, Crisovan E, Wiegert-Rininger K, Vaillancourt B, et al. De novo genome assembly of *Camptotheca acuminata*, a natural source of the anti-cancer compound camptothecin. Gigascience. 2017;6 9:1-7. doi:10.1093/gigascience/gix065.
13.   Sahu SK, Thangaraj M and Kathiresan K. DNA Extraction Protocol for Plants with High Levels

of Secondary Metabolites and Polysaccharides without Using Liquid Nitrogen and Phenol. ISRN Mol Biol. 2012;2012:205049. doi:10.5402/2012/205049.

14. Etherington GJ, Heavens D, Baker D, et al. 10x Genomics Library Construction. protocols.io 2020. https://dx.doi.org/10.17504/protocols.io.bd3vi8n6

15. Huang J, Liang X, Xuan Y, et al. BGISEQ-500 WGS library construction. protocols.io. 2018. http://dx.doi.org/10.17504/protocols.io.ps5dng6

16. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30 15:2114-20. doi:10.1093/bioinformatics/btu170.

17. Edmunds S. RNA extraction for plant samples using CTAB-pBIOZOL. protocols.io. 2017. http://dx.doi.org/10.17504/protocols.io.gsnbwde

18. Simbolo M, Gottardi M, Corbo V, Fassan M, Mafficini A, Malpeli G, et al. DNA qualification workflow for next generation sequencing of histopathological samples. PLoS One. 2013;8 6:e62692. doi:10.1371/journal.pone.0062692.

19. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv. 2013. arXiv:1308.2012v2

20. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27 5:757-67. doi:10.1101/gr.214874.116.

21. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1 1:18. doi:10.1186/2047-217X-1-18.

22. Seppey M, Manni M and Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962:227-45. doi:10.1007/978-1-4939-9173-0_14.

23. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. arXiv:1303.3997v2

24. Tarailo-Graovac M and Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009;Chapter 4:Unit 4 10. doi:10.1002/0471250953.bi0410s25.

25. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110 1-4:462-7. doi:10.1159/000084979.

26. Price AL, Jones NC and Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21 Suppl 1:i351-8. doi:10.1093/bioinformatics/bti1018.

27. Xu Z and Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35 Web Server issue:W265-8. doi:10.1093/nar/gkm286.

28. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27 2:573-80. doi:10.1093/nar/27.2.573.

29. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO and Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33 20:6494-506. doi:10.1093/nar/gki937.

30. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8 8:1494-512. doi:10.1038/nprot.2013.084.

31. Campbell MS, Holt C, Moore B and Yandell M. Genome Annotation and Curation Using

MAKER and MAKER-P. Curr Protoc Bioinformatics. 2014;48:4 11 1-39. doi:10.1002/0471250953.bi0411s48.

32.   Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59. doi:10.1186/1471-2105-5-59.

33.   Lowe TM and Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25 5:955-64. doi:10.1093/nar/25.5.955.

34.   Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43 Database issue:D130-7. doi:10.1093/nar/gku1063.

35.   Nawrocki EP, Kolbe DL and Eddy SR. Infernal 1.0: inference of RNA alignments. Bioinformatics. 2009;25 10:1335-7. doi:10.1093/bioinformatics/btp157.

36.   Bairoch A and Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28 1:45-8. doi:10.1093/nar/28.1.45.

37.   Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28 1:27-30. doi:10.1093/nar/28.1.27.

38.   Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res. 2005;33 Web Server issue:W116-20. doi:10.1093/nar/gki442.

39.   Liu H; Wei J; Yang T; Mu W; Song B; Yang T; Fu Y; Wang X; Hu G; Li W; Zhou H; Chang Y; Chen X; Chen H; Cheng L; He X; Cai H; Cai X; Wang M; Li Y; Yang J; Wang Y; Mu R; Liu J; Zhao J; Huang Z; Liu X (2019): Genomic data of Nanmaohu Park vascular plant specimen, RL0289. GigaScience Database. http://dx.doi.org/10.5524/101352

40.   Liu H, Wei J, Yang T, Mu W, Song B, Yang T, et al. Molecular digitization of a botanical garden: high-depth whole-genome sequencing of 689 vascular plant species from the Ruili Botanical Garden. Gigascience. 2019;8 4 doi:10.1093/gigascience/giz007.

41.   Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux PM, et al. 10KP: A phylodiverse genome sequencing plan. Gigascience. 2018;7 3:1-9. doi:10.1093/gigascience/giy013.

42.   Mu W, Wei J, Yang T, Fan Y, Cheng L, Yang J, Mu R, Liu J, Zhao J, Sun W, Xu X, Liu X, Drmanac R, Liu H. Genome data for the draft assembly of the Chinese tertiary relict tree, *Nyssa yunnanensis*. *GigaScience* Database. 2020. http://doi.org/10.5524/100752

Tables

Table 1: Statistics of the *N. yunnanensis* genome assembly.

Parameters	Scaffold		Contig	
	Length (bp)	**Number**	**Length (bp)**	**Number**
Maximal length (bp)	17,928,324		636,521	
N90	3,639	38,762	2,907	67,003
N80	8,911	13,177	6,795	37,603
N70	27,779	2,699	12,643	22,971
N60	308,718	527	21,401	14,910
N50	985,593	280	32,329	9,853
N40	1,699,409	166	4,5737	6,370
N30	2,587,311	96	62,625	3,875
N20	3,773,602	49	86,821	2,057
N10	6,183,248	17	127,405	771
Total length (bp)	1,474,960,449		1,330,978,457	
Number ≥ 100bp		288,519		320,818
Number ≥ 2000bp		57,709		79,919
Percentage of N content		9.76%		

Table 2. Statistics of repetitive sequences identified in the *N. yunnanensis* genome.

Category	Total repeat length (bp)	% of assembly
DNA	133,291,367	9.04%
LINE	47,813,283	3.24%
SINE	1,053,993	0.07%
LTR	690,959,337	46.87%
Tandem repeats	6,108	0.0004%
Unknown	135	0.000009%
Combined	799,507,629	54.24%

Note: DNA: DNA transposon; LINE: long interspersed nuclear element; SINE: short interspersed nuclear elements; LTR: long terminal repeat.

Table 3. Summary of protein-coding genes annotated in the *N. yunnanensis* genome.

Characteristics of protein-coding genes	
Total number of protein-coding genes	39,803
Mean gene size (bp)	2576.05
Mean CDS length (bp)	957.64
Mean exon number per gene	4.16
Mean exon length (bp)	230.26
Mean intron length (bp)	512.32
Functional annotation by searching public databases	
% of proteins with hits in Swiss-Prot database	76.01
% of proteins with hits in NCBI nr database	96.30
% of proteins with hits in KEGG database	72.15
% of proteins with hits in TrEMBL database	95.90
% of proteins with hits in Interpro database	70.99
% of proteins with functional annotation (combined)	96.57

Table 4: Statistics of predicted ncRNA in the *N. yunnanensis* genome.

Type		Number	Average length (bp)	Total length (bp)	% of genome
miRNA		175	120.8629	21,151	0.0014
tRNA		1,130	75.2221	85,001	0.0058
rRNA	**rRNA**	1,502	142.5366	214,090	0.0145
	18S	469	231.5117	108,579	0.0074
	28S	458	106.0764	48,583	0.0033
	5.8S	116	111.0603	12,883	0.0009

	5S	459	95.9586	44,045	0.0030
snRNA	**snRNA**	3,106	108.3973	336,682	0.0228
	CD-box	2,911	106.8636	311,080	0.0211
	HACA-box	51	129.3333	6,596	0.0004
	splicing	144	131.9861	19,006	0.0013

Table 5: BUSCO assessment of *N. yunnanensis* genome.

BUSCO benchmark	Number of genes	Percentage
Complete BUSCOs	1244	90.5
Complete and single-copy BUSCOs	1126	81.9
Complete and duplicated BUSCOs	118	8.6
Fragmented BUSCOs	63	4.6
Missing BUSCOs	68	4.9
Total	1375	/