Comparative Analyses of 35 Complete Chloroplast Genomes from the Genus *Dalbergia* (Fabacaeae) and the Identification of DNA Barcodes for Tracking Illegal Logging and Counterfeit Rosewood

Zhou Hong
Chinese Academy of Forestry
Dan Peng
AGIS
Wenchuang He
AGIS
Ningnan Zhang
Chinese Academy of Forestry
Zengjiang Yang
Chinese Academy of Forestry
Luke R Tembrock
Colorado State University
Zhiqiang Wu (wuzhiqiang@caas.cn)
AGIS https://orcid.org/0000-0002-4238-7317
Xuezhu Liao
AGIS
Daping Xu
Chinese Academy of Forestry

Research Article

Keywords: forensic biology, phylogeny, CITES, endangered species, forest conservation

Posted Date: May 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-271391/v1

Version of Record: A version of this preprint was published at Forests on April 16th, 2022. See the published version at https://doi.org/10.3390/f13040626.

Abstract

The genus Dalbergia contains more than 120 species several of which are trees that produce traditional medicines and extremely high value timber commonly referred to as rosewood. Due to the rarity of these species in the wild, the high value of the timber, and a growing international illicit trade CITES has listed the entire genus in appendix II and the species D. nigra in appendix I because it is considered threatened with extinction. Given this and the fact that species or even genus level determination is nearly impossible from cut timber alternative molecular methods are needed to identify and track intercepted rosewood. In order to improve molecular identification of rosewood, we sequenced and assembled eight chloroplast genomes including D. nigra as well as conducted comparative analyses with all other available chloroplast genomes in Dalbergia and closely related lineages. From these analyses numerous repeats including simple sequence repeats (SSR) and conserved nucleotide polymorphisms unique to subclades within the genus were detected. From phylogenetic analysis using the CDS of 77 coding genes the groups Siam rosewood and scented rosewood based mainly on wood characteristics were supported as monophyletic. In addition, several instances of paraphyly and polyphyly resulting from mismatch between taxonomic determinations and phylogenetic tree topology were identified. Ultimately, the highly variable regions in the chloroplast genomes will provide useful plastid markers for further studies regarding the identification, phylogeny, and population genetics of Dalbergia species including those frequently intercepted in illegal trade.

1. Introduction

The genus *Dalbergia* L.f. in the tribe Dalbergieae (DC.) Cardoso and family Fabaceae contains approximately 275 species of trees, shrubs, and lianas. The genus is widely distributed worldwide with species occurring in the tropics and subtropics of South and Central America, Africa, Madagascar, and Asia. Several tree species in the genus are highly valued for producing premium darkly colored, dense, and sometimes fragrant wood used in a variety of applications such as the production of musical instruments, traditional medicine, fine furniture, cabinetry, and veneers. By way of example rosewood furniture in China can range from thousands of dollars for an ornate chair to millions of dollars ($USD) for a bed frame (Zhu, 2020). The main species used to produce high quality timber are loosely grouped into the three categories (made up of several species in each group) black rosewood, scented rosewood, and Siam rosewood based mainly on characteristics of the wood. The black and Siam rosewoods are the most highly valued.

The extremely high value of rosewood has led to large scale illegal logging and international trade in a number of different rosewood species involving numerous countries where rosewood is native. The main centers of rosewood logging are found in Central and South America (Espinoza et al., 2015; Vardeman and Velásquez Runk 2020), Africa and Madagascar (Innes 2010; Abdul-Rahaman et al., 2016), and Southeast Asia (Siriwat and Nijman 2018; Nhung et al., 2020). Madagascar is an area of particular concern given the high number of endemic *Dalbergia* species (42 spp.) and the presence of organized syndicates taking advantage of the political and economic instability in the country (DuPuy 2002; Patel

2007; Randriamalala and Liu 2010) Rosewood is considered the most trafficked group of endangered species in the world with the value of global seizures exceeding ivory, rhino horn, and big cats combined (United Nations Office on Drugs and Crime, 2016).

Because of illegal trade, over exploitation, and the similarity of wood between some species the entire *Dalbergia* genus is protected under CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora) appendix II which requires permitting for export of most parts of the plant including wood from this genus (CITES, 2020). The Brazilian species *Dalbergia nigra* (Vell.) Benth endemic to the Bahia interior forest ecoregion, is considered threatened with extinction and has been listed in appendix I of CITES (2020) prohibiting all international trade. In addition to listing in appendix II of CITES *D. fusca* Pierre and *D. odorifera* T.C. Chen are also listed in China's key protection list. Given the illegal trade in rosewood and difficulty in differentiating Dalbergieae species based on differences in confiscated wood, a means of effectively identifying different rosewood species is needed. In parallel with tracking illegal trade of rosewood, programs that could be developed to produce licit rosewood timber will also benefit from routine molecular identification and tracking as part of a robust certification program.

In studies of plant biology, many different kinds of molecular markers have been developed, including AFLP (Amplified Fragment Length Polymorphism), SSRs (Simple Sequence Repeats), and a variety of short (500-1000bp) DNA sequences referred to as barcodes for the identification of species and populations (Thiel *et al.*, 2003). The cost of sequencing has rapidly decreased in recent years with concurrent increases in throughput making the discovery and development of diagnostic markers more efficient and cost effective even for a large number of species (Wu *et al.*, 2020). Given the utility of DNA barcode-based-identification numerous different barcode regions from different cellular compartments have been proposed in plants (CBOL Plant Working Group, 2009). In recent years, super DNA barcodes have been applied to entire genome sequences, as in the comparative analyses of complete plant chloroplast genomes (Hollingsworth *et al.*, 2016; Zhang *et al.*, 2019; Zhang *et al.*, 2020). Unlike the relatively large and complex nuclear genome, the chloroplast genome has several advantages including uniparental inheritance, high information content (in variable sites), very low recombination rates, and high copy number, making chloroplast genomes and chloroplast barcodes ideally suited for studies in plant systematics (Wang *et al.*, 2011; Fu *et al.*, 2019), population genetics (Wang *et al.*, 2011; Zhou *et al.* 2021) and plant taxonomy (CBOL Plant Working Group, 2009). The conserved gene content and structural arrangement of chloroplast genomes in two inverted repeating regions (IRs), a long single-copy region (LSC) and, a short single-copy region (SSC) make assembly and alignment of chloroplast genomes more complete and less error prone than with plant nuclear or mitochondrial genomes. Additionally, because a large number of complete chloroplast genomes are available in public databases comparative analyses and searches are more accurate and thorough than with any other complete genome data in plants. Based on the advantages outlined above the whole chloroplast genome and barcodes derived therefrom would be ideal for identifying *Dalbergia* species from wood samples.

To isolate informative molecular markers useful for species identification from wood material we conducted the following analyses: 1) sequenced, assembled, and annotated the chloroplast genomes of eight *Dalbergia* species (including the CITES listed appendix I *D. nigra*) using next-generation sequencing methods; 2) conducted a phylogenetic analysis to infer the relationships among *Dalbergia* species and; 3) comprehensively analyzed the highly variable regions and conserved nucleotide sites unique to each clade in *Dalbergia* that could be employed for DNA barcode-based-identification.

2. Results And Discussion

2.1 Complete Chloroplast Genomes

The complete chloroplast genome lengths of the eight *Dalbergia* species sequenced as part of this study range from 155,330 bp to 156,697 bp and were all typically circular in structure (Figure 1, Supplemental Figure 1, Table 1). The length of each chloroplast genome was *D. nigra* (MT644130) 155,330 bp, *D. hupeana* Hance (MT644129) 155,829 bp, *D. fusca* (MT644128) 156,033 bp, *D. odorifera* (MT644131) 156,064 bp, *D. tonkinensis* Prain (MT644133) 156,087 bp, *D. bariensis* Pierre (MT644134) 156,544 bp, *D. cochinchinensis* Pierre (MT644135) 156,576 bp, and *D. oliveri* Prain (MT644132) 156,697 bp. The IR regions are also similar in length, ranging from 25,469 to 25,720 bp, separated by an LSC region (85,110-86,036 bp) and an SSC region (18,856-19,427 bp; Table 1; Supplemental Table 1). In general, the genome sequences of species in *Dalbergia* were similar in length.

In addition to being similar in genome length and overall structure, the number and types of genes are also very similar among the newly sequenced genomes. The annotated coding genes included 45 photosynthesis genes (*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *cemA*, *ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*, *petA*, *petB*, *petD*, *petG*, *petL*, *petN*, *psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*, and *rbcL*), 20 ribosomal protein genes (*rpl14*, *rpl16*, *rpl2*, *rpl20*, *rpl23*, *rpl32*, *rpl33*, *rpl36*, *rps11*, *rps12*, *rps14*, *rps15*, *rps16*, *rps18*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*, and *rps8*), 4 transcription/translation genes (*rpoA*, *rpoB*, *rpoC1*, and *rpoC2*), 4 miscellaneous protein genes (*accD*, *ccsA*, *clpP* and *matK*), and four conserved ORFs, (*ycf1-4*). Introns were found in the 12 coding genes, *atpF*, *clpP*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl16*, *rpl2*, *rpoC1*, *rps12*, *rps16* and *ycf3*. Among the intron containing genes, *ycf3* and *clpP* contained 2 introns, while the 10 other genes contained a single intron.

Table 1 Summary of the complete chloroplast genomes sequenced for this study.

GB ID	Species	TL	Genes	tRNA	rRNA	GC%	LSC	IRB	SS	IRA
MT644128	*Dalbergia fusca*	156,033	83	37	8	36.08	85,475	25,711	19,131	25,716
MT644129	*Dalbergia hupeana*	155,829	83	37	8	36.19	85,304	25,680	19,168	25,677
MT644130	*Dalbergia nigra*	155,330	82	37	8	36.05	85,110	25,469	19,282	25,469
MT644131	*Dalbergia odorifera*	156,064	83	37	8	36.09	85,804	25,702	18,856	25,702
MT644132	*Dalbergia oliveri*	156,697	82	37	8	35.96	86,036	25,691	19,278	25,692
MT644133	*Dalbergia tonkinensis*	156,087	83	37	8	36.09	85,763	25,720	18,884	25,720
MT644134	*Dalbergia bariensis*	156,544	83	37	8	35.94	85,765	25,675	19,427	25,677
MT644135	*Dalbergia cochinchinensis*	156,576	83	37	8	36.08	85,886	25,682	19,326	25,682

Along with the eight *Dalbergia* chloroplasts newly sequenced in this study, an additional 27 published genomes from *Dalbergia*, four from *Arachis* L., two from *Pterocarpus* Jacq. (which also produce a rosewood-type timber), one from *Tipuana tipu* (Benth.) Kuntze, and an outgroup species *Amorpha fruticosa* L. were employed where larger comparative analysis were appropriate (Supplemental Table 1 and 2). In comparing these 43 chloroplast genomes they contained either 82 or 83 coding genes, 37 tRNA genes, and 8 rRNA genes. Among which the genes *rpl2*, *rpl23*, *ycf2*, *ndhB*, *rps7*, and *rps12* have two copies due to duplication in the IR region. The exception to this is *D. oliveri* (MT644132) in which only a single copy of *ycf2* is found with the second copy having been truncated (Table 1; Supplemental Table 1 and 3).

The *ndhF* gene spans IRB and SSC in most of *Dalbergia* species (Supplemental Table 1). The SSC and IRA junction is spanned by *ycf1* in all 43 chloroplast genomes, while no gene spans the IRA and LSC junction. The intergenic region that spans the IRA and LSC junction varies widely across the 43 chloroplast genomes with the distance between *trnN-GUU* and *ndhF* ranging from 864 bp to 1,668 bp. Even within a species this region was found to vary as in *D. tonkinensis* with the distance ranging from 882 bp to 900 bp. Similarly, the intergenic region between *ycf1* and *trnN-GUU* varied from 411 bp to 429 bp across the *D. tonkinensis* chloroplast genomes (Figure 2, Supplemental Figure 2, Supplemental Table 1). The expansion and contraction of the SSC region across the samples used in this study might make this region a suitable candidate for marker development.

2.2 Repeat Analysis

Nucleotide repeats in chloroplast genomes can be very useful markers for identifying populations and/or species given the high rates of mutation in these regions. All 43 chloroplast genomes were analyzed using Reputer software using the limitation that repeats must have the length of 8 bp or more. Four types of repeats were considered, including forward (direct) (F), reverse (R), complement (C), and palindromic (P). Based on different motif types the number of each type was counted based on grouping by sequence length (Figure 3). Almost all the repeats were in the 20-29 bp length range, followed by 30-39 bp, then 50+ bp, with the fewest in the 40-49 bp range. No C and R repeats where detected above 40bp in length and they were rare even in the smaller size ranges. In the 30-39 bp group, C and R repeats were only found in a few species. The R repeats only appear in six *Dalbergia* species, and C is only in *Pterocarpus*. For F and P repeat types, they are absent in the range of 40-49 bp in the scented rosewood species, while in 20-29 bp group, the C type repeats are lowest in Siam rosewood species (Figure 3). Given these differences in type

and abundance, markers for identification of species and/or lineages could be devised based on different repeats. Fixed differences in repeat location, abundance, and type in a genome have provided ideal signatures for species or clade identification in previous studies (Saltonstall and Lambertini, 2012; Brazda *et al.*, 2018; Zhang *et al.*, 2019).

In chloroplasts, SSRs are often used for population genetics and/or phylogenetic analysis. Among all 43 chloroplast genomes, 86.7% (5579/6423) of the SSRs were single nucleotide A/T motifs. The genus *Arachis* was found to have less than half the number of A/T SSRs than the other species used in this study suggesting that indels are more abundant in long A/T stretches for these species. Similarly, the number of nearly all other, save AG/CT and AGAT/ATCT, SSR motif types were far fewer in *Arachis* than the other species (Figure 4). Within *Dalbergia*, the number of SSRs vary greatly to the point where certain motifs such as AAT/ATT are present in the scented rosewoods and absent in the black rosewoods. Additionally, the presence or absence of certain SSRs differed within a given species such as AT/AT in *D. odorifera* and C/G in *D. tonkinensis*. The results from the SSR analyses suggests that these genomic regions might be useful for the identification of populations, species, and clades of *Dalbergia* if the data from different SSR motif types and length difference is combined in a nested analysis.

2.3 Genome Sequence Divergence

To further characterize the differences between chloroplast genomes, we employed mVISTA to find regions of greatest difference between conserved regions in the eight newly sequenced genomes and *Pterocarpus indicus* Willd. (an outgroup species that also produces high-quality rosewood type lumber). The intergenic and intragenic regions were found to have the least similarity between chloroplast genomes in *Dalbergia* and *P. indicus* (Figure 5), especially in LSC (from *psbA* to *rps19*) and SSC regions (from *ndhF* to *ycf1*). Given these results there are numerous intergenic and intragenic regions for developing markers to differentiate *Pterocarpus* from *Dalbergia* and fewer, but a sufficient number of regions to use for within *Dalbergia* species or clade level differentiation.

In order to more comprehensively assess regions of dissimilarity we used T-coffee to compare all 43 samples using complete genome sequences (Figure 6, Supplemental Table 4). As in other plant lineages (Gu *et al.*, 2019) CDS regions had very high identity scores across all 43 chloroplast genomes while the lowest identity scores were found in *rps8-rpl14* (score 767; length 524bp), *trnR-UCU-trnG-UCC* (score 788; length 592bp), *accD-psaI* (score 790; length 824bp), *psbA-trnK-UUU* (score 797; length 658bp), and *ndhG-ndhI* (score 804; length 1308bp). The discovery of these hypervariable intergenic regions provide candidate regions for the development of genetic markers. While most intergenic regions are more variable *psaA-psaB* (score 1000; length 25bp), *psbL-psbF* (score, 1000; length, 22bp), *psbF-psbE* (score, 1000; length, 9), and *ndhA-ndhH* (score, 1000; length, 1) where noted as being very similar and short in length. As such these regions should be excluded from further consideration for marker development but might be useful in providing priming sites adjacent to variable regions in the development of molecular assays. Given that these photosynthesis genes are clustered in a single operon strong selection of function has resulted in nucleotide conservation even in the intergenic regions. However, in general

sufficient nucleotide divergence has been found for the development of genetic markers in *Dalbergia* and to closely related genera that produce lumber of a similar appearance.

In a third approach to find regions of fixed differences for identifying *Dalbergia* species we analyzed the dN (nonsynonymous substitution rates), dS (synonymous substitution rates) and the ratio of dN/dS (quantify strength of selection) of all genes with PAML. This approach is particularly useful in finding fixed differences that persist through a given lineage because mutations that are undergoing different modes of selection can be detected. From this analysis the dN of all genes was relatively low, while for the dS, for two ribosomal genes (*rps16* and *rpl32*) were outliers. Except for *ycf2* (1.02), the dN/dS of all the other genes are less than 1, indicating that they are subject to purification selection, especially photosynthesis genes, which are lower on average than the other four gene categories (Figure 7).

2.4 Codon usage in *Dalbergia* chloroplast genes

Relative Synonymous Codon Usage (RSCU) is often used to analyze the frequency of codon usage, with the higher values scaling with usage frequency (Figure 8). Given the conserved nature of codon usage, if a mutation is detected, it general remains fixed and as such can provide a useful marker locus. Given this we analyzed the codon usage of *Dalbergia* chloroplast genes. There are no large differences in RSCU between *Dalbergia* chloroplast genes, indicating conservation in codon usage across *Dalbergia*, and a very limited number of loci for clade or species identification.

2.5 Phylogenetic analysis and barcode selection

In order to assess chloroplast genome divergence in an evolutionary context and find synapomorphies (and ultimately barcodes) for given clades we conducted a phylogenetic analysis using 77 coding genes in 43 chloroplast genomes (Figure 9; Supplemental Figure 3). The resulting phylogenetic tree resolved several instances of polyphyly and paraphyly in regard to the species names applied to a given NCBI accession. For instance, *D. sissoo* MN936016 resolved in an early diverging position to a clade of *D. vietnamesis* P.H. Hô & Niyomdhan + *D. yunnanensis* Franch. + *D. tonkinensis* + *D. hainanensis* Merr. & Chun + *D. odorifera* with high support (BS = 100) and *D. sissoo* DC. MN251242 resolved in an early diverging position to a clade of *D. cochinchinensis* + *D. hupeana* with high support (BS = 100) making this species polyphyletic. Another clear example of polyphyly is *D. hainanensis* NC_036961 resolving in a clade with *D. odorifera* while *D. hainanensis* MN251246 resolved in an early diverging position to a clade of *D. hupeana* + *D. balanense* Prain making *D. hupeana* polyphyletic. Instances of paraphyly are also present as in the branching order of *D. sissoo*, *D. cochinchinensis*, and *D. hupeana* in the Siam rosewood category with other examples found elsewhere in the tree (Figure 9). These discrepancies between taxonomy and phylogeny indicate that some of the species identifications for *Dalbergia* NCBI accessions are probably incorrect. Alternatively, some of the examples of polyphyly might be the result of interspecific hybridization where a chloroplast genome was maternally inherited from a more distantly related species. Ultimately, in either case improper identification (of hybrids and/or species) has led to discordance between the topology and taxonomic designations. These discrepancies should be addressed through identification of samples by taxonomic experts where possible as well as

resequencing checks using nuclear loci (e.g. ITS) and reanalysis as well as increasing the number of different species used in phylogenetic analyses. Furthermore, comparative phylogenetic approaches can be used to isolate which taxa are most likely misidentified in the chloroplast data. For instance, *D. sissoo* MN936016 is probably correctly identified based on the location it resolved in the phylogenetic tree (early diverging to a clade containing *D. tonkinensis* in both trees) in this paper compared to the position in the phylogenetic analyses of Vatanparast *et al.* (2013) and Hassold *et al.* (2016). Issues of polyphyly in the Siam rosewood clade were further verified using whole chloroplast comparisons. Differences are apparent between the phylogenetic tree presented here and previously published trees as is expected given the different genomic regions and species sampled. That said consistency is found among the two trees in the membership of important taxa in separate clades and branching order. For example, *D. nigra* has an early diverging position to clade V in Vatanparast *et al*. (2013) and to a clade with similar membership in the phylogenetic tree presented here. These similarities in phylogenetic topology suggest that both ITS and chloroplast barcoding could be employed to identify wood samples. However, much greater within species sampling is needed to compare the stability of polymorphisms within species for each locus before a molecular assay can be deployed. This is especially true of ITS where the large number tandem copies can be found in multiple ribotypes (Teruel *et al.* 2013) and potentially affect the accuracy of molecular assays. Lastly it should be noted that in our analyses the black rosewood category does not form a monophyletic grouping whereas scented and Siam rosewood categories do form monophyletic groupings provided the issue of incorrect labeling and polyphyly can be corrected.

From a MAFFT alignment of the coding regions from 43 chloroplast genomes, we scanned for loci rich in SNVs (single nucleotide variants) and INDELS (INsertions DELetions) to identify potential barcode loci (Supplemental File 1 and 2). The MAFFT alignment after manual correction was 219,992 sites in total length. From this alignment 58 SNV loci and 10 INDEL loci were found to identify Siam rosewoods, 29 SNVs and 13 INDELs in scented rosewood, and 2 SNVs and 1 INDEL for black rosewood if *D. nigra* was taken into account. If *D. nigra* was removed 229 SNVs and 49 INDELs were found for the identification of the clade *D. cultrata* Benth. + *D. fusca*. However, since *D. nigra* is the highest priority *Dalbergia* species in CITES a set of 94 SNVs and 69 INDELs were isolated for the identification of this species from all other *Dalbergia* used in this study (Table 2; Supplemental Table 5 and 6).

Table 2 Number of candidate molecular markers for black, Siam and scented rosewood.

Group		Unique INDEL	Unique SNV
Black rosewood		1	2
Black rosewood	*Dalbergia fusca* *Dalbergia cultrata*	49	229
Black rosewood	*Dalbergia nigra*	69	94
Siam rosewood		10	58
Scented rosewood		13	29
Scented rosewood	*Dalbergia odorifera*	3	7
Scented rosewood	*Dalbergia tonkinensis*	3	0

3. Methods

3.1 Tissue samples and DNA extraction

We collected fresh leaves from eight species of *Dalbergia* for DNA extraction. The leaf material from seedling plants were collected at the Experimental Station of the Research Institute of Tropical Forestry, Chinese Academy of Forestry, Jianfeng Town, Ledong Li Autonomous County, China (18.69°N, 108.79°E). *Dalbergia odorifera*, 40 years old originated from Hainan, China, *D. cochinchinensis* 13 years old from Baan Subprik, Muak-Lek, Saraburi, Thailand, *D. tonkinensis* from Lạng Sơn, Vietnam 5 years old, *D. oliveri* 10 years old from Pursat, Cambodia, *D. bariensis* Pierre 8 years old from Muang district, Khon Kaen, Thailand, *D. fusca* 35 years old from Yunnan, China, *D. hupeana* 30 years old from Jiangsu, China, and *D. nigra* 2 years of age, which originated from Curitiba, Brazil. To obtain the chloroplast genome sequences, the genomic DNA was extracted by QIAGEN DNeasy Plant Maxi Kit (Cat. NO 68163) for Illumina paired-end sequencing.

3.2 Genome sequencing and assembly

The Illumina HiSeq 2500 platform was used to sequence the raw data with insert sizes of 500 bp, for 150 bp paired-end read lengths. Raw data was quality control filtered with the following criteria: filtered reads with adapters, filtered reads with N bases >10%, and filtered reads with low-quality bases (≤5) >50%, which yielded 2 Gb of clean reads for each species by Trimmomatic (Bolger *et al.*, 2014).

All paired-end clean reads were aligned to the chloroplast database (containing all published chloroplast genomes from NCBI by the date November 26, 2019) with bwa v0.7.17-r1188 (Li, 2013) software, and then the Picard v2.20.3 program was used to select chloroplast reads. The selected chloroplast reads were assembled by Spades v3.14.0 (Nurk *et al.*, 2013) with default parameters and the output scaffolds (GFA file) were imported into Bandage v0.8.1 (Wick *et al.*, 2015) to generate the final chloroplast genome for each species.

3.3 Genome annotation

All 43 chloroplast genomes, which contained 8 newly assembled *Dalbergia* chloroplast genomes, and 27 other published genomes of Dalbergieae, four *Arachis* L. species, two *Pterocarpus*, one *Tipuana* Benth. species, and an outgroup species *Amorpha fruticose* L., were compiled for analysis, with all NCBI accession numbers listed in supplemental table 1. The *D. tonkinensis* genome was used as a reference, with all genes delimited manually to provide a complete reference template. The re-annotation of all species was then executed using Plastid Genome Annotator (PGA, https://github.com/quxiaojian/PGA, Qu *et al.*, 2019), and the visualization of genome structure was implemented by the Draw Organelle Genome Maps online software (OGDRAW v1.3.1 https://chlorobox.mpimp-golm.mpg.de/OGDraw.html, Greiner *et al.*, 2019).

3.4 Genome Structure Analysis

Four repeat types in 43 chloroplast genomes, F (forward), P (palindrome), R (reverse), and C (complement) were identified using REPuter Kurtz *et al.*, 2001) with default settings. Simple sequence repeats (SSRs) were detected using the Perl script MISA (Thiel *et al.*, 2003; Beier *et al.*, 2017), with 10, 6, 5, 5, 5, and 5 repeat units set for mono-, di-, tri-, tetra-, penta-, and hexa-motif microsatellites set as the minimum threshold respectively. CodonW v1.4.4 (Sharp and Li, 1987) was employed to assess codon distribution on the basis of relative synonymous codon usage (RSCU) ratio.

3.5 Genome Nucleotide Diversity

Analyses of genome sequence diversity was done using an online software mVISTA (http://genome.lbl.gov/vista/mvista/submit.shtml, Frazer *et al.*, 2004) to compare the 8 newly assembled *Dalbergia* species using Shuffle-LAGAN (Brudno *et al.*, 2003) alignment program with the *P. indicus* chloroplast genome used as a reference. All 43 chloroplast genomes were split into several parts based on annotation files, and the overall consistency score of each part was calculated with multiple sequence alignment tools using T-Coffee (Notredame *et al.*, 2000) in default mode.

3.6 Phylogenetic Analysis and Nucleotide Substitutions

The whole chloroplast genome sequence alignment of 43 chloroplast genomes were generated using MAFFT v7.464 (Katoh *et al.*, 2002; Katoh and Standley, 2013) software, with TrimAL v1.4(Capella-Gutierrez *et al.*, 2009) used to trim the poorly aligned positions. The longest CDS sequences of 77 protein-coding genes were extracted from each genome according to the annotation files, and also aligned using MAFFT (Katoh *et al.*, 2002; Katoh and Standley, 2013). The nucleotide sequence alignments of 77 protein-coding genes were concatenated. This data set was further used to resolve the phylogenetic tree using IQ-TREE v2.0 (Nguyen *et al.*, 2015; Minh *et al.*, 2020) with 1000 ultrafast bootstrap replicates to assess branch support with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree) used for tree visualization. CODEML in PAML v4.9 (Yang, 1997) was used to estimate the nonsynonymous (dN), synonymous (dS) and the ratio of nonsynonymous to synonymous nucleotide substitutions (dN/dS) for each gene.

Declarations

Acknowledgements

We sincerely thank Shanghai BIOZERON Biotechnology Co., Ltd. for performing the high throughput sequencing.

Author contributions

Zhou Hong conceived and designed the study. Dan Peng, Wenchuang He, Zhiqiang Wu and Xuezhu Liao performed the experiments and data analysis. Ningnan Zhang and Zengjiang Yang contributed materials. Zhiqiang Wu and Xuezhu Liao wrote the paper. Luke R. Tembrock, Zhiqiang Wu, Xuezhu Liao and Daping Xu revised the paper. All authors approved the final manuscript.

Funding

This study was co-supported by the Research Funds for the Central Non-profit Research Institution of Chinese Academy of Forestry (CAFYBB2017ZA001-7) and National Natural Science Foundation of China (31500537).

Conflicts of Interest

The authors declare no conflict of interest.

References

Abdul-Rahaman I, Kabanda J, Braimah MM. 2016. Desertification of the Savanna: Illegal Logging of Rosewood, Causes and Effects on the People of Kabonwule, Northern Region. *Saudi Journal of Humanities and Social Sciences* 1: 48-54.

Beier S, Thiel T, Münch T, Scholz U, Mascher M. 2017. MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33: 2583-2585.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Brazda V, Lysek J, Bartas M, Fojta M. 2018. Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. *Biomed Research International* 1-10.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou, S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* 13: 721-731.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972-1973.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America* 106: 12794-12797.

CITES. 2020. The CITES species. https://cites.org/eng/disc/species.php

Du Puy, D.J. 2002. The Leguminosae of Madagascar. Royal Botanic Gardens, Kew, UK.

Espinoza EO, Wiemann MC, Barajas-Morales J, Chavarria GD, McClure PJ. 2015. Forensic Analysis of Cites-Protected *Dalbergia* Timber from the Americas. *IAWA Journal* 36: 311-325.

Fu CN, Wu CS, Ye LJ, Mo ZQ, Liu J, Chang YW, Li DZ, Chaw SM, Gao LM. 2019. Prevalence of isomeric plastomes and effectiveness of plastome super-barcodes in yews (*Taxus*) worldwide. *Scientific Reports* 9: 2773.

Gu C, Ma L, Wu Z, Chen K, Wang Y. 2019. Comparative analyses of chloroplast genomes from 22 Lythraceae species: inferences for phylogenetic relationships and genome evolution within Myrtales. *BMC Plant Biology* 19: 281.

Hassold S, Lowry PN, Bauert MR, Razafintsalama A, Ramamonjisoa L, Widmer A. 2016. DNA Barcoding of Malagasy Rosewoods: Towards a Molecular Identification of CITES-Listed Dalbergia Species. *PLoS One* 11: e157881.

Hollingsworth PM, Li DZ, van der Bank M, Twyford AD. 2016. Telling plant species apart with DNA: from barcodes to genomes. *Philosophical Transactions of the Royal Society B-Biological Sciences* 371.

Innes J. 2010. Madagascar rosewood, illegal logging and the tropical timber trade, *Madagascar Conservation & Development* 5: 6-10.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059-3066.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30: 772-780.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2 [q-bio.GN]*.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37: 1530-1534.

Nguyen L, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32: 268-274.

Nhung NP, Chi NM, Thu PQ, Thuong BH, Ban DV, Dell B. 2020. Market and policy setting for the trade in *Dalbergia tonkinensis*, a rare and valuable rosewood, in Vietnam, Trees, Forests and People 1: 100002.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of Molecular Biology* 302: 205-217.

Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel SR, Woyke T, McLean JS, Lasken R, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology* 20: 714-737.

Patel ER. 2007. Logging of Rare Rosewood and Palisandre (*Dalbergia* spp.) within Marojejy National Park, Madagascar, *Madagascar Conservation & Development* 2: 11-16.

Randriamalala H, Liu Z. 2010. Rosewood of Madagascar: Between democracy and conservation, *Madagascar Conservation & Development* 5: 11-22.

Saltonstall K, Lambertini C. 2012. The value of repetitive sequences in chloroplast DNA for phylogeographic inference: a comment on Vachon & Freeland 2011. *Molecular Ecology Resources* 12: 581-585.

Sharp PM, Li WH. 1987. The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research* 15: 1281-1295.

Siriwat P, Nijman V. 2018. Using online media-sourced seizure data to assess the illegal wildlife trade in Siamese rosewood, *Environmental Conservation* 45: 352-360.

Teruel M, Ruíz-Ruano FJ, Marchal JA, Sánchez A, Cabrero J, Camacho JP, Perfectti F. 2014. Disparate molecular evolution of two types of repetitive DNAs in the genome of the grasshopper *Eyprepocnemis plorans*. *Heredity* 112: 531-542.

Thiel T, Michalek W, Varshney RK, Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics* 106: 411-422.

Thiel T, Michalek W, Varshney R, Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theoretical and Applied Genetics* 106: 411-422.

United Nations Office on Drugs and Crime. 2016. World wildlife crime report: Trafficking in protected species. New York: United Nations Office on Drugs and Crime.

Vardeman E, Velásquez Runk J. 2020. Panama's illegal rosewood logging boom from *Dalbergia retusa*. *Global Ecology and Conservation* 23: e01098.

Vatanparast M, Klitgård BB, Adema FACB, Pennington RT, Yahara T, Kajita T. 2013. First molecular phylogeny of the pantropical genus *Dalbergia*: implications for infrageneric circumscription and biogeography. *South African Journal of Botany* 143-149.

Wang L, Wu ZQ, Bystriakova N, Ansell SW, Xiang QP, Heinrichs J, Schneider H, Zhang XC. 2011. Phylogeography of the Sino-Himalayan fern *Lepisorus clathratus* on "the roof of the world". *PLoS One* 6: e25896.

Wang Q, Yu QS, Liu JQ. 2011. Are nuclear loci ideal for barcoding plants? A case study of genetic delimitation of two sister species using multiple loci and multiple intraspecific individuals. *Journal of Systematics and Evolution* 3: 182-188.

Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31: 3350-3352.

Wu ZQ, Liao XZ, Zhang XN, Tembrock LR, Broz A. 2020 Genomic architectural variation of plant mitochondria—A review of multichromosomal structuring. *Journal of Systematics and Evolution* 1-9.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS* 13: 555-556.

Zhang R, Ge F, Li H, Chen Y, Zhao Y, Gao Y, Liu Z, Yang L. 2019. PCIR: a database of Plant Chloroplast Inverted Repeats. *Database (Oxford)* 1: 2019-2127.

Zhang W, Sun Y, Liu J, Xu C, Zou X, Chen X, Liu Y, Wu P, Yang X, Zhou S. 2020. DNA barcoding of *Oryza*: conventional, specific, and super barcodes. *Plant Molecular Biology* 1-14.

Zhou J, Zhang S, Wang J, Shen H, Ai B, Gao W, Zhang C, Fei Q, Yuan D, Wu Z, Tembrock LR, Li S, Gu C, Liao X. 2021. Chloroplast genomes in *Populus* (Salicaceae): comparisons from an intensively sampled genus reveal dynamic patterns of evolution. *Scientific Reports* 11: 9471.

Zhu AL. 2020. China's Rosewood Boom: A Cultural Fix to Capital Overaccumulation. *Annals of the American Association of Geographers* 1: 277-296.

Figures

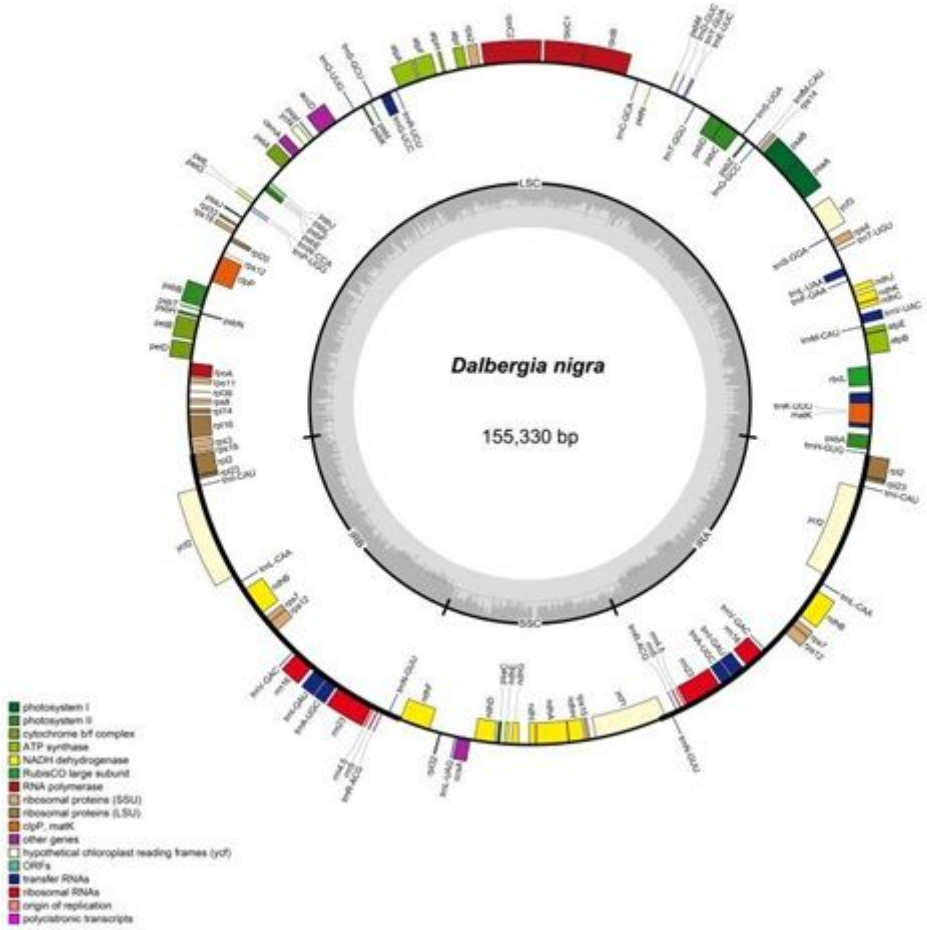

Figure 1

Gene map of the D. nigra chloroplast genome. Genes shown inside the circle are transcribed clockwise, and those outsides are transcribed counterclockwise. The darker gray color in the inner circle corresponds to the GC content. The IRA and IRB (two inverted repeating regions); LSC (long single-copy region); and SSC (short single-copy region) are indicated outside of the GC content.

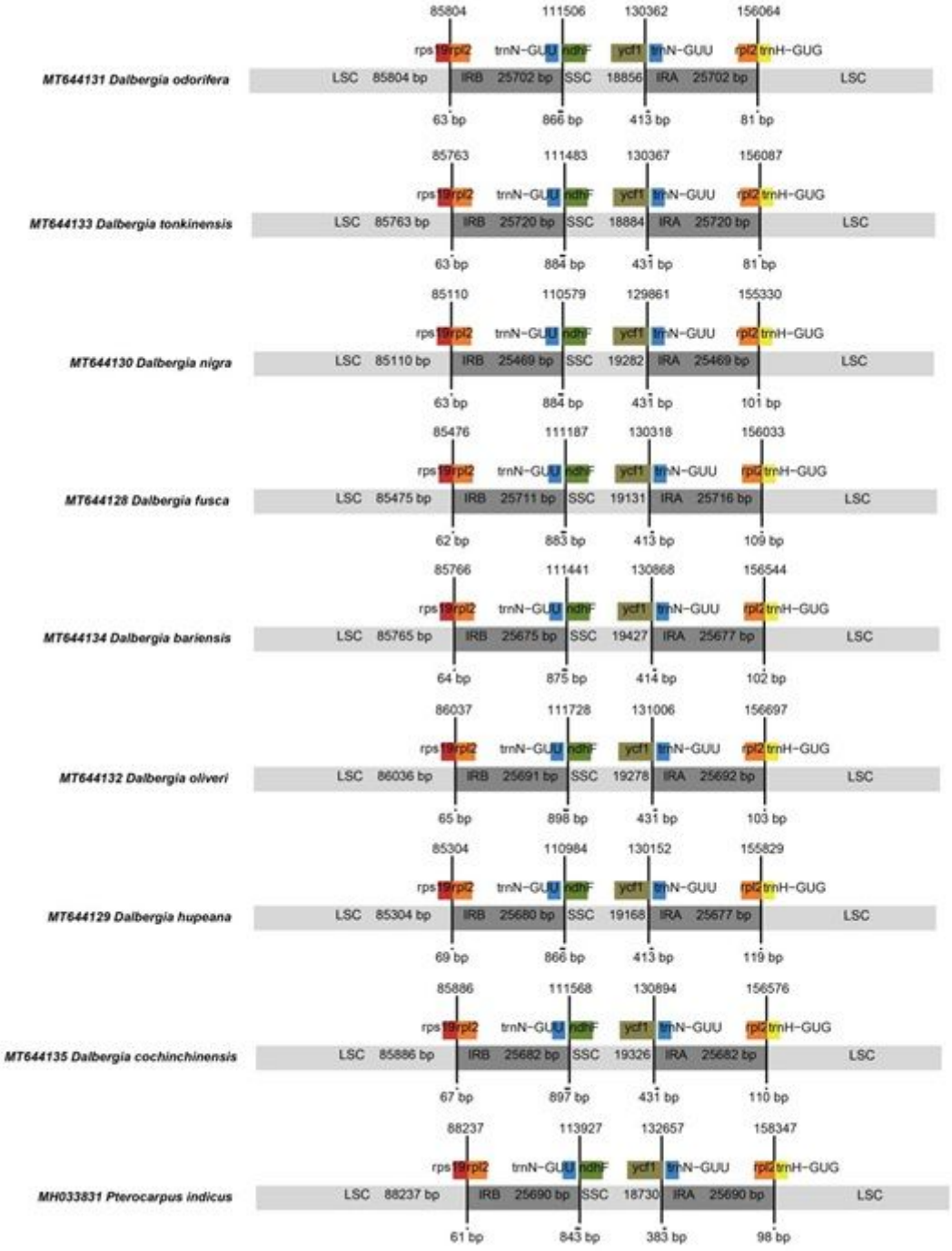

Figure 2

Comparison of junctions between the LSC, SSC, and IR regions among eight newly assembled Dalbergia species and the Pterocarpus indicus chloroplast genome. Figure is not to scale. (LSC Large single-copy, SSC Small single-copy, IR inverted repeat).

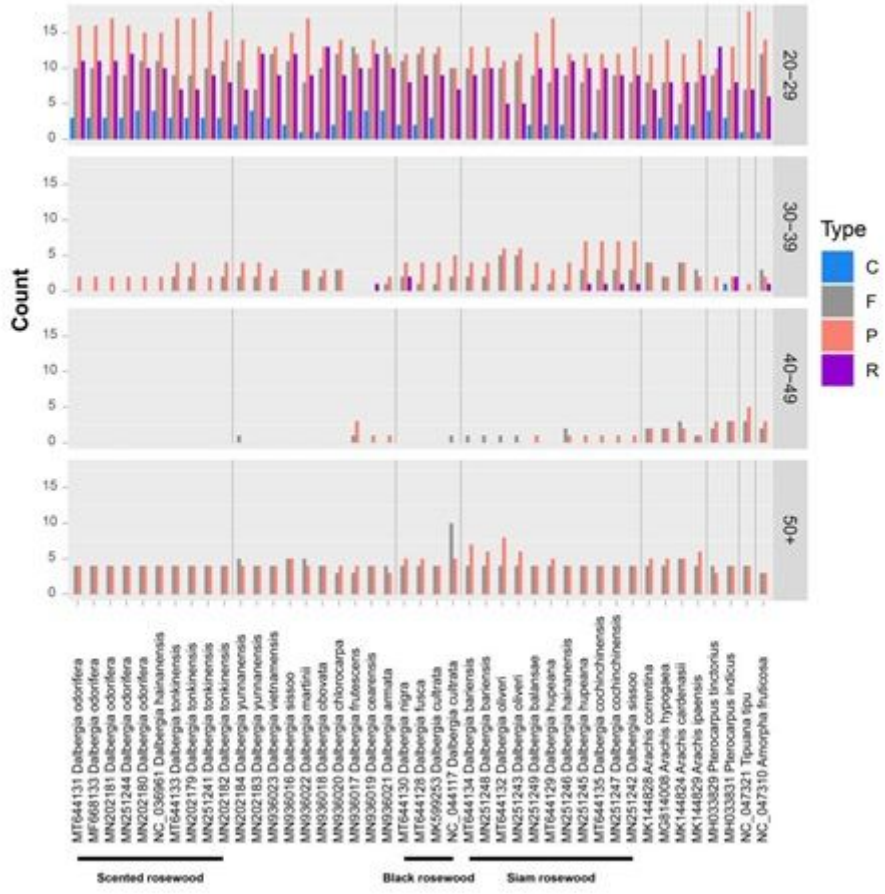

Figure 3

Variation of repeat abundance and type in 43 chloroplast genomes.

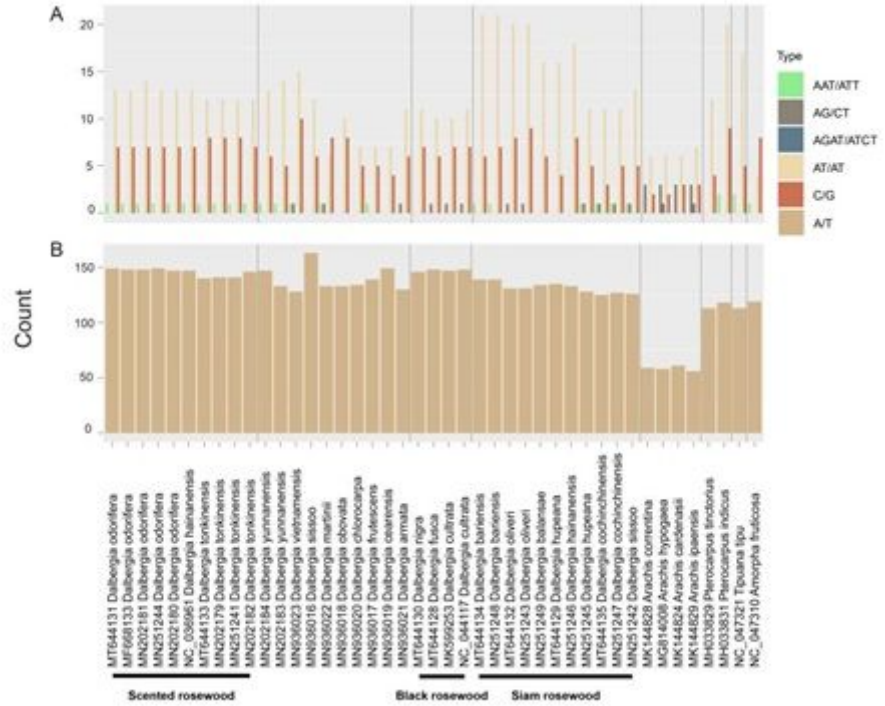

Figure 4

The number of Simple Sequence Repeat (SSR) of different types from 43 chloroplast genomes. A. the number of AAT/ATT, AG/CT, AGAT/ATCT, AT/AT, C/G type SSRs; B, the number of A/T type SSRs in each sample.

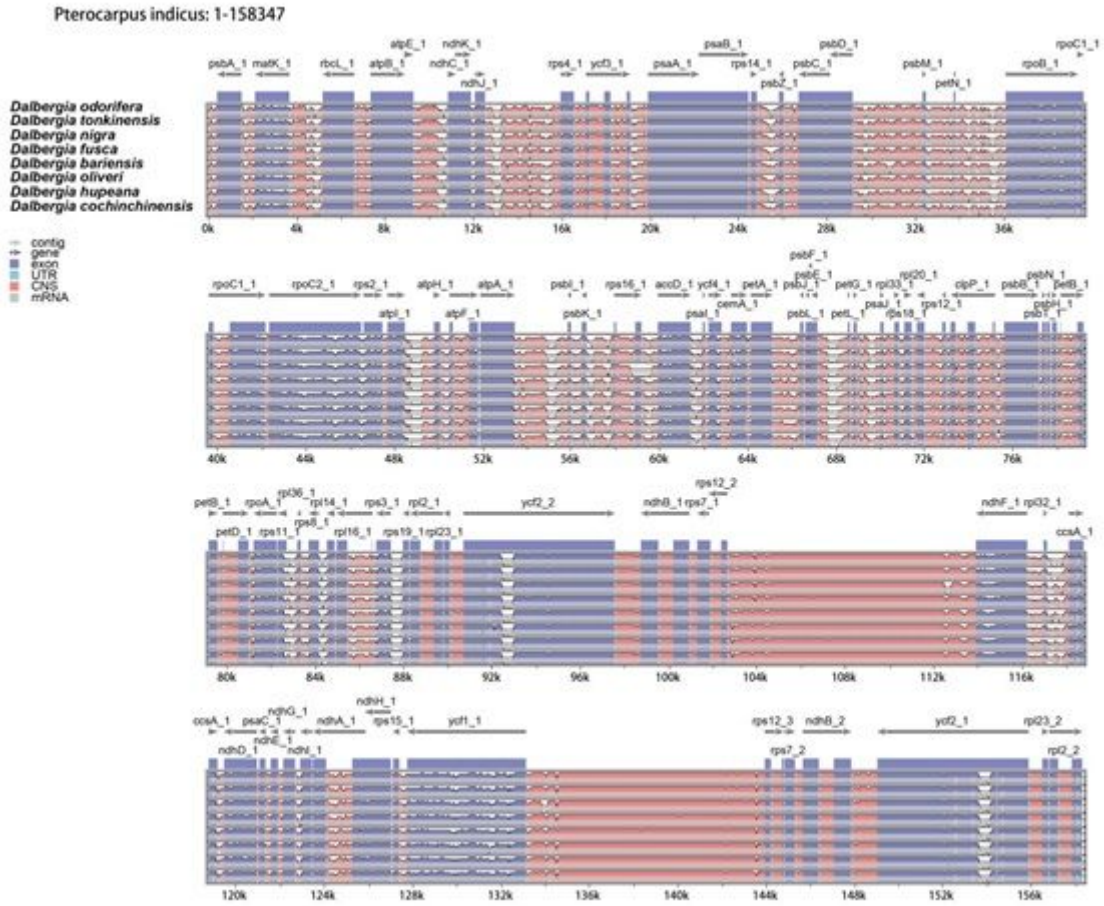

Figure 5

Global alignment of eight newly assembled Dalbergia chloroplast genomes using mVISTA with P. indicus as reference. Y-axis shows the range of sequence identity (50–100%). tRNA and rRNA genes were not present in this figure.

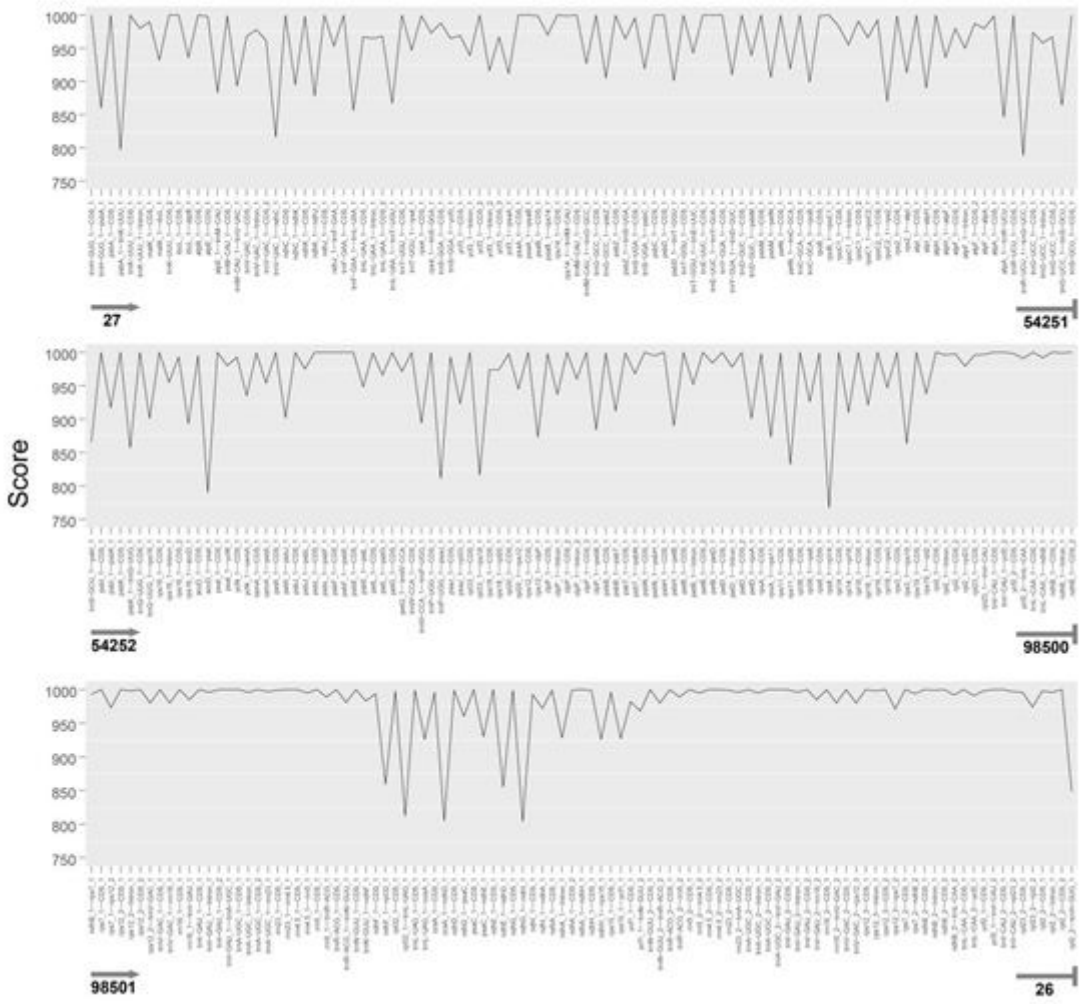

Figure 6

Sequence identity among coding and non-coding regions based on the alignment from 43 chloroplast genomes. T-Coffee was used to calculate the score of identity.

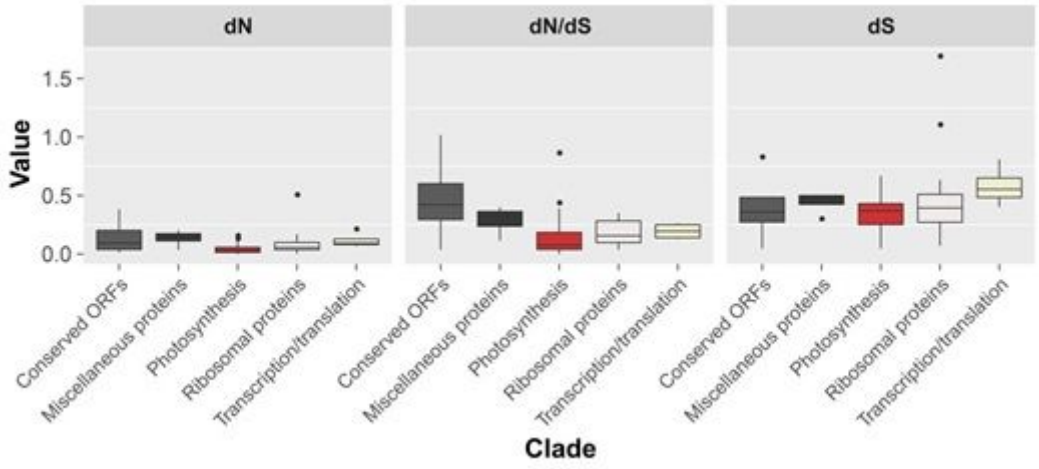

Figure 7

The mode and strength of selection among 77 chloroplast protein coding genes in Dalbergia.

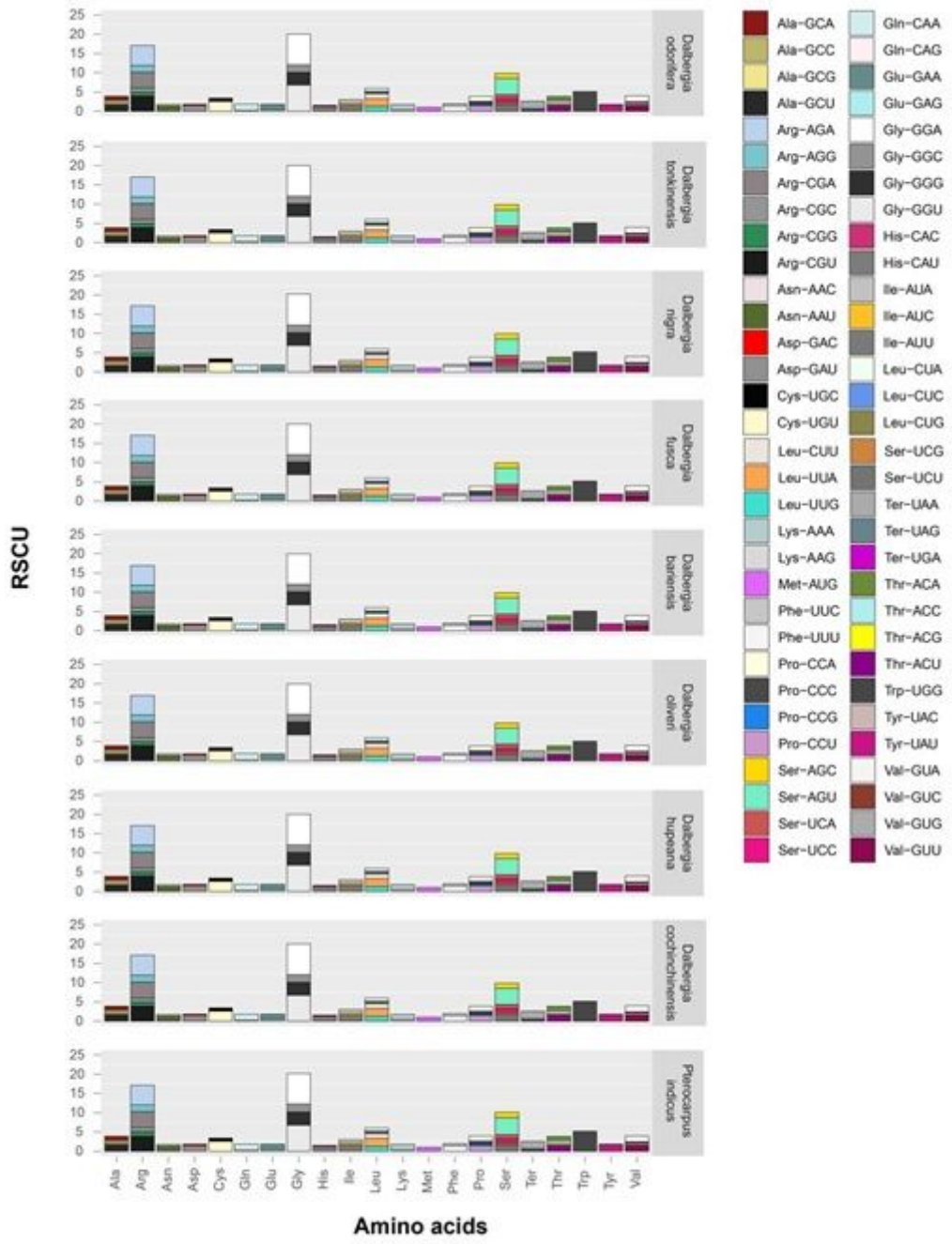

Figure 8

Codon content of 21 amino acids and a stop codon of 77 coding genes of 8 newly assembled Dalbergia chloroplast genome. Color of the histogram corresponds to the color of codons in legend.

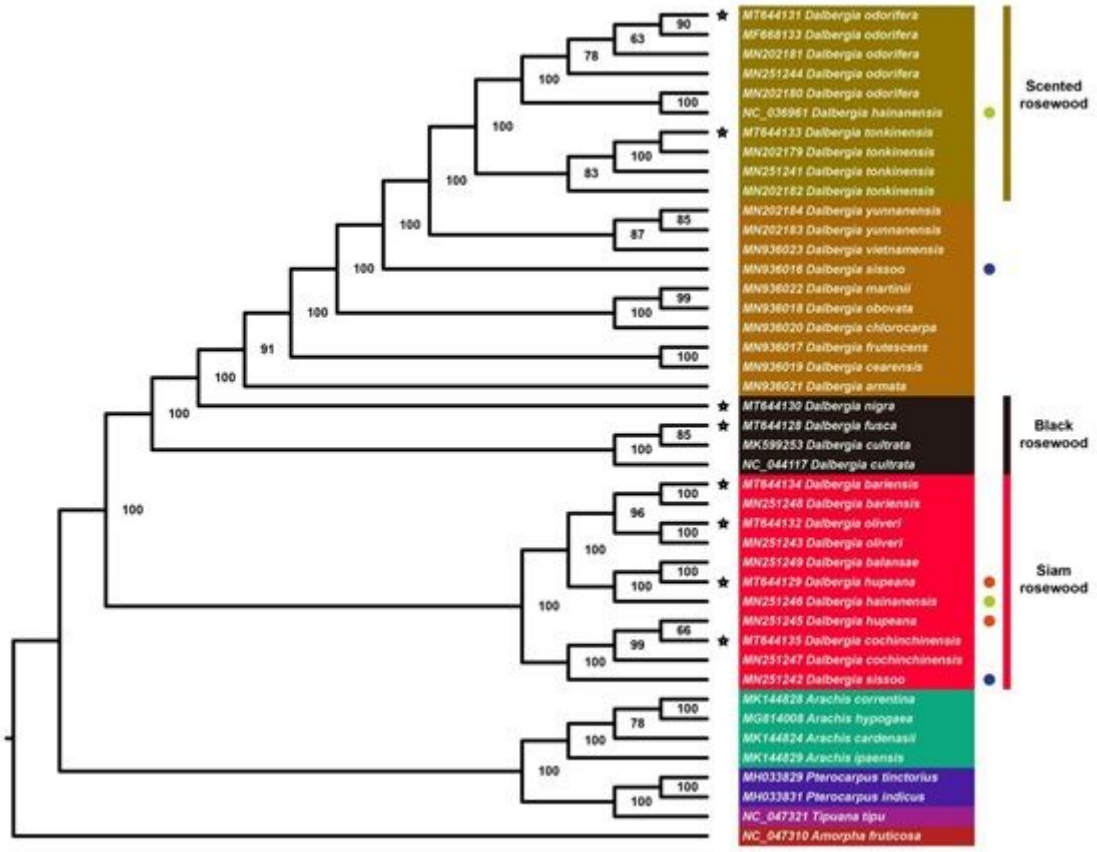

Figure 9

Phylogenetic tree of Dalbergia based on the CDS alignment of 77 chloroplast coding genes. Numbers treat nodes indicate the ultrafast bootstrap values generated by IQ-TREE. Species noted with stars are newly assembled chloroplast genomes from this study. Polyphyletic species are indicated with a dot.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigure1.pdf
- SupplementalFigure2.pdf
- SupplementalFigure3.pdf
- SupplementalFile1.aln
- SupplementalFile2.pl
- SupplementalTable1.xls
- SupplementalTable2.xls
- SupplementalTable3.xls
- SupplementalTable4.xls
- SupplementalTable5.xls

- SupplementalTable6.xls