Supplementary notes for

Transcript assembly, phylogenetic tree construction and test datasets generation

Jun-Jie Wu[a,†], Yu-Wei Han[b,†], Chen-Feng Lin[a,†], Jing Cai[b,*], Yun-Peng Zhao[a,*]

[a]Systematic & Evolutionary Botany and Biodiversity Group, MOE Key Laboratory of Biosystem Homeostasis and Protection, College of Life Sciences, Zhejiang University, Hangzhou 310058, China

[b]School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710129, China

This PDF file includes:

De novo transcriptome assembly

The 204 open gymnosperm transcriptomes downloaded from NCBI were initially in binary format (**Table S2**), which was converted to fastq format using the "fasterq-dump" function in sratoolkit (v3.0.1, https://github.com/ncbi/sra-tools). *De novo* assembly was performed using the Trinity (v2.10.0)[1] with default parameters for each of the 204 transcriptomes. cd-hit (v 4.8.1, https://github.com/weizhongli/cdhit) was then used to filter transcripts by clustering and comparing their nucleotide sequences. Finally, the set of coding genes predicted by the transcriptome and the set of translated proteins are obtained using TransDecoder (v 5.5.0, https://github.com/TransDecoder/). In addition, to ensure the completeness of the transcript, sequences featuring a stop codon within the middle of the sequence (early termination), lacking a start codon, or lacking a stop codon at the final position, were meticulously removed.

Phylogenetic analyses

For concatenation-based maximum likelihood (ML) tree inference, the proteins of 204 gymnosperm transcriptomes and 7 gymnosperm genomes were evaluated respectively in Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.4.7), and we used 1,603 BUSCO genes from our gymnosperm benchmark gene set as a set of reliable markers of diverse lineages. Multiple sequence alignments were performed by MAFFT (v7.310)[2] with "--auto" parameter. We used trimAl[3] (v1.4) to trim poorly alignment with the "-automated1" parameter. IQ-TREE[4] (v 2.1.4) was used to construct maximum likelihood trees with the "-m MF" and "-B 1000" parameters. Finally, the coalescence-based phylogeny was inferred by ASTRAL-III (v5.7.1)[5].

Generation of test datasets

To assess the reliability of the gymnosperm benchmark gene set, we generated multiple datasets for BUSCO analyses by randomly deleting different proportions of input sequences. To save computing time, the genome of *Gnetum montanum* was used in subsequent analysis[6,7]. We randomly removed 10/30/50% of the annotated protein sequences or the single copy orthologs by "sample" function in SeqKit[8] with different seed, and masked the corresponding regions of the genome by coding as missing

information (N) using “maskfasta” in BEDTools (v2.31.0)[9].

References:

1. Haas, B.J. et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494-512 (2013).
2. Katoh, K., Misawa, K., Kuma, K.I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066 (2002).
3. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
4. Minh, B.Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530-1534 (2020).
5. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinform.* **19**, 153 (2018).
6. Wan, T. et al. The *Welwitschia* genome reveals a unique biology underpinning extreme longevity in deserts. *Nat. Commun.* **12**, 4247 (2021).
7. Wan, T. et al. A genome for gnetophytes and early evolution of seed plants. *Nature Plants* **4**, 82-89 (2018).
8. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**, e0163962 (2016).
9. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).