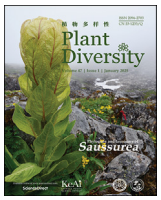
Plant Diversity 47 (2025) 115–126

Contents lists available at ScienceDirect

Plant Diversity

journal homepage: http://www.keaipublishing.com/en/journals/plant-diversity/
http://journal.kib.ac.cn



Research paper

Assembly-free reads accurate identification (AFRAID) approach outperforms other methods of DNA barcoding in the walnut family (Juglandaceae)

Yanlei Liu [a, 1], Kai Chen [a, 1], Lihu Wang [a, 1], Xinqiang Yu [a], Chao Xu [b], Zhili Suo [b], Shiliang Zhou [b, *], Shuo Shi [c, ***], Wenpan Dong [d, **]

[a] School of Landscape and Ecological Engineering, Hebei University of Engineering, Handan 056038, China
[b] State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[c] College of Life Science, Hebei Normal University, Shijiazhuang 050024, China
[d] School of Ecology and Nature Conservation, Beijing Forestry University, Beijing 100083, China

ARTICLE INFO

Article history:
Received 16 February 2024
Received in revised form
30 September 2024
Accepted 10 October 2024
Available online 16 October 2024

Keywords:
DNA barcode
Species identification
Random DNA barcode
Juglandaceae
Assembly-free

ABSTRACT

DNA barcoding has been extensively used for species identification. However, species identification of mixed samples or degraded DNA is limited by current DNA barcoding methods. In this study, we use plant species in Juglandaceae to evaluate an assembly-free reads accurate identification (AFRAID) method of species identification, a novel approach for precise species identification in plants. Specifically, we determined (1) the accuracy of DNA barcoding approaches in delimiting species in Juglandaceae, (2) the minimum size of chloroplast dataset for species discrimination, and (3) minimum amount of next generation sequencing (NGS) data required for species identification. We found that species identification rates were highest when whole chloroplast genomes were used, followed by taxon-specific DNA barcodes, and then universal DNA barcodes. Species identification of 100% was achieved when chloroplast genome sequence coverage reached 20% and the original sequencing data reached 500,000 reads. AFRAID accurately identified species for all samples tested after 500,000 clean reads, with far less computing time than common approaches. These results provide a new approach to accurately identify species, overcoming limitations of traditional DNA barcodes. Our method, which uses next generation sequencing to generate partial chloroplast genomes, reveals that DNA barcode regions are not necessarily fixed, accelerating the process of species identification.

1. Introduction

Species is the basic unit of biological taxonomic research (Simpson, 1951). However, plant species are commonly misidentified (Le et al., 2020; Shirai et al., 2022), partly due to difficulties in understanding the species concept or a lack of reliable identification methodologies (Cracraft, 1983; Hong, 2016; Liu, 2016). DNA barcoding, which relies on a short, fixed, easily amplified DNA sequence to distinguish species (Hebert et al., 2003), has been extensively used in species identification (Kress et al., 2005; Barberán et al., 2015; Liu et al., 2018; Lv et al., 2023; Duan et al., 2024), new species discovery (Liu et al., 2011; Tyagi et al., 2019), and biodiversity assessment (Hajibabaei et al., 2007; McFadden et al., 2014; Barberán et al., 2015; Liu et al., 2023). Researchers have long sought to find a universal plant DNA barcode (Kress and Erickson, 2007; Hollingsworth et al., 2009, 2011; Li et al., 2011a; Yu et al., 2011; Dong et al., 2014; Xu et al., 2015), however, candidates for this type of barcode have provided insufficient resolution, and alternative approaches are needed.

Several strategies have been proposed to overcome limitations of conventional DNA barcoding sequences, including combinations of multiple DNA fragments, whole chloroplast genomes, the so-

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: slzhou@ibcas.ac.cn (S. Zhou), shishuo@hebtu.edu.cn (S. Shi), wpdong@bjfu.edu.cn (W. Dong).
Peer review under the responsibility of Editorial Office of Plant Diversity.
[1] Yanlei Liu, Kai Chen and Lihu Wang were contributed equally in this paper.

https://doi.org/10.1016/j.pld.2024.10.002

called super barcode (Li et al., 2015; Chen et al., 2018; Zhang et al., 2019, 2021; Wu et al., 2021), and taxon-specific DNA barcodes (Selvaraj et al., 2015; Dong et al., 2021; Zhang et al., 2021; Govender et al., 2022). However, these DNA barcoding methods are impractical when trying to identify species from highly degraded or mixed samples due to the probability of PCR amplification or sanger sequencing failures. Next generation sequencing (NGS) technology has closed the gap between conventional DNA barcodes and super DNA barcodes and is able to overcome some of these limitations. For example, NGS has been used to identify species from highly degraded specimens and environmental samples (Galan et al., 2012; Shokralla et al., 2014; Xu et al., 2015; Prosser et al., 2016). In addition, NGS has lowered the cost and increased the efficiency of DNA barcoding reference library construction (Liu et al., 2021). Regrettably, the application of DNA barcoding with NGS technology has been rare, and critical technical inquiries, such as determining the optimal data size for collection, remain unresolved.

Here, we propose a new DNA barcoding method for plants. We evaluated this new approach by testing whether it accurately identifies species in Juglandaceae, world-renowned for walnuts and pecans (Guo et al., 2020; Zhou et al., 2021). Juglandaceae has about 60 species in nine genera (Song et al., 2020; Zhou et al., 2021): *Alfaroa* (five species), *Carya* (about 15 species, including *Annamocarya*), *Cyclocarya* (one species), *Engelhardia* (about 15 species, including monospecific *Alfaropsis*), *Juglans* (about 20 species, including *Wallia*), *Oreomunnea* (one species), *Platycarya* (two species), *Pterocarya* (about eight species), and *Rhoiptelea* (one species). The evolution of the Juglandaceae remains a difficult problem, hypothesized to have both ancient and recent speciation, extinctions and radiations (Lu, 1982; Manchester, 1989; Zhou et al., 2021; Zhang et al., 2022). This evolutionary history, as well as the size of the family, make it ideal for testing a new DNA barcoding method (Zhou et al., 2021; Yang et al., 2023; Yan et al., 2024).

In this study, we use plant species in Juglandaceae to evaluate an Assembly-Free Reads Accurate IDentification (AFRAID) method of species identification, a novel approach for precise species identification in plants. Specifically, we determined (1) the accuracy of DNA barcoding approaches in delimiting species in Juglandaceae, (2) the minimum size of chloroplast dataset for species discrimination, and (3) minimum amount of NGS data required for species identification.

2. Materials and methods

2.1. Data preparation

A total of 119 whole chloroplast genome sequences, representing 54 species of all nine genera in Juglandaceae, were used in this study. A total of 91 whole chloroplast genome sequences from NCBI were included (Table S1). In addition, we newly sequenced 28 whole chloroplast genomes (Fig. 1). All specimens were collected in collaboration with the Herbarium of the Institute of Botany, Chinese Academy of Sciences (PE). The same plant materials have been accurately identified previously using morphological traits (Table S1). Total genomic DNA was extracted following Li et al. (2013), purified by a Wizard DNA cleanup system, quantified by spectrophotometry, and checked using a 1.5% (w/v) agarose gel.

Total DNA was fragmented to about 350 bp by ultrasound. A paired-end library was constructed for each sample using the NEBNext Ultra™ DNA library prep kit. PE150 sequencing was performed on the Illumina HiSeq X Ten platform. NGS QC toolkit was used for quality control and to filter low-quality reads (Patel and Jain, 2012). Chloroplast genomes were assembled following the method of Dong et al. (2022). Contigs were assembled from the quality-controlled paired-end reads by using the SPAdes v.3.6.1 program (Bankevich et al., 2012). Chloroplast genome contigs were picked out by the Blast2$^+$ program using the chloroplast genome of *Juglans regia* as a reference (Altschul et al., 1990). Chloroplast contigs were assembled into chloroplast genomes using Sequencher v.5.4.5 (Gene Codes Corporation, Ann Arbor, MI, USA). Chloroplast genomes were annotated with Plann (Huang and Cronk, 2015) using *J. regia* as a reference, and the missing genes or errors were checked manually according to the results of Geneious Prime 2022.2.1 (GraphPad Software, LLC, USA).

2.2. Phylogenetic analysis of chloroplast genome data

Phylogenetic analysis based on the whole chloroplast genome dataset was conducted using maximum likelihood (ML) and Bayesian inference (BI) methods, with four samples of *Morella* as outgroups. Whole chloroplast genome sequences were aligned using the MAFFT online version (Katoh and Standley, 2013) and the unreliably aligned regions were trimmed with Gblocks v.0.91b (Castresana, 2000). The optimal model TVM + F + I + G4 was calculated by ModelFinder based on the BIC standard (Kalyaanamoorthy et al., 2017). ML analysis was performed using the IQ-tree (Nguyen et al., 2015), and sampling was repeated 1000 times. Bayesian inference was conducted with Phylosuite (Zhang et al., 2020) and two independent Markov chain Monte Carlo (MCMC) analyses were performed in MrBayes (Ronquist et al., 2012), each with four chains for 20,000,000 generations, sampling every 1000 trees. Each chain started with a random tree. A majority rule consensus tree was constructed after discarding the first 25% of sampled trees as burn in. Posterior probabilities (PP) were computed to estimate the reliabilities of branches.

2.3. Comparative analysis of DNA barcodes

The nucleotide diversity of the chloroplast genome was calculated based on the sliding window analysis using DnaSP v.6.12.03 software (Librado and Rozas, 2009). The window length was set to 600 bp with a 200 bp step size. Juglandaceae taxon-specific DNA barcodes were selected according to nucleotide diversity calculated using DnaSP v.6.12.03. Then, conventional DNA barcodes (*matK*, *rbcL*, *trnH-psbA*, and *trnL*-intron) and taxon-specific DNA barcodes were extracted from the aligned whole chloroplast genome data. We compared the ability of conventional DNA barcodes and taxon-specific barcodes to identify species by using the tree-building methods described above, which were implemented in PhyloSuite (Zhang et al., 2020).

2.4. Determination of minimum chloroplast genome sequence length for species identification

To determine the minimum chloroplast genome sequence length for species discrimination, we divided the chloroplast genome dataset into sub-datasets of 100 bp and then randomly sampled the subsets from 5% to 95% at a 5% interval using random-sampling.py (https://github.com/Mycroft-behind/random-sampling/tree/main). The samplings were repeated six times. The sampled sub-datasets were concatenated and phylogenetic trees were constructed using the ML method (Zhang et al., 2020). ML trees were manually compared and the species discrimination rates (ratio of species discriminated to the total species*100%) were calculated. The success of species identification was based on two criteria: (1) Whether different individuals of the same species cluster together? If they do, species identification is considered successful; if not, it was deemed unsuccessful; (2) For species represented by a single individual, the evaluation focused on the differentiation from other species

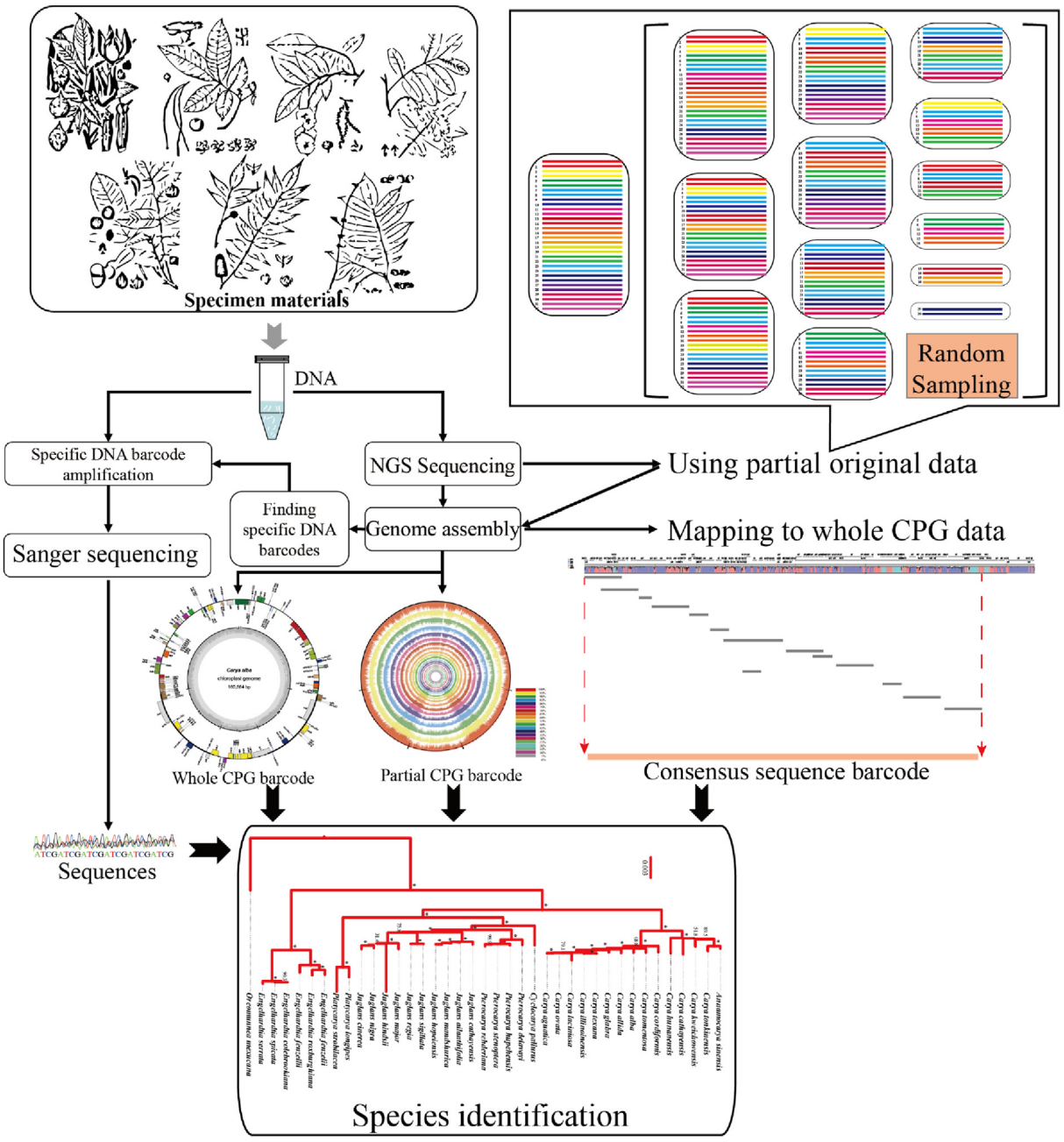


Fig. 1. Diagram of experimental design. DNA was extracted from the Juglandaceae plant materials. Following next generation sequencing, samples were used for data assembly or to determine the minimum amount of data necessary to assemble a chloroplast genome capable of identifying Juglandaceae species. For data assembly, the assembled genome was either merged with existing data to develop Juglandaceae taxon-specific DNA barcodes, utilized for the identification of Juglandaceae species, or to explore the proportion of chloroplast genome data required for Juglandaceae species identification.

within the phylogenetic tree. This differentiation is primarily reflected in branch length, which refers to the length of a branch in a phylogenetic tree, typically representing the amount of evolutionary change that has occurred along that branch. If all branches are relatively short, it becomes challenging to identify these species. However, if the branch lengths exhibit sufficient variation, precise species identification can be achieved.

2.5. Species identification using incomplete chloroplast genome

In practice, genome sequences of both query samples and the reference library are commonly incomplete. To understand the effects of NGS sequencing depth on species discrimination power, we resampled data from 27 NGS clean reads (raw data after NGS QC Toolkit quality control) (belonging to 27 undoubtable species, Table S3 which contains nine samples from public data and 18 samples from this study) in gradient and assembled draft chloroplast genomes. Read sampling was conducted with Geneious Prime 2022.2.1 (Kearse et al., 2012). A total of 100,000 to 1,000,000 reads were sampled from paired clean reads (about 14 million reads each sample) at increments of 100,000 reads. The corresponding whole chloroplast genomes were used as controls. Draft chloroplast genome assembly followed the same methods as that of chloroplast genome assembly. Gaps (or holes) were treated as “missing”

(marked as "?"). Incomplete chloroplast genomes were checked using the assembly function in Geneious Prime 2022.2.1 and confirmed by blast2^{+} against the 27 corresponding chloroplast genomes (Table S4).

2.6. Species identification using assembly-free clean reads

To determine whether assembly-free clean NGS reads can correctly identify species, a total of 500,000 NGS clean reads from 27 species were used to blast search 97 Juglandaceae whole chloroplast genomes. To reduce the computational complexity of the data, the blast algorithm was used to first filter out the chloroplast genome-related reads based on relatively loose parameters. Results were analyzed using the following two parameters. First, we considered the number of reads. If only one reference had the highest bit-score value, only this reference was retained and treated as the final blast result. If several references had the same highest bit-score, these references were retained and treated as the final results, and were marked as one. Second, to determine whether the correct species received the highest bit-score value, bit-score values of each retained read were summed (Fig. 2). Statistical analysis of blast results mainly utilized AFRAID (Assembly-Free Reads Accurate IDentification), which was developed for this study (https://github.com/Mycroft-behind/classify/tree/main).

Currently, assembly-free NGS data is routinely used to calculate genetic distances between samples, using algorithms such as MIKE and Skmer (Sarmashghi et al., 2019; Wang et al., 2024). In this study, we used both MIKE and Skmer to calculate genetic distances in the 500,000 NGS clean reads (PE150) dataset of 27 species (the same samples as used in AFRAID) (Table S3). We followed the workflow recommended by the two algorithms. Each algorithm was repeated 100 times to obtain its supporting rate. The results were manually compared with the interspecies Jaccard genetic relationships obtained using the whole chloroplast genome in Mega X (Kumar et al., 2018).

3. Results

3.1. Variation and phylogeny of Juglandaceae based on whole chloroplast genomes

Analysis of 119 Juglandaceae chloroplast genomes indicated that Juglandaceae genome lengths range from 158,223 to 161,713 bp, the LSC region from 87,898 to 91,058 bp, the IRa region from 24,029 to 26,242 bp, and the SSC region from 18,174 to 20,554 bp (for details, see Table S1). Furthermore, the average number of tRNAs was 39, the average number of CDS was 86, and the average number of rRNAs was eight (Table S2).

We constructed a phylogenetic tree of Juglandaceae primarily to verify the reliability of chloroplast genome data and to assess the feasibility of using the complete chloroplast genome for the identification of Juglandaceae species. Phylogenetic analysis based on chloroplast genome data confirmed the monophyly of five genera in Juglandaceae (*Cyclocarya*, *Juglans*, *Platycarya*, *Rhoiptelea* and *Pterocarya*) (Fig. 3).

Most species were monophyletic, with a few exceptions, i.e., *Carya*, *Engelhardia*, *Juglans* and *Pterocarya*. We also found that whole chloroplast genomes are feasible for species identification within Juglandaceae. Specifically, whole chloroplast genomes could be used to identify species when we clustered the same species within the entire phylogenetic tree and examined genetic differences (branch lengths) between different species.

The overall nucleotide diversity (π) was 0.00077 across whole chloroplast genomes. However, different regions of chloroplast genomes exhibited considerable variation in nucleotide diversity. The SSC region exhibited the highest π value, and the IR had the lowest (Fig. S1a). In total, eight regions in the Juglandaceae chloroplast genomes were identified with $\pi > 0.04$: *matK-rps16*, *rps16-trnQ*, *trnE-trnT*, *ndhE-ndhG*, *ycf1*, *trnE-trnT*, *ndhA-intron*, and *rrn5-rrn4.5* (Fig. S1a). Similarly, hypervariable regions were estimated at the generic level (Fig. S1). The most highly variable regions in *Pterocarya* were *ycf1-ndhF* and *trnN-trnR* (Fig. S1b, $\pi > 0.02$). In *Platycarya*, the most variable regions were *trnT-psbD*, *trnL*-intron, *ndhC-trnV*, *accD-psaI*, *petA-psbJ*, *rps7-trnV*, *ycf1-ndhF*, *ndhF*, *ycf1*, and *rrn5-rrn4.5* (Fig. S1c, $\pi > 0.05$). In *Juglans*, the highly variable regions were *trnS-trnG*, *psbZ-trnG*, *trnT-psbD*, *accD-psaI*, *ycf1-ndhF*, *ndhD-psaC*, *ndhA-intron*, and *ycf1* (Fig. S1d, $\pi > 0.0025$). In *Engelhardia,* the highly variable regions were *matK-rps16*, *rps16-trnQ*, *trnS-trnG*, *trnE-trnT*, *trnF-ndhJ*, *accD-psbI*, and *rpl32-trnL* (Fig. S1e, $\pi > 0.05$). Lastly, the highly variable regions within *Carya* were found to be *trnK-rps16*, *ndhC-trnV*, *atpF-atpH*, *trnE-trnT*, *trnM-atpE*, *rpl32-trnL*, *ndhA*-intron, and *rrn5-rrn4.5* (Fig. S1f, $\pi > 0.05$). Taken together, the following regions of the chloroplast genome are variable in more than one genus and may serve as DNA barcodes for Juglandaceae: *accD-psaI*, *ndhA-intron*, *ndhC-trnV*, *rpl32-trnL*, *rrn5-rrn4.5*, *trnE-trnT*, *trnS-trnG*, *trnT-psbD*, *ycf1*, and *ycf1-ndhF*.

3.2. Assessment of DNA barcodes

After removal of incorrectly identified samples and correction of non-monophyletic genera as indicated in Fig. 3, we used the ML tree to assess the ability of the following conventional DNA barcodes (i.e., *rbcL*, *matK*, *trnH-psbA*, and *trnL*-intron) to identify species. The *rbcL* barcode identified 9.6% of species (nine samples and five species, Fig. S2); *matK* identified 39.4% (37 samples and 21 species, Fig. S3); *psbA-trnH* identified 18.1% (17 samples and 11 species, Fig. S4); and *trnL*-intron identified 6.4% (six samples and four species, Fig. S5). The concatenated data of all universal DNA barcodes were able roughly resolve 34.0% of species (32 samples and 21 species, Fig. S6).

When using taxon-specific DNA barcodes at the generic level, the eight *Carya*-specific DNA barcodes resolved 75% of species (12/16, Fig. S7). In *Engelhardia*, the ML tree constructed by combining seven newly discovered highly variable DNA barcode regions showed that all samples could be distinguished according to differences in branch length, with a success rate of 100% (Fig. S8). In *Juglans*, the ML tree was constructed jointly with eight newly discovered highly variable DNA barcode regions. All samples were identified based on differences in branch length (including 36 samples and nine species, Fig. S9). In *Platycarya*, ML trees constructed by combining ten newly discovered highly variable DNA barcode regions identified five samples (Fig. S10). In *Pterocarya*, the ML tree constructed by combining two newly discovered highly variable DNA barcoding regions identified only three species successfully, with a success rate of 18.75% (3/16, Fig. S11). However, even for genera where complete identification can be achieved using multiple taxon-specific DNA barcodes, it is still not possible to achieve complete species identification within the genus using a single highly variable DNA barcode. Furthermore, the DNA barcodes of eight highly variable taxa at the family level of Juglandaceae were clearly distinguishable between genera, except for *Oreomunnea*, which was embedded in *Engelhardia*. Combined Juglandaceae taxon-specific DNA barcodes identified 87.63% (85) of samples and 84% (42) of species. The unidentified species were mainly distributed in *Peterocarya* (Fig. S12).

3.3. Assessment of partial super DNA barcodes

The super DNA barcode using the whole chloroplast genomes resolved all species except *Pterocarya*, a genus in which chloroplast

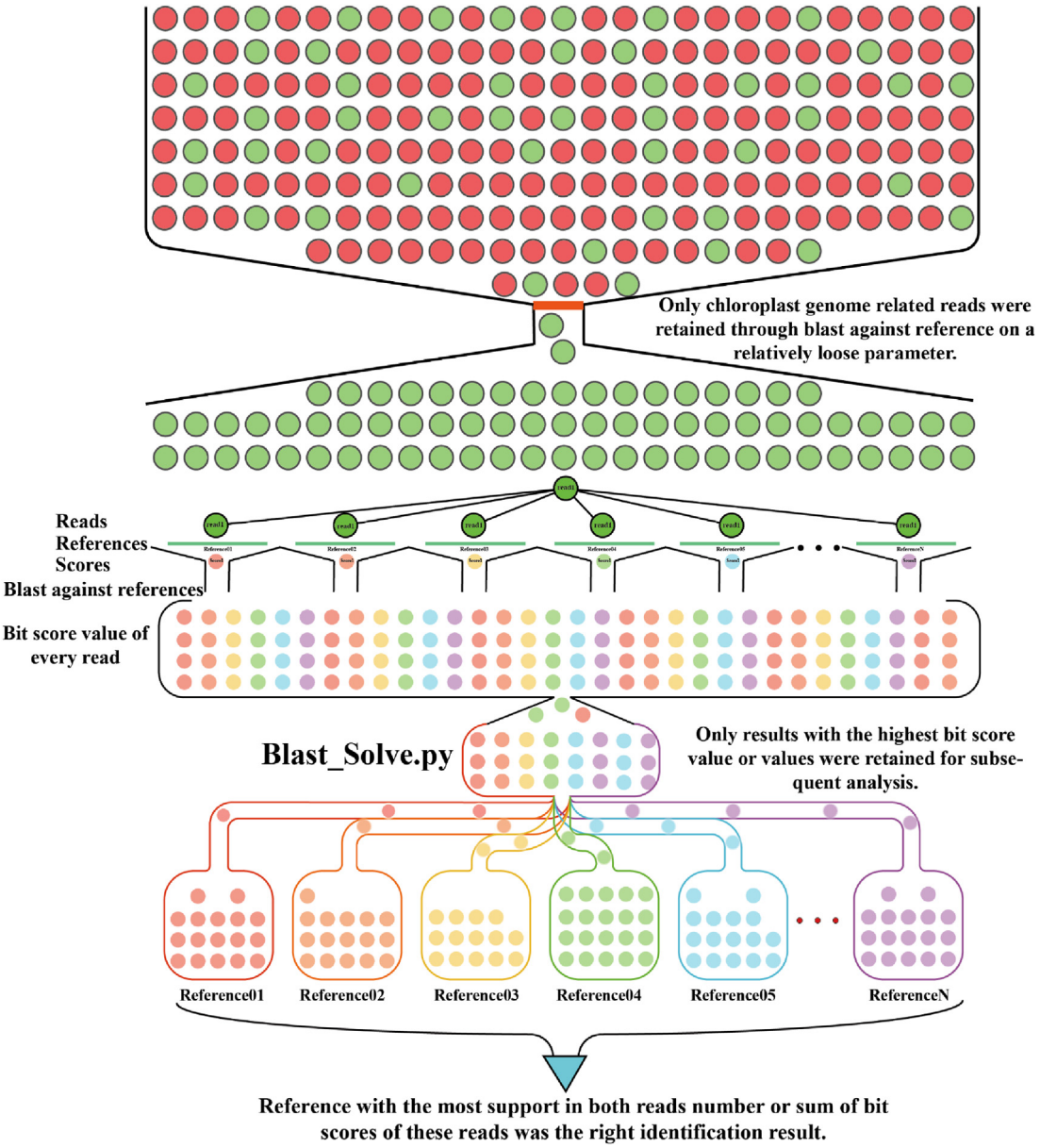


Fig. 2. Assembly-free reads accurate identification (AFRAID) for plant species identification. To conserve computational resources, AFRAID first selectively screens chloroplast genome-related clean reads using a reference sequence at a lower sequence similarity. Then, these sequences are subjected to a BLAST search against a reference sequence database provided by the researchers. By organizing all BLAST results, the test sample is determined to have the highest similarity or the greatest sum of bit scores with a specific reference sequence.

genome captures seem common (Fig. 3). We used distances between ML trees and the reference tree (which was constructed using the whole chloroplast genome) to assess the ability of partial super DNA barcodes to identify species. Genetic distances decreased as the size of sampled genome increased (Fig. 4a). However, no obvious inflection point was observable. When 5%–15% of genome data was used, about one-third of species were identified; many phylogenetic branches had no branch lengths and unstable phylogenetic positions. When 20%–55% of genome data were used, the positions of some of the larger phylogenetic tree branches were still less certain. When 60%–70% of genome data were used, the positions of only two small branches remained unstable. When 75%–85% of genome data used, only small phylogenetic tree branch positions continued to change. Finally, when the coverage reached 90%, the topologies of the phylogenetic trees became stable. These results indicate that construction of robust phylogenetic relationships requires data from nearly the whole chloroplast genome, however, species discrimination requires only 20% of random whole chloroplast genome. The number of reads belonging to chloroplast genomes became steadily stable with the increase in sample sizes (Fig. 4b). These estimates were made using 27 samples with NGS from 100,000 to 1,000,000 reads (Table S3).

3.4. Minimum next generation sequencing depth for species identification

We next determined the minimum reads needed for species identification with 20% of whole chloroplast genomes. Reads were assembled into contigs. The mean values of the number of contigs decreased sharply when the sample size reached 400,000 reads

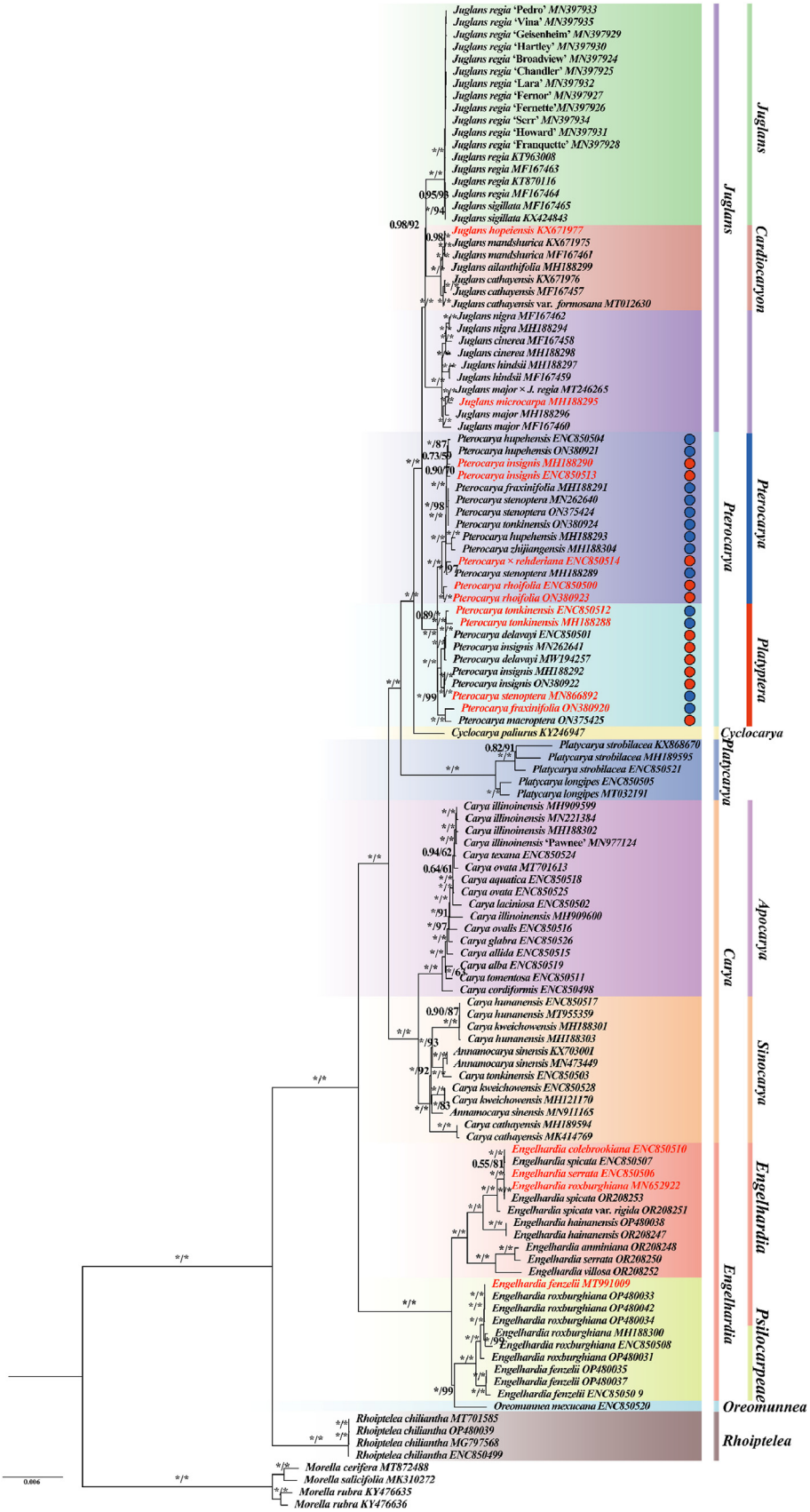


Fig. 3. Maximum likelihood strict consensus tree of Juglandaceae based on the whole chloroplast genome sequences. ML bootstrap values and BI posterior probability values are shown beside the branches. Star represents the supported value as 1.00 or posterior probability as 100. Species names in red were misidentified.

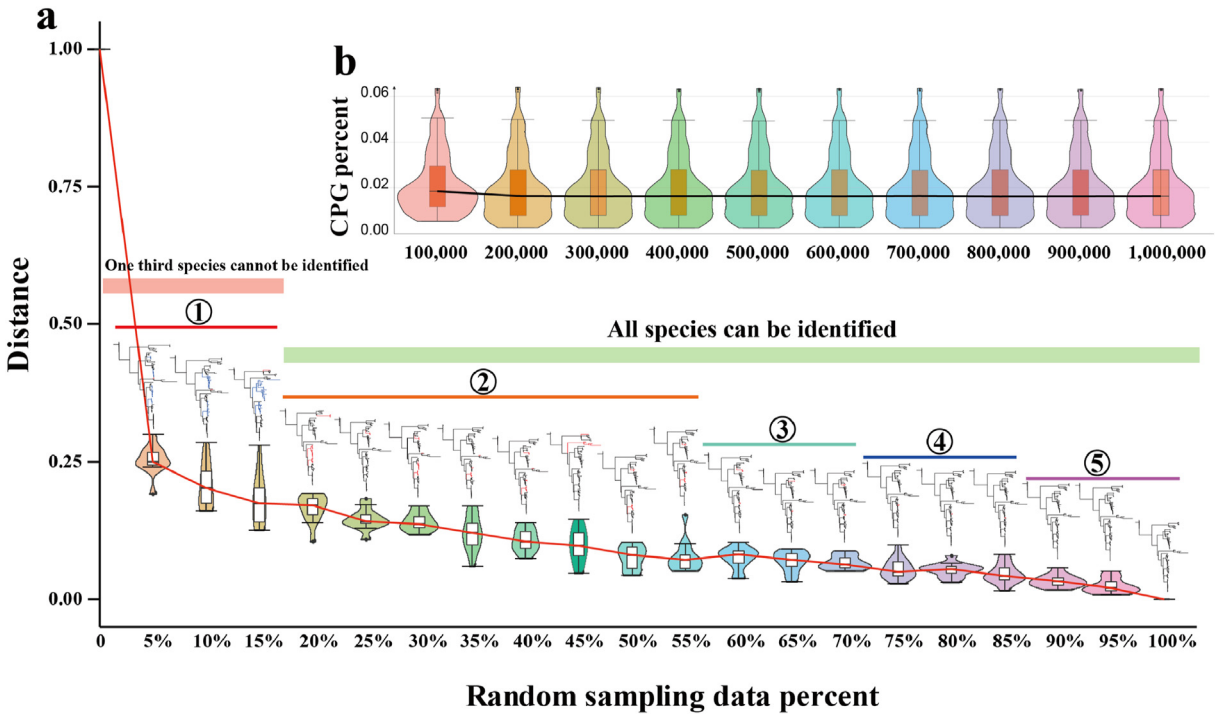


Fig. 4. ML tree distances indicate resolution of species identity changes as percentage of chloroplast genome coverage increases in the walnut family. **a**: Genetic distance calculated between ML tree constructed based on the random chloroplast genome samples and ML tree constructed from whole chloroplast genome. ML trees constructed for each dataset are shown above violin plots. Manually identified regions with unstable phylogenetic positions are shown in red. Branches with species that cannot be accurately identified are shown in blue. **b**: Proportion of chloroplast genome reads with the increase of sample sizes. The numbers in the circles in the figure represent: ①: Several branches were gathered together, the phylogenic relationship among them was not clear; ②: Several big branches' phylogenic positions were not clear; ③: Two small branches' phylogenic positions were not clear; ④: One small branch's phylogenic position was not clear; ⑤: Phylogenic structure became stable.

and stabilized between 1 and 3. When the number of random samples was small, the proportion of chloroplast genome raw data to the total data volume fluctuated, which was more prominent when the random sampling was 100,000 reads (Fig. 5a). At increased sampling depth, the number of contigs generated gradually decreased as the amount of data increased, even though the number of contigs fluctuated when the data volume was small. Additionally, in the transition from 400,000 to 500,000 reads, the decrease in the number of generated contigs slowed and nearly reached a steady state at 700,000 reads. When sampling reads reached 1,000,000, 1 to 3 longer contigs were generated, from which the whole chloroplast genome could be assembled. This was strong support for the maximum sampling volume for our study (Fig. 5a). The reliability of the assembled draft chloroplast genomes was parameterized using ω (ω = right base number/genome total length*100%). We also observed that as the number of reads increased, the credibility of the data increased, and when the data reached 500,000, it was already very close to the reference sequence (Fig. 5b).

The sequence with the largest score value compared to the reference sequence was considered the identification result. Species identification was considered successful and recorded as value one only when the species names of the two were exactly the same. When the species names were different, they were recorded as 0. As the random data increased, the number of identified samples (Fig. 5c in pink) increased and misidentified samples (Fig. 5c in gray) decreased. Moreover, the number of misidentified random samples fluctuated only in a small range of less than five random samples. The identification results tended to be constant, and the correct rate was nearly 100% (1,000,000, 98.02%; 900,000, 99.21%; 800,000, 99.60%; 700,000, 99.60%; 600,000, 99.21%; 500,000, 99.21%).

Distances between test trees and reference tree decreased. Specifically, when the sampling data reached 500,000, the distances flattened out. When the data volume reached 900,000, the average distance among test samples and reference tree reached a minimum value (Fig. 6a). The UPGMA (Unweighted Pair Group Method with Arithmetic Mean) tree results show that when data reached 1,000,000, samples were identified at the highest level (27/27, 100%); after 600,000 reads, identification success rate reached 96.30% (Fig. 6b). The UPGMA tree results also show as clean data increased, sub-data sets of the same sample gradually clustered together (Fig. 6c). When the data reached 600,000, the majority of samples were clustered together (26/27, 96.30%).

3.5. Species identification using assembly-free original reads based on bit-score value

AFRAID accurately identified species for all 27 query samples after 500,000 clean reads (Figs. 7 and S13). One clean read was not identified at species level, likely owing to limited available information sites.

Skmer generated consistent phylogenetic relationships based on genetic distances and Jaccard distances. Genetic distances obtained from MIKE were consistent with these findings. However, Jaccard distances generated from MIKE differed from those generated by Skmer. Results supported by the majority of the interspecific analyses were compared with the genetic relationships constructed based on complete chloroplast genomes using Jaccard distance.

AFRAID supports the placement of *Oremunnea* in *Engelhardia*, *Platycarya* situated between *Engelhardia* and *Carya*, and *Cyclocarya* as a sister group to *Pterocarya*. These genetic relationships differ from previously suggested relationships obtained by whole chloroplast genome data. However, the genetic variations displayed

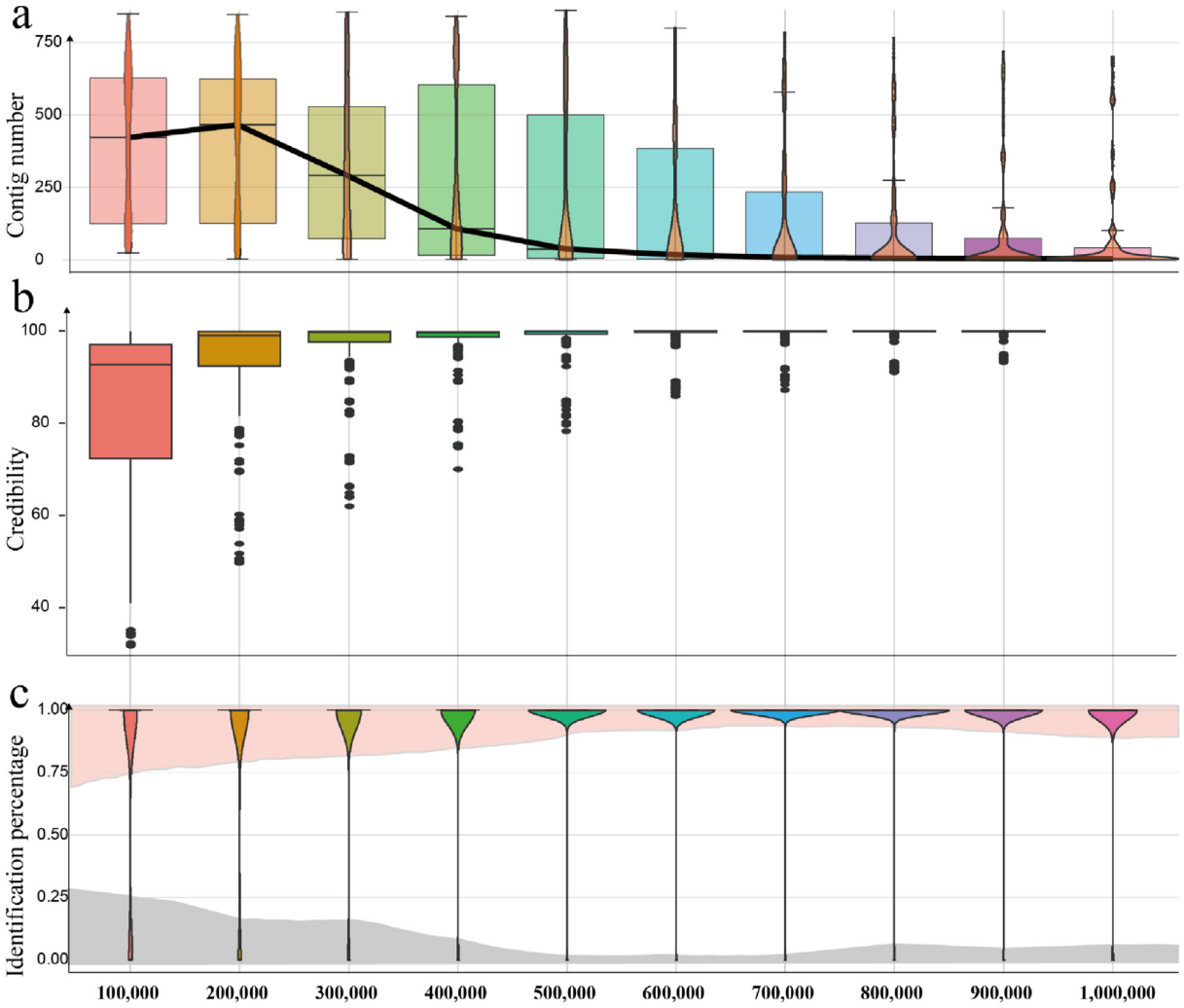


Fig. 5. Changes in contig numbers and similarities of draft genomes to reference genomes with the increase of sampled reads of 27 NGS datasets of Juglandaceae. **a**: Number of contigs assembled by chloroplast genome related reads. **b**: Similarities between draft genome sequences and reference genome sequences measured by ω (= right base number/genome total length*100%). **c**: Effects of sampling different raw data gradients on the identification results of species in the walnut family. Pink background indicates successfully identified random samples; gray background indicates unsuccessfully identified random samples.

between species are sufficient to meet the criteria for species identification (Fig. S14).

AFRAID takes approximately 20 min to run on a system with 10 cores and 32 GB of RAM, whereas MIKE takes around 2 h, and Skmer takes about 3 h.

4. Discussion

4.1. Whole chloroplast genomes may be unnecessary for plant identification

Here, we found that only 20% of whole chloroplast genome data was necessary for complete species identification in Juglandaceae using branch lengths (Fig. 4). Several studies on taxon-specific DNA barcodes (e.g., partial chloroplast genome data) strongly support this conclusion (Li et al., 2015; Wu et al., 2021; Zhang et al., 2021). Our finding suggests that sequencing whole chloroplast genomes may be unnecessary for identification of most plant species.

Chloroplast DNA, however, may not provide enough resolution to distinguish closely related species, especially if they share a recent common ancestor (such as some species in *Juglans*, Fig. 3), which reduces its effectiveness in species identification (Palmer et al., 1988; Shaw et al., 2007). In our study, chloroplast genome data failed to distinguish several species. In some cases, plant materials were misidentified, e.g., *Carya kweichowensis* MH188301 is a misidentification of *C. hunanensis*; *Engelhardia roxburghiana* MN652922, *Engelhardia serrata* ENC850506, *Engelhardia colebrookiana* ENC850510, and *Engelhardia fenzelii* MT991009 were also likely misidentified. In other cases (e.g., species in *Pterocarya*), identification may have been complicated by chloroplast genome capture or incomplete lineage sorting during speciation (Rieseberg and Soltis, 1991). Our analysis may also have clarified previously unknown relationships. For example, *J. regia* may be a cultivated form of *Juglans sigillata*. We may also have found something entirely new, i.e., *Annamocarya sinensis* MN911165 is quite distinct.

Chloroplast DNA is typically maternally inherited, which means it does not account for genetic contributions from both parents, potentially overlooking important genetic variations (Wolfe et al., 1987). This may explain differences in the results between MIKE and Skmer. In addition, selective pressures may lead to convergent evolution in chloroplast genomes, causing unrelated species to appear genetically similar (Clegg et al., 1994). In cases where

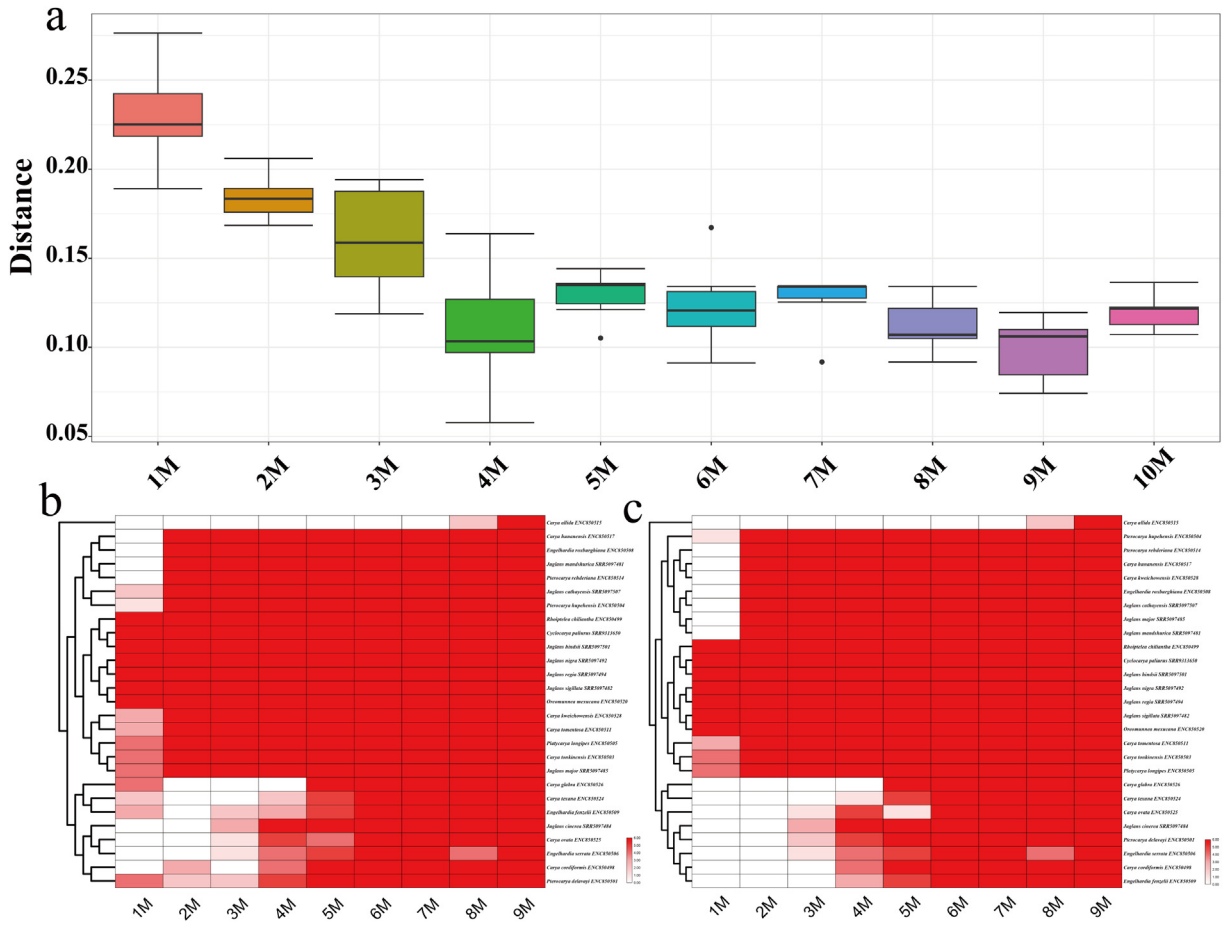


Fig. 6. Results of tree comparison and UPGMA clustering. **a**: Tree comparison results among test samples and reference tree; **b**: Sample UPGMA clustering results in different datasets; **c**: Species UPGMA identification results in different datasets. M represents 100,000 reads.

chloroplast genomes are not sufficient to completely differentiate species, additional genetic markers (e.g., nuclear data) or whole-genome sequencing should be used (Bock et al., 2014).

4.2. Total DNA raw data for efficient data acquisition and species identification

DNA barcoding has been extensively used to identify plant species (Selvaraj et al., 2015; Dong et al., 2021; Zhang et al., 2021; Govender et al., 2022). However, current methods have several limitations. Universal DNA barcoding has a low capacity to identify species (see Figs. S2-S6). Similarly, taxon-specific DNA barcodes are not effective at identifying species in some genera (Figs. S7-S12) and fail to meet the needs of current research. Taxon-specific DNA barcodes also require additional design of relevant primer sequences, which complicates the construction of subsequent databases. Moreover, the length of current DNA barcode design and existing next generation sequencing platforms are mismatched, making it challenging to obtain these DNA barcode sequences using relatively cheaper NGS techniques. Furthermore, DNA barcoding methods that rely on conventional PCR require high-quality DNA sequences for the relevant barcodes; however, many samples have severely degraded DNA or low DNA content (Li et al., 2011b).

Here, we addressed these limitations by sampling raw data in a random gradient. When the data volume exceeded 500,000 reads, this approach identified nearly 100% of species in Juglandaceae (Fig. 5c), with less than five incorrectly identified random samples. The current minimum sequencing volume of a single sample for NGS is 500,000 reads. Thus, our results indicate that our proposed approach to identify plant species is promising (Figs. 2 and 7). This approach overcomes the uncertainty of previous DNA barcode PCR amplification methods, is highly inclusive of DNA profiles of identified samples, and coupled with the current use of NGS, the assembly-free identification method should take less than 15 days (Li et al., 2021). Furthermore, our study shows that when the total DNA of degraded samples is obtained, an ideal identification method would be to use direct library building to obtain NGS data, then use the chloroplast-related sequences in the total DNA for precise identification in both assembled or assembly-free methods.

4.3. Traditional DNA barcoding needs to be expanded

Our results show that DNA barcoding regions are no longer fixed in the genome. From universal DNA barcoding and taxon-specific DNA barcoding to whole chloroplast genomes for species identification, these data for species identification are identified for a fixed region, regions, or the whole chloroplast genome (Coissac et al., 2016). This is identical to the traditional DNA barcoding concept, differing only in the amount of data used for sample identification. Traditional DNA barcoding can still play a significant role in many areas (e.g., providing DNA barcode reference database support, normal plant material identification), including in some ambitious projects that have been launched over the past few years [e.g.,

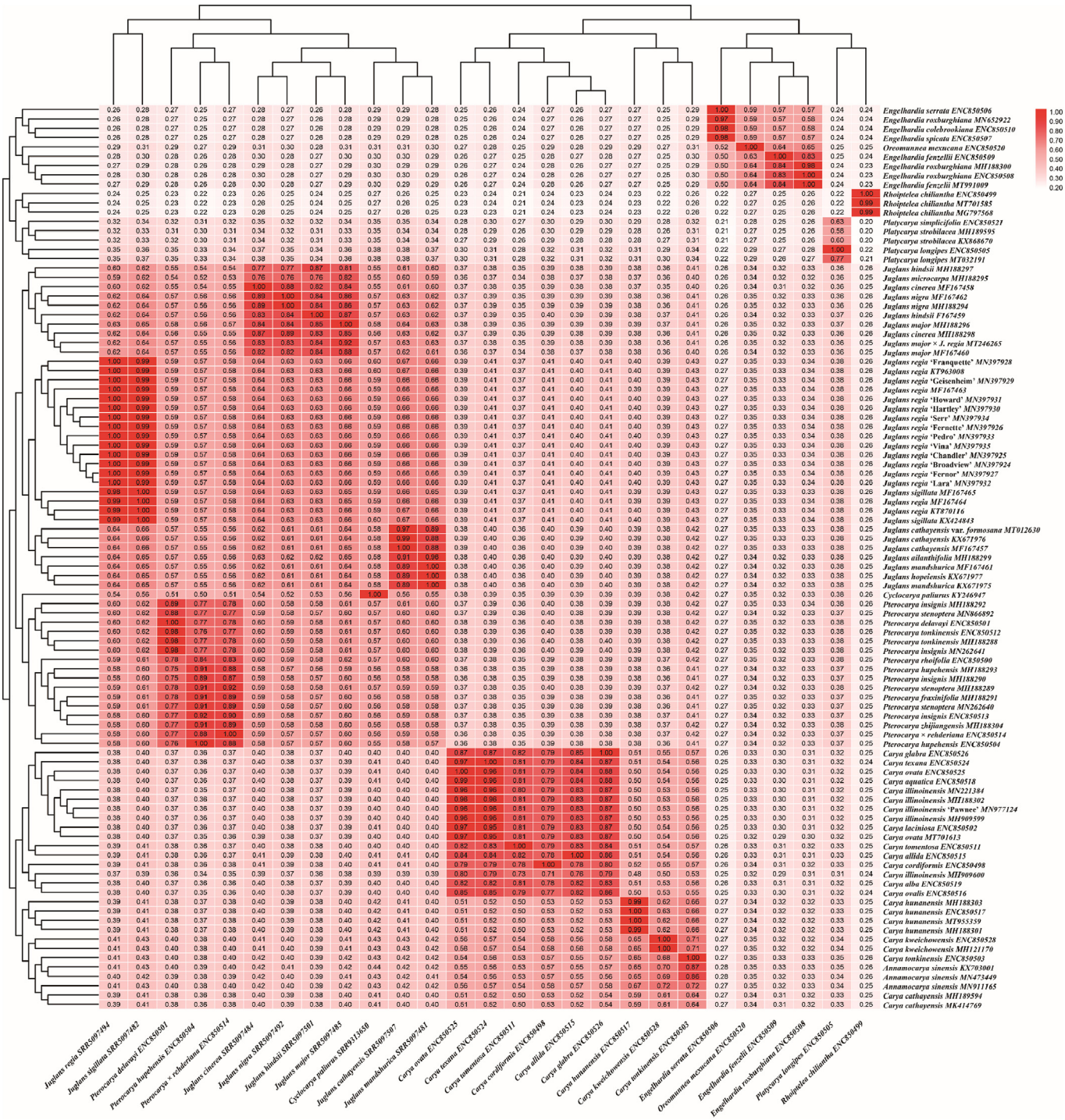


Fig. 7. Identification of species with assembly-free original reads based on bit-score value. Shades of red represent the degree of similarity between the samples, with the similarity index ranging from 0 to 1.00.

BARCODE 500K (https://ibol.org), BIOSCAN (Hobern and Hebert, 2019), and ISHAM- ITS (Irinyi et al., 2016)].

Our study aimed to determine whether the traditional DNA barcoding concept of using fixed regions for species identification is necessary. Our results show that both random chloroplast genome sampling, random raw data sampling, and assembly-free identification show a high success rate of species identification. More specifically, 20% of random chloroplast genomes can fully identify species (Fig. 4a), and both 500,000 clean sequences assemble or assembly-free methods can achieve accurate species identification (Figs. 5c and 7). A distinctive feature of both data types is a high degree of species identification success. Furthermore, in both, the chloroplast genomic regions used are random and uncertain, suggesting that these data should be classified as a type of DNA barcoding, a vital complement to the traditional DNA barcoding concept.

The results of direct species identification using original reads by AFRAID, MIKE and Skmer indicate that for unknown samples,

statistical analysis using the identification results of reads can identify the samples to the species level, which enriches the statistical methods for identifying species using DNA barcodes (Figs. 6, 7 and S14). These extensions will reduce the requirement of sample DNA for species identification and increase the success rate of data acquisition and species identification. However, due to the limitations of traditional DNA barcoding, expanding DNA barcoding concepts to study species diversity in mixed samples still requires researchers to further develop data analysis methods.

4.4. Assembly-free identification will accelerate species identification

The major discovery of this study is that our proposed method (AFRAID) provides fast, accurate species identification. This method overcomes the limitations of using single or multiple DNA barcodes to identify species. At present, this method is based on the chloroplast genome. We also found that MIKE and Skmer algorithms based on all genetic data have tremendous potential in species identification (Fig. S14) (Sarmashghi et al., 2019; Bohmann et al., 2020; Paula et al., 2022; Wang et al., 2024). We believe that the use of complete nuclear genome data will increase the accuracy of AFRAID in species identification. However, analyzing larger genomic datasets will increase computation burdens and require a larger nuclear genome database, which may be a worthy goal for future research. AFRAID is also user-friendly, accessible to beginners and the general public, allowing data analysis to be completed with a single command line. Thus, the proposal and validation of the assembly-free original reads identification method provide a new approach for species identification for community scientists, expanding opportunities for species identification.

5. Conclusions

This study used newly sequenced whole chloroplast genomes of 28 Juglandaceae species to construct a solid phylogenetic relationship for identification. By clarifying the relationship between species identification and species phylogeny, we determined that partial chloroplast genome sequences (i.e., 20% of genome) can be used to identify species. The proposed assembly-free reads accurate identification (AFRAID) method provides a new approach for species identification that overcomes the limitations of traditional DNA barcodes. AFRAID requires less runtime and less raw data than either MIKE or Skmer, with the same species identification rates. This study improves plant species identification, injecting new vitality into traditional DNA barcoding.

CRediT authorship contribution statement

Yanlei Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kai Chen:** Writing – original draft, Visualization, Validation, Formal analysis. **Lihu Wang:** Writing – original draft, Methodology, Investigation, Formal analysis. **Xinqiang Yu:** Software. **Chao Xu:** Visualization, Resources, Methodology. **Zhili Suo:** Resources, Methodology, Investigation. **Shiliang Zhou:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources. **Shuo Shi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources. **Wenpan Dong:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources.

Data accessibility

All 28 CPG data have been deposited in GenBank (the accession numbers of each genome were listed in Table S1) and China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gsub/submit/biosample/subSAM125150/finishedOverview). Researchers can download AFRAID tool for further plant species identification here (https://github.com/Mycroft-behind/classify/tree/main).

Declaration of competing interest

There are no conflicts of interest to declare.

Acknowledgements

This study was partly supported by the funds from Natural Science Foundation of Hebei Province (C2022402017). We are very grateful to Drs. Jianwen Zhang and Guojin Zhang for their suggestions on the accuracy of this article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pld.2024.10.002.

References

Altschul, S.F., Gish, W., Miller, W., et al., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Bankevich, A., Nurk, S., Antipov, D., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.
Barberán, A., Ladau, J., Leff, J.W., et al., 2015. Continental-scale distributions of dust-associated bacteria and fungi. Proc. Natl. Acad. Sci. U.S.A. 112, 5756–5761.
Bock, D.G., Kane, N.C., Ebert, D.P., et al., 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: neither from Jerusalem nor an artichoke. New Phytol. 201, 1021–1030.
Bohmann, K., Mirarab, S., Bafna, V., et al., 2020. Beyond DNA barcoding: the unrealized potential of genome skim data in sample identification. Mol. Ecol. 29, 2521–2534.
Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Phylogenet. Evol. 17, 540–552.
Chen, X., Zhou, J., Cui, Y., et al., 2018. Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. Front. Pharmacol. 9, 695.
Clegg, M.T., Gaut, B.S., Learn, G.H., et al., 1994. Rates and patterns of chloroplast DNA evolution. Proc. Natl. Acad. Sci. U.S.A. 91, 6795–6801.
Coissac, E., Hollingsworth, P.M., Lavergne, S., et al., 2016. From barcodes to genomes: extending the concept of DNA barcoding. Mol. Ecol. 25, 1423–1428.
Cracraft, J., 1983. Species concepts and speciation analysis. In: Johnston, R.F. (Ed.), Current Ornithology, vol. 1. Springer, New York, pp. 159–187.
Dong, W., Cheng, T., Li, C., et al., 2014. Discriminating plants using the DNA barcode *rbcL*b: an appraisal based on a large data set. Mol. Ecol. Resour. 14, 336–343.
Dong, W., Li, E., Liu, Y., et al., 2022. Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. BMC Biology 20, 92.
Dong, W., Liu, Y., Xu, C., et al., 2021. Chloroplast phylogenomic insights into the evolution of *Distylium* (Hamamelidaceae). BMC Genomics 22, 293.
Duan, H.N., Jiang, Y.Z., Yang, J.B., et al., 2024. Skmer approach improves species discrimination in taxonomically problematic genus *Schima* (Theaceae). Plant Divers. 45, https://doi.org/10.1016/j.pld.2024.06.003.
Galan, M., Pagès, M., Cosson, J.F., 2012. Next-generation sequencing for rodent barcoding: species identification from fresh, degraded and environmental samples. PLoS One 7, e48374.
Govender, A., Singh, S., Groeneveld, J., et al., 2022. Experimental validation of taxon-specific mini-barcode primers for metabarcoding of zooplankton. Ecol. Appl. 32, e02469.
Guo, W., Chen, J., Li, J., et al., 2020. Portal of Juglandaceae: a comprehensive platform for Juglandaceae study. Hortic. Res. 7, 35.
Hajibabaei, M., Singer, G.A.C., Clare, E.L., et al., 2007. Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. BMC Biology 5, 24.
Hebert, P.D.N., Cywinska, A., Ball, S.L., et al., 2003. Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270, 313–321.
Hobern, D., Hebert, P.D.N., 2019. Bioscan - revealing eukaryote diversity, dynamics, and interactions. Biodivers. Inf. Sci. Stand. 3, e37333.
Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., et al., 2009. A DNA barcode for land plants. Proc. Natl. Acad. Sci. U.S.A. 106, 12794–12797.

Hollingsworth, P.M., Graham, S.W., Little, D.P., 2011. Choosing and using a plant DNA barcode. PLoS One 6, e19254.
Hong, D., 2016. Biodiversity pursuits need a scientific and operative species concept. Biodivers. Sci. 24, 979–999.
Huang, D.I., Cronk, Q.C.B., 2015. Plann: a command-line application for annotating plastome sequences. Appl. Plant Sci. 3, 1500026.
Irinyi, L., Lackner, M., de Hoog, G.S., et al., 2016. DNA barcoding of fungi causing infections in humans and animals. Fungal Biol. 120, 125–136.
Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., et al., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Phylogenet. Evol. 30, 772–780.
Kearse, M., Moir, R., Wilson, A., et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649.
Kress, W.J., Erickson, D.L., 2007. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. PLoS One 2, e508.
Kress, W.J., Wurdack, K.J., Zimmer, E.A., et al., 2005. Use of DNA barcodes to identify flowering plants. Proc. Natl. Acad. Sci. U.S.A. 102, 8369–8374.
Kumar, S., Stecher, G., Li, M., et al., 2018. Mega X: molecular evolutionary genetics analysis across computing platforms. Mol. Phylogenet. Evol. 35, 1547–1549.
Le, D.T., Zhang, Y.Q., Xu, Y., et al., 2020. The utility of DNA barcodes to confirm the identification of palm collections in botanical gardens. PLoS One 15, e0235569.
Li, M., Cao, H., But, P.P.H., et al., 2011b. Identification of herbal medicinal materials using DNA barcodes. J. Syst. Evol. 49, 271–283.
Li, D.Z., Gao, L.M., Li, H.T., et al., 2011a. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc. Natl. Acad. Sci. U.S.A. 108, 19641–19646.
Li, J., Wang, S., Yu, J., et al., 2013. A modified CTAB protocol for plant DNA extraction. Chin. Bull. Bot. 48, 72–78.
Li, N., Cai, Q., Miao, Q., et al., 2021. High-throughput metagenomics for identification of pathogens in the clinical settings. Small Methods 5, 2000792.
Li, X., Yang, Y., Henry, R.J., et al., 2015. Plant DNA barcoding: from gene to genome. Biol. Rev. 90, 157–166.
Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25, 1451–1452.
Liu, J., 2016. The integrative species concept and species on the speciation way. Biodivers. Sci. 24, 1004–1008.
Liu, J., Milne, R.I., Möller, M., et al., 2018. Integrating a comprehensive DNA barcode reference library with a global map of yews (*Taxus* L.) for forensic identification. Mol. Ecol. Resour 18, 1115–1131.
Liu, J., Möller, M., Gao, L.M., et al., 2011. DNA barcoding for the discrimination of Eurasian yews (*Taxus* L., Taxaceae) and the discovery of cryptic species. Mol. Ecol. Resour. 11, 89–100.
Liu, Y., Xu, C., Dong, W., et al., 2023. What determines plant species diversity along the Modern Silk Road in the east? iMeta 2, e74.
Liu, Y., Xu, C., Sun, Y., et al., 2021. Method for quick DNA barcode reference library construction. Ecol. Evol. 11, 11627–11638.
Lu, A., 1982. On the geographical distribution of the Juglandaceae. Acta Phytotaxon. Sinica 20, 257–274.
Lv, S.Y., Ye, X.Y., Li, Z.H., et al., 2023. Testing complete plastomes and nuclear ribosomal DNA sequences for species identification in a taxonomically difficult bamboo genus *Fargesia*. Plant Diver. 45, 147–155.
Manchester, S.R., 1989. Early history of the Juglandaceae. In: Ehrendorfer, F. (Ed.), Woody Plants - Evolution and Distribution since the Tertiary. Springer, Vienna, pp. 231–250.
McFadden, C.S., Brown, A.S., Brayton, C., et al., 2014. Application of DNA barcoding in biodiversity studies of shallow-water octocorals: molecular proxies agree with morphological estimates of species richness in Palau. Coral Reefs 33, 275–286.
Nguyen, L.T., Schmidt, H.A., von Haeseler, A., et al., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Phylogenet. Evol. 32, 268–274.
Palmer, J.D., Jansen, R.K., Michaels, H.J., et al., 1988. Chloroplast DNA variation and plant phylogeny. Ann. Mo. Bot. Gard. 75, 1180–1206.
Patel, R.K., Jain, M., 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7, e30619.
Paula, D.P., Barros, S.K.A., Pitta, R.M., et al., 2022. Metabarcoding versus mapping unassembled shotgun reads for identification of prey consumed by arthropod epigeal predators. GigaScience 11, giac020.
Prosser, S.W.J., deWaard, J.R., Miller, S.E., et al., 2016. DNA barcodes from century-old type specimens using next-generation sequencing. Mol. Ecol. Resour. 16, 487–497.
Rieseberg, L.H., Soltis, D., 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trends Plants 5, 65–84.
Ronquist, F., Teslenko, M., van der Mark, P., et al., 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61, 539–542.
Sarmashghi, S., Bohmann, K., Gilbert, P.M.T, et al., 2019. Skmer: assembly-free and alignment-free sample identification using genome skims. Genome Biol. 20, 34.
Selvaraj, D., Sarma, R.K., Shanmughanandhan, D., et al., 2015. Evaluation of DNA barcode candidates for the discrimination of the large plant family Apocynaceae. Plant Syst. Evol. 301, 1263–1273.
Shaw, J., Lickey, E.B., Schilling, E.E., et al., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am. J. Bot. 94, 275–288.
Shirai, M., Takano, A., Kurosawa, T., et al., 2022. Development of a system for the automated identification of herbarium specimens with high accuracy. Sci. Rep. 12, 8066.
Shokralla, S., Gibson, J.F., Nikbakht, H., et al., 2014. Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Mol. Ecol. Resour. 14, 892–901.
Simpson, G.G., 1951. Horses. Oxford University Press, New York and London, pp. 1–3.
Song, Y.G., Fragnière, Y., Meng, H.H., et al., 2020. Global biogeographic synthesis and priority conservation regions of the relict tree family Juglandaceae. J. Biogeogr. 47, 643–657.
Tyagi, K., Kumar, V., Kundu, S., et al., 2019. Identification of Indian spiders through DNA barcoding: cryptic species and species complex. Sci. Rep. 9, 14033.
Wang, F., Wang, Y., Zeng, X., et al., 2024. MIKE: an ultrafast, assembly-, and alignment-free approach for phylogenetic tree construction. Bioinformatics 40, btae154.
Wolfe, K.H., Li, W.H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. U.S.A. 84, 9054–9058.
Wu, L., Wu, M., Cui, N., et al., 2021. Plant super-barcode: a case study on genome-based identification for closely related species of *Fritillaria*. Chin. Med. 16, 52.
Xu, C., Dong, W., Shi, S., et al., 2015. Accelerating plant DNA barcode reference library construction using herbarium specimens: improved experimental techniques. Mol. Ecol. Resour 15, 1366–1374.
Yan, H., Zhou, P., Wang, W., et al., 2024. Biogeographic history of *Pterocarya* (Juglandaceae) inferred from phylogenomic and fossil data. J. Syst. Evol. jse, 13055.
Yang, Y., Forsythe, E.S., Ding, Y.-M., et al., 2023. Genomic analysis of plastid-nuclear interactions and differential evolution rates in coevolved genes across Juglandaceae species. Genome Biol. Evol. 15, evad145.
Yu, J., Xue, J.H., Zhou, S.L., 2011. New universal *matK* primers for DNA barcoding angiosperms. J. Syst. Evol. 49, 176–181.
Zhang, D., Gao, F., Jakovlić, I., et al., 2020. PhyloSuite: an integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol. Ecol. Resour 20, 348–355.
Zhang, Q., Ree, R.H., Salamin, N., et al., 2022. Fossil-informed models reveal a boreotropical origin and divergent evolutionary trajectories in the walnut family (Juglandaceae). Syst. Biol. 71, 242–258.
Zhang, W., Sun, Y., Liu, J., et al., 2021. DNA barcoding of *Oryza*: conventional, specific, and super barcodes. Plant Mol. Biol. 105, 215–228.
Zhang, Z., Zhang, Y., Song, M., et al., 2019. Species identification of *Dracaena* using the complete chloroplast genome as a super-barcode. Front. Pharmacol. 10, 1441.
Zhou, H., Hu, Y., Ebrahimi, A., et al., 2021. Whole genome based insights into the phylogeny and evolution of the Juglandaceae. BMC Ecol. Evol. 21, 191.