 doi:10.20944/preprints202005.0417.v3

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. 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 family Nyssaceae within the order Cornales. As owning only eight individuals in two sites recorded in Yunnan province of China, this species was listed as the China's national grade-I protection species in

 doi:10.20944/preprints202005.0417.v3

1999, and also as one of 120 PSESP (Plant Species with Extremely Small Populations) in *Implementation Plan of Rescuing and Conserving China's Plant Species with extremely Small Populations(PSESP)* (2011-2-15). *N. yunnanensis* was also been evaluated as Critically Endangered in IUCN red list and Threatened Species List of China's Higher Plants. Hence understanding the genomic characteristics of this highly endangered Tertiary relict tree species is essential, especially for developing conservation strategies. 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. 799.51 Mb of the assembled genome was identified as repetitive elements, accounting for 54.24% of the sequenced genome. And a total of 39,803 protein-coding genes were annotated.

The genomic data of *N. yunnanensis* provided in this study will provide basic information for future genomic and evolutionary studies. With the genomic characteristics of *N. yunnanensis* available, our study might also facilitate in future conservation biology studies to help protecting this extremely threatened tree species.

Data Description

Nyssa yunnanensis belonging to the family Nyssaceae is an extremely threatened range-restricted tree species evaluated as Critically Endangered (CR) in the IUCN Red List of Threatened Species [1], as well as a national key protected species under grade I protection in China [2]. *N. yunnanensis* was also listed as one of 120 PSESP (Plant Species with Extremely Small Populations) in *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]. It is a canopy tree species able to reach 30m in height and functionally dioecious, consisting two types of individuals. One bearing staminate flowers while the other type bearing flowers which are morphologically normal but producing inaperturate and inviable pollen grains. *N. yunnanensis* does not appear to exhibit parthenogenesis [5].

 doi:10.20944/preprints202005.0417.v3

A survey on *N. yunnanensis* population status and ecological characteristics had been carried out, suggested that this species is under highly risk of extinction, having only two natural populations with eight individuals exist in Yunnan, China. This might due to both ecological and human factors [6]. Since 2009, an integrated PSESP conservation strategy of *N. yunnanensis* was initiated. After over seven years of implementation, the natural populations are now under secure protection. 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]. The phylogenetic study 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 evidences, the result suggested that only *N. sinensis, N. yunnanensis and N. javanica* should be recognized [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 as well as to improve conservation strategies [9]. *Acer yangbiense* is another plant species with extremely small populations endemic to Yunnan Province, which had been sequenced in 2019. The sequenced *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 genome of *Nyssa sinensiss* along with the genome of *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 of 403.17 Mb in length with an N50 scaffold size of 1753 Kb [12] respectively.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

Methods

Sample collection

The wild individual of *Nyssa yunnanensis* (NCBI: txid161873) sequenced was 70cm in height (Fig. 1), and sampled growing in Ruili, Yunan province, China. (97°56′20.99″ N, 24°03′02.72″ E, altitude 843 M). Fresh young leaf samples were collected for DNA extraction. Voucher specimens and images were collected and stored in the CNGB herbarium (Fig. 2). The extracted DNA is stored in the BGI-sample center.

DNA extraction

Leaf samples of *N. yunnanensis* were used for DNA extraction using the cetyl-triethylammonium bromide (CTAB) method [13]. Quality control was done using a Sage Science Pippin pulse electrophoresis system and high-molecular-weight (HMW) gDNA with a length of around 50kb was obtained.

Library preparation and sequencing

The HMW gDNA was loaded onto a Chromium Controller chip with 10X Chromium reagents and gel beads, and rest of the library preparation procedures were carried out according to the manufacturer's protocol [14]. Subsequently the sequencing was performed on a BGISEQ-500 platform according to the manufacturer's instructions [15] using the whole-genome shotgun sequencing strategy and a total of 1.64 Gb of raw data (150 bp, paired-end) was eventually generated which covered about 100× of the 1.64Gb estimated genome size.

Genome size estimation

The raw data of the *N. yunnanensis* was trimmed with Trimmomatic-0.38 [16] with the parameters "ILLUMINACLIP: 2:35:4 HEADCROP:5 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:15 MINLEN:50". The 1.64 Gb *N. yunnanensis* genome size was then estimated by k-mer frequency analysis software gce-1.0.0 (GCE, RRID:SCR_017332) [17] with the clean data. In addition, the genome size estimation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

performed automatically by the supernova-2.0.0 assembler software (Supernova assembler, RRID:SCR_016756) [18] resulted in an estimated genome size of 1.43 Gb.

De novo genome assembly

De novo assembly was carried out using supernova-2.0.0 software [18] with the "--maxreads 691040000" parameter. Linked read data without trimming were used as the software recommended. Then the gaps within the scaffolds were filled by GapCloser version 1.12 (GapCloser, RRID:SCR_015026) [19] with the parameters "-l 150 -t 32" using barcode trimmed pair-end reads.

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 [20] were applied, aligning the *N. yunnanensis* genome sequences against the Repbase v16.10 [21] and to identify the known repetitive elements. For *de novo* prediction, RepeatModeler v1.0.5 (RepeatModeler, RRID:SCR_015027) [22] was first executed to build a de novo repeat library and using the *N. yunnanensis* genome assembly. Then RepeatMasker v3.3.0 [20] was employed to align the *N. yunnanensis* genome sequences against the *de novo* repeat library to identify the repetitive elements. LTR_FINDER v1.05 (LTR_Finder, RRID:SCR_015247) [23] was used for *ab initio* LTR retrotransposon finding and Tandem Repeats Finder v4.07 [24] was used for Tandem repeats identification respectively.

Gene prediction

The *N. yunnanensis* genome with repetitive regions masked were further used to carry out gene prediction. Protein-coding genes were predicted based on both homology and *de novo* prediction evidence. Protein sequences of *Camptotheca acuminate* and *Arabidopsis thaliana* were used as the homology evidence, while self-training was done using Genemark-ES v4.21 (GeneMark, RRID:SCR_011930) [25]. MAKER-P v2.31

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

(MAKER, RRID:SCR_005309) [26] was used to perform the prediction on the basis of the evidence above. The first round of MAKER-P was run with the "protein2genome" parameter set to "1" to obtain protein-supported gene models. SNAP [27] was then applied to train with these gene models. Then MAKER-P was run for the second round, combining all the results to generate the final gene models with default parameters.

Functional annotation

The predicted gene models were further functionally annotated by searching the protein sequences against public databases of Swiss-Prot [28], NCBI non-redundant (NR), KEGG [29] and TrEMBL with BLASTP (BLASTP, RRID:SCR_001010) (E-value ≤ 1e-05). InterProScan v5.21 (InterProScan, RRID:SCR_005829) [30] was further used to search the protein motifs and domains against public domain databases including the PFAM, PANTHER, PRINTS, PROSITE, ProDom, and SMART databases. tRNAscan-SE v1.23 [31] was used for tRNA genes identification and 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) for rRNA genes identification respectively. miRNAs and snRNAs were annotated by searching the sequences against the Rfam database [32] using INFERNAL (Infernal, RRID:SCR_011809) [33] software.

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 BGISEQ-500 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, respectively. The *N. yunnanensis* genome size we assembled is also close to the estimated genome size of 1.23 Gb from the digitization of Ruili Botanical Garden project [34]. 9.76% of the genome regions were presented as Ns (Table 1). The GC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

content of the *N. yunnanensis* assembly excluding gaps was 42.18%, 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 the functional annotation successfully annotated 96.57% of the *N. yunnanensis* gene loci (Table 3). Non-coding genes were also annotated, identifying 175 microRNA, 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 completeness of *N. yunnanensis* assembly was estimated with two strategies. Firstly, we performed the completeness assessment using BUSCO ((BUSCO; v3.0.2, RRID:SCR_015008)) [35] with Embryophyta_odb10 database. The result 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). Secondly, the RNA of the *N. yunnanensis* was extracted and sequenced generating 8.85Gb raw data, which was further filtered using SOAPfilter v2.2 (SOAP, RRID:SCR_000689) with following parameters "-q 33 -i 200 -g 1 -M 2 -Q 20". Then all the clean reads were aligned to the *N. yunnanensis* assembly using BWA-MEM (BWA, version 0.7.16, RRID:SCR_010910) [36] with default parameters. In total, 98.95% of the reads could be successfully mapped back to the assembled genome assembly and 83.74% of them were properly paired. These results demonstrated the high completeness of the *N. yunnanensis* assembled genome.

Re-use potential

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 a part of the 10KP (10,000 Plants) Genome Sequencing Project [37], 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

 doi:10.20944/preprints202005.0417.v3

population structure of *N. yunnanensis,* the genomic resources released in this study will be supportive for further researches on the conservation biology studies of this highly endangered species, as well as other PSESP species.

Abbreviations

PSESP: Plant species with extremely small populations; CR: Critically Endangered; CTAB: cetyl-triethylammonium bromide; HMW: high-molecular-weight; BUSCO: Benchmarking Universal Single-Copy Orthologs

Availability of supporting data

The raw sequencing reads are deposited in NCBI under the BioProject accession PRJNA438407, with SRA accession number SRX8345787 and SRX8373586. The raw reads are also deposited in the CNGB Nucleotide Sequence Archive (CNSA) with accession number CNP0001048. Genome assembly, protein-coding gene and repeat annotations are deposited in the *GigaScience* GigaDB [38].

Funding

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

Competing interests

The authors declare that they have no competing interests.

 doi:10.20944/preprints202005.0417.v3

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 the inputs from all authors. All authors read and approved the final manuscript.

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, Sun W: **China's conservation program on Plant Species with Extremely Small Populations (PSESP): Progress and perspectives**. *Biological Conservation* 2020, **244**.
5. Sun B-L, Zhang C-Q, Lowry PP, 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):672-684, 613.
6. Chen W, Shi F, Yang W, Zhou Y, Chen H: **Population status and ecological characteristics of Nyssa yunnanensis**. *Journal of Northeast Forestry University* 2011, **39**(9):17-19, 61.
7. Yang W-Z, Zhang S-S, Wang W-B, Kang H-M, Ma N: **A sophisticated species conservation strategy for Nyssa yunnanensis, a species with extremely small populations in China**. *Biodiversity and Conservation* 2017, **26**(4):967-981.
8. Wang N, Milne RI, Jacques FMB, Sun B-L, Zhang C-Q, 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):344-354.
9. Supple MA, Shapiro B: **Conservation of biodiversity in the genomics era**. *Genome Biol* 2018, **19**(1):131.
10. Yang J, Wariss HM, Tao L, Zhang R, Yun Q, Hollingsworth P, Dao Z, Luo G, Guo H, Ma Y *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).
11. Yang X, Kang M, Yang Y, Xiong H, Wang M, Zhang Z, Wang Z, Wu H, Ma T, Liu J *et al*: **A chromosome-level genome assembly of the Chinese tupelo Nyssa sinensis**. *Sci Data* 2019, **6**(1):282.
12. Zhao D, Hamilton JP, Pham GM, Crisovan E, Wiegert-Rininger K, Vaillancourt B,

DellaPenna D, Buell CR: **De novo genome assembly of Camptotheca acuminata, a natural source of the anti-cancer compound camptothecin**. *Gigascience* 2017, **6**(9):1-7.

13. Sahu SK, Thangaraj M, 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.
14. **10X Genomics Genome Reagent Kit User Guide** [https://support.10xgenomics.com/de-novo-assembly/sample-prep/doc/user-guide-chromium-genome-reagent-kit-v1-chemistry]
15. **BGISEQ-500 WGS library construction**
16. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
17. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W: **Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects**. *arXivorg* 2013.
18. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB: **Direct determination of diploid genome sequences**. *Genome Res* 2017, **27**(5):757-767.
19. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y *et al*: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler**. *Gigascience* 2012, **1**(1):18.
20. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences**. *Curr Protoc Bioinformatics* 2009, **Chapter 4**:Unit 4 10.
21. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements**. *Cytogenet Genome Res* 2005, **110**(1-4):462-467.
22. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes**. *Bioinformatics* 2005, **21 Suppl 1**:i351-358.
23. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W265-268.
24. Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic Acids Res* 1999, **27**(2):573-580.
25. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm**. *Nucleic Acids Res* 2005, **33**(20):6494-6506.
26. Campbell MS, Holt C, Moore B, Yandell M: **Genome Annotation and Curation Using MAKER and MAKER-P**. *Curr Protoc Bioinformatics* 2014, **48**:4 11 11-39.
27. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:59.
28. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Res* 2000, **28**(1):45-48.
29. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
30. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W116-120.
31. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA**

genes in genomic sequence. *Nucleic Acids Res* 1997, **25**(5):955-964.

32. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J *et al*: **Rfam 12.0: updates to the RNA families database**. *Nucleic Acids Res* 2015, **43**(Database issue):D130-137.
33. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments**. *Bioinformatics* 2009, **25**(10):1335-1337.
34. Liu H, Wei J, Yang T, Mu W, Song B, Yang T, Fu Y, Wang X, Hu G, Li W *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).
35. Seppey M, Manni M, Zdobnov EM: **BUSCO: Assessing Genome Assembly and Annotation Completeness**. *Methods Mol Biol* 2019, **1962**:227-245.
36. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM**. *arXivorg* 2013.
37. Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux PM, Li FW, Melkonian B, Mavrodiev EV, Sun W *et al*: **10KP: A phylodiverse genome sequencing plan**. *Gigascience* 2018, **7**(3):1-9.
38. W M, J W, T Y, Y F, L C, J Y, R M, J L, J Z, X X *et al*: **Genome data for the draft assembly of the Chinese tertiary relict tree, Nyssa yunnanensis**. In: *GigaScience Database.* 2020.

Figures

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

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

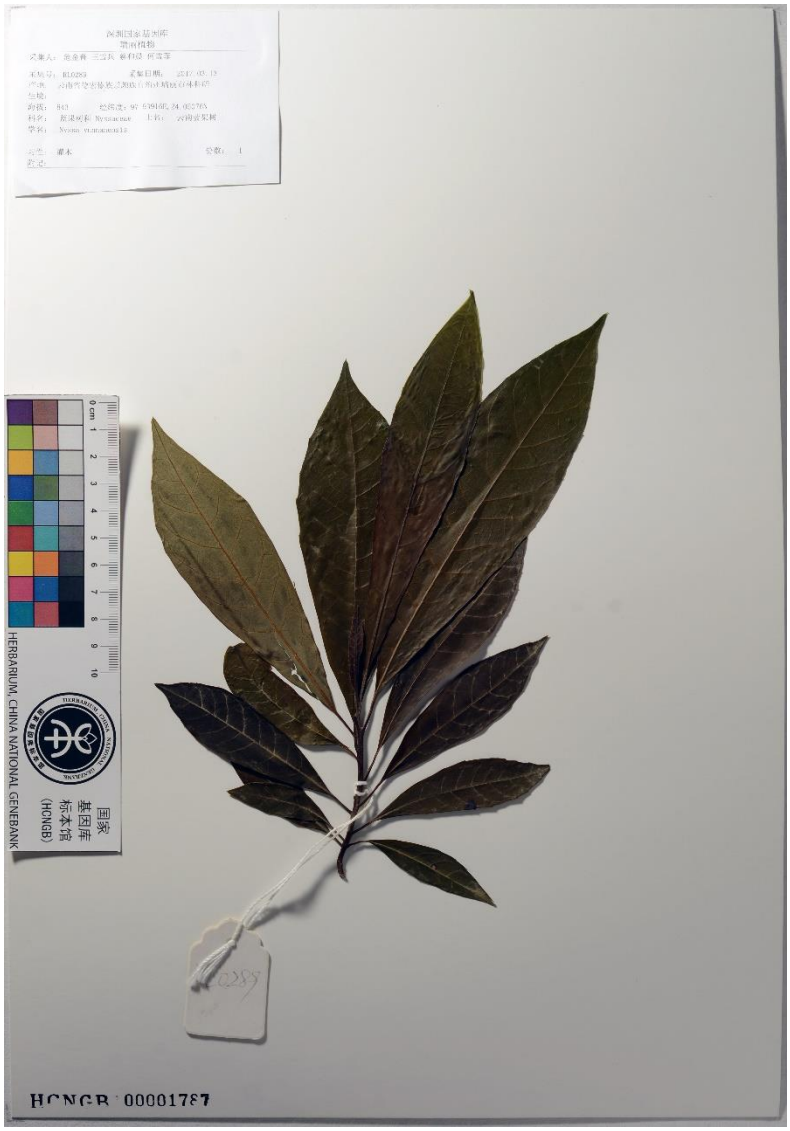

Figure 2. Photograph of the voucher specimens of *Nyssa yunnanensis*, stored in the CNGB herbarium (voucher RL0289)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

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%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2020 doi:10.20944/preprints202005.0417.v3

	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	/